УДК 519.766.4

Н.В. Кузнецова, П.И. Бидюк

МОДЕЛИРОВАНИЕ КРЕДИТНЫХ РИСКОВ НА ОСНОВЕ ТЕОРИИ ВЫЖИВАНИЯ

Введение

Актуальная задача деятельности банковского сектора — анализ финансового кредитного риска. Кредитование — один из основных источников доходов в банковском деле, инструмент стимулирования экономического развития и вместе с тем — источник вероятных потерь. Оценить, предсказать и предотвратить эти потери возможно на этапе выдачи кредита и в ходе его обслуживания.

Традиционная постановка задачи оценки финансовых (кредитных) рисков для банковского сектора — это оценка новых клиентов при обработке заявок на выдачу кредита. Часть банков до сих пор решают эту задачу исключительно в статическом плане, оценивая вероятность возврата кредита и объем возможных потерь на текущий момент. Идея динамической оценки кредита и клиента заключается в периодической проверке клиента с точки зрения выполнения его обязательств, чтобы предусмотреть возможные проблемы в обслуживании им кредита и своевременной уплате ежемесячной задолженности. Она отличается от стандартного подхода к построению скоринговых моделей [1–3], поскольку позволяет оценивать кредиты до окончания срока, на который они были выданы, своевременно реагировать и разрабатывать актуальные механизмы и сценарии действий в случае появления таких проблемных кредитов.

В настоящей статье предлагается построение математической модели при анализе кредитного риска новым способом, который предполагает динамическое оценивание клиентов. Наряду с традиционными характеристиками клиента и кредита предусматривается возможность прогнозирования момента времени (в месяцах) $0 < t < 12$, предшествующего наступлению неблагоприятного события — момента появления задолженности (просрочки) по кредиту. Таким образом, предполагается итерационная процедура построения скоринговых карт, которые будут предоставлять информацию о поведении заемщиков в момент обслуживания кредита, а также построения так называемой «скоринговой карты поведения».

Скоринговые карты и модели

Скоринговая карта поведения — это математическая модель со свойственной ей совокупностью входящих факторов (характеристик) клиента и кредита, изменяющихся во времени и влияющих на целевую характеристику — переменную, описывающую возможность своевременной уплаты кредита в текущем месяце.

Такая поведенческая скоринговая карта строится для различных типов «типичных» заемщиков и позволяет оценивать вероятность ежемесячной уплаты задолженности. Параллельно с этим банки заинтересованы в объективной информации о возможных потерях в случае неуплаты по кредитам. В соответствии с Базелем II для оценки кредитных рисков банков используется IRB-подход (Internal Ratings-Based Approach) с учетом внутренних рейтингов заемщиков, т.е. рейтингов, устанавливаемых самими банками [4, 5]. Такой подход предоставляет возможность рассчитать сумму, уплаченную по кредиту, и сумму, непокрытую по кредиту, для каждого конкретного кредита в конкретный момент времени.

Введем понятие «цикл успешного обслуживания кредита», которое определяется как количество месяцев или дней, когда осуществляется оплата кредита без просрочек, т.е. $0 < t_{\text{goodcredit}} < 12$ при условии, что задержка $\text{delay} = 0$ дней, $P(\text{delay} = 0) = 1$.

Просрочка более трех, но менее 30 дней, определяется специальным маркером «подозрительного» поведения $I(behavioral\_debts) = 1$ и считается поводом для включения таких клиентов в периодический (более частый) мониторинг с использованием скоринговых карт поведения и возможных средств защиты или противодействия появлению дальнейших просрочек по кредиту.

Теория выживания: общие сведения и предположения для предметной области прогнозирования времени «успешности» кредита

Традиционно модели анализа выживания используются для исследования момента гибели некоторой популяции. Время до наступления этого момента называется временем выживания.

Модели анализа выживания предшествовало создание таблиц смертности, которые использовались в страховании жизни и демографических науках в XVII в. Это привело к употреблению слова «выживание» в контексте уровня смертности. Изначально метод таблиц смертности базировался на широких временных промежутках и больших объемах данных. В 1950-х Каплан и Мейер [4] предложили статистическую оценку кривой выживания. Они разработали метод для коротких временных отрезков и меньших выборок по сравнению с теми, которые использовалась в демографических исследованиях.

Д. Кокс [2] предложил метод, позволяющий добавлять ковариантны к анализу подобных данных, известный как «модель пропорциональных рисков Кокса» (proportional hazards — PH). Такая модель использует регрессоры, не зависящие от времени, или статические переменные и предполагает, что появление рисков не меняется с течением времени. Однако в реальных данных часто возникают характеристики, изменяющиеся со временем. Такие переменные нарушают предположение о постоянстве отношения, поэтому модель Кокса была модифицирована и дополнена. Известны ее стратифицированная и обобщенная модификации.

Применение теории анализа выживания для моделирования кредитных рисков предложено недавно. Так, в работе [4] указываются преимущества методов анализа выживания по сравнению с общепринятыми статистическими методами. Более передовая методология выживания использует большее количество информации, чем обычные модели, поскольку она позволяет детализировать поведение путем цензуры и за счет использования переменной времени, что нельзя непосредственно применить ни в линейной, ни в логистической регрессии. К тому же не нужно делать никаких предположений относительно распределения переменной выхода. Именно такие рекомендации стали исходными для более глубокого исследования авторами методов

анализа выживания и их усовершенствования в контексте применения для анализа времени платежеспособности клиентов — владельцев кредитных карт (КК). В частности, в данной работе выполнена формализация постановки задачи прогнозирования времени беспроблемного обслуживания кредита и экспериментальные исследования с применением моделей пропорциональных рисков.

Правыми цензурированными наблюдениями называют такие, которые прекращаются до наступления события.

Наблюдение называется цензурированным слева, если исследуется до начала периода наблюдения.

Интервальными цензурированными наблюдениями называются наблюдения, если известна лишь информация о том, что время выживания распределено между переменными a и b ($t \in [a, b]$).

Типы правого цензурирования:

1) субъекты исследования выжили до конца исследования; время цензуры фиксированное;

2) субъекты исследования выжили до конца исследования; время цензуры наступит, когда произойдет предварительно определенное количество событий;

3) случайные наблюдения прекращаются по причинам, которые не могут быть контролируемыми исследователем.

Для исследования кредитных карточек определим правила цензурирования следующим образом. КК, по которым были просрочены менее трех платежей на сумму не менее 100 грн., считаются дефолтными, т.е. «плохими» в контексте данного исследования. Все остальные результаты отсекаются, т.е. считаются не дошедшими до своего логического конца.

Формализация задачи прогнозирования времени беспроблемности обслуживания банковского кредита. Условная функция выживания, используемая для моделирования кредитного риска, открывает интересную перспективу для изучения дефолта. Вместо того, чтобы определять есть ли дефолт, оцениваем время его наступления, учитывая кредитную информацию клиентов (эндогенные ковариванты) и рассматривая индикаторы для экономического цикла (экзогенные ковариванты). Таким образом, риск дефолта измеряется посредством условного распределения случайной переменной времени до дефолта, T, заданного вектором ковариант, X. Из-за механизма цензурирования переменная T не является полностью наблюдаемой.

Поскольку на практике доля просроченных кредитов маленькая, доля цензурированных данных большая, это может привести к плохой производительности статистических методов. С другой стороны, размер выборки обычно очень большой. Это облегчает проблему значительной доли цензурирования [2].

Условный анализ выживания в кредитном риске

Использование методов анализа выживания для изучения кредитного риска и, в частности, для модели вероятности дефолта (PD), можно иллюстрировать с помощью рис. 1, где представлены три распространенные ситуации, которые могут возникнуть на практике, когда кредитная компания соблюдает «срок» кредита. Рассмотрим интервал PD $[0, \tau]$ в качестве горизонта исследования. Случай *a* показывает кредит с дефолтом до конечной точки времени при исследовании (τ). В этом случае время жизни кредита, T, которое является временем до дефолта кредита, — наблюдаемая переменная. Случаи (*b*) и (*c*) показывают две различные ситуации. В обоих случаях невозможно наблюдать момент, когда кредит вступает в дефолт, что вызывает нехватку информации, поступающей из правого отсечения.

В случае (*b*) это только время от начала кредитования до конца исследования, тогда (*c*) учитывает ситуации, когда ожидается отказ от кредита или погашения кредита до наступления дефолта.

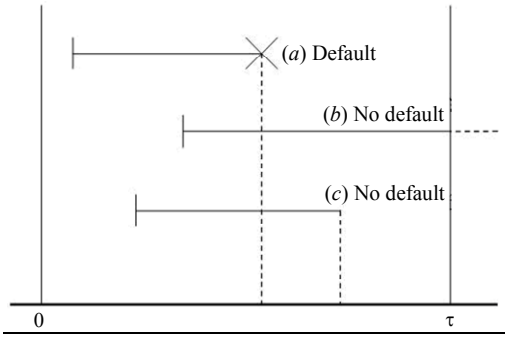

Рис. 1

Доступной информацией для моделирования PD является выборка из n независимых одинаково распределенных величин $\{(Y_1, X_1, \delta_1), ..., (Y_n, X_n, \delta_n)\}$, случайного вектора $\{Y, X, \delta\}$, где $Y = \min\{T, C\}$ — наблюдаемая зрелость; T — время до наступления дефолта; C — время до окончания исследования или предполагаемого аннулирования кредита; $\delta = (T \le c)$ — индикатор нецензурирования и X — вектор поясняющих ковариант. Здесь предполагается независимость величин T и X, а также условная независимость величин T и C для данного X.

С учетом предыдущих допущений можно полностью характеризировать условное распределение случайной величины T, используя некоторые общие соотношения в анализе выживаемости. Таким образом, функция условной выживаемости — $S(t \mid x)$, условный уровень опасности — $\lambda(t \mid x)$, условная кумулятивная функция риска — $\Lambda(t \mid x)$, условная функция распределения — $F(t \mid x)$, связаны следующим образом [2]:

$$S(t \mid x) = P(T > t \mid X = x) = \int_t^{\infty} f(u \mid x)du,$$

$$\lambda(t \mid x) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t, X = x)}{\Delta t} = \frac{f(t \mid x)}{S(t \mid x)},$$

$$\Lambda(t \mid x) = \int_0^t \lambda(u \mid x)du = \int_0^t \frac{f(t \mid x)}{S(t \mid x)}du,$$

$$S(t \mid x) = e^{-\Lambda(t|x)},$$

$$F(t \mid x) = 1 - S(t \mid x).$$

В данной статье используются различные подходы к моделированию PD, применяя условный анализ выживания. Все модели основаны на записи PD с точки зрения условной функции распределения времени до дефолта. Таким образом, PD можно оценить как с помощью моделей логистической регрессии, так и пропорциональных рисков Кокса, в которой оценка функции выживаемости получена решением уравнений частичного правдоподобия. Регрессионная модель Кокса дает $\hat{\text{PD}}^{PHM}$ с помощью обобщенной линейной модели с параметрами, оцененными методом максимального правдоподобия; в результате получаем модель $\hat{\text{PD}}^{GLM}$.

Моделирование вероятности дефолта функцией условного распределения. В соответствии с требованиями Базеля II [5] модели кредитного скоринга используются для измерения вероятности дефолта на горизонте времени $t + b$ со временем

зрелости t. Типичное значение $b = 12$ (в месяцах). Таким образом должна быть вычислена следующая вероятность [2]:

$$\mathrm{PD}(t \mid x) = P(t \le T < \mathrm{t} + \mathrm{b} \mid \mathrm{T} \ge \mathrm{t}, \mathrm{X} = \mathrm{x}) = \frac{P(T < \mathrm{t} + \mathrm{b} \mid \mathrm{X} = \mathrm{x}) - \mathrm{P}(\mathrm{T} \le \mathrm{t} \mid \mathrm{X} = \mathrm{x})}{\mathrm{P}(\mathrm{T} \ge \mathrm{t} \mid \mathrm{X} = \mathrm{x})} =$$

$$= \frac{F(t+b \mid x) - F(t \mid x)}{1 - F(t \mid x)} = 1 - \frac{S(t+b \mid x)}{S(t \mid x)}, \qquad (1)$$

где t — наблюдаемый срок погашения кредита, x — значение ковариационного вектора X для этого кредита.

Модели пропорциональных рисков. В данной статье используется полупараметрический подход к оцениванию пропорциональных рисков Кокса для функции условного выживания $S(t \mid x)$, оценивающей совокупную условную функцию риска, — $L(t \mid x)$, с использованием метода максимального правдоподобия. Необходимо разработать условную модель для индивидуального $S(t \mid x)$, которая определена в терминах $L(t \mid x)$. Для того чтобы описать $\hat{\mathrm{PD}}^{PHM}$, приведем некоторые определения из теории Кокса [2, 6].

Оценка функции условного уровня риска определяется следующим образом:

$$\hat{\lambda}(t \mid x) = \hat{\lambda}_0(t) \exp(x^{\mathrm{T}} \hat{\beta}), \qquad (2)$$

где $\hat{\lambda}_0(t)$ — оценка базовой функции уровня риска $\lambda_0(t)$, $\hat{\beta}$ — оценка вектора параметров β.

Таким образом, в предположении о существовании модели пропорциональных рисков PD оценивается как

$$\hat{\mathrm{PD}}^{PHM}(t \mid x) = \frac{\hat{F}_{\hat{\beta}}(t+b \mid x) - \hat{F}_{\hat{\beta}}(t \mid x)}{1 - \hat{F}_{\hat{\beta}}(t \mid x)} = 1 - \frac{\hat{S}_{\hat{\beta}}(t+b \mid x)}{\hat{S}_{\hat{\beta}}(t \mid x)}, \qquad (3)$$

где $1 - \hat{F}_{\hat{\beta}}(t \mid x) = \hat{S}_{\hat{\beta}}(t \mid x) = \exp(-\hat{\Lambda}(t \mid x))$.

Метод оценивания для этой модели состоит из двух шагов. На первом шаге интегральная функция базового риска $\Lambda_0(t)$ оценивается так:

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \frac{1\{Y_i \le t, \delta_i = 1\}}{\sum_{j=1}^{n} 1\{Y_j \ge Y_i\}}. \qquad (4)$$

Тогда параметр β:

$$\hat{\beta}^{PHM} = \arg\max_{\beta} L(\beta), \qquad (5)$$

где частичная функция правдоподобности задается выражением

$$L(\beta) = \prod_{i=1}^{n} \frac{\exp(x_i^T \beta)}{\left(\sum_{j=1}^{n} 1_{\{Y_j > Y_i\}} \exp(x_j^T \beta) \right)}. \qquad (6)$$

Таким образом, оценка условной интегральной функции риска вычисляется по формуле

$$\hat{\Lambda}(t \mid x) = \int_0^t \hat{\lambda}(s \mid t)ds = \exp(x^T \hat{\beta}^{PHM})\hat{\Lambda}_0(t) . \quad (7)$$

Асимптотические свойства этой оценки подробно изложены в [7]. Аналогичные соотношения можно получить для оценки PD, определенной в (3).

В контексте потребительских кредитов популяция по теории выживания состоит из индивидов с кредитами в форме КК или других займов, живущих по следующим правилам:

— пользователь кредита перестает выполнять свои обязательства (переходит в состояние дефолта) по погашению задолженности, это считается его гибелью;

— время выживания измеряется, начиная с даты открытия счета;

— если клиент никогда не переходит в дефолт в течение периода наблюдения, то он подлежит цензуре в точке наблюдения, т.е. фиксируется и не используется при построении модели.

Предварительная подготовка данных

Входные данные, использованные для экспериментальных исследований, включали в себя информацию о КК, выданных в разные периоды времени с 2013 по 2015 годы. Для того чтобы использовать как можно больше доступных записей, решено рассматривать КК в разрезе продолжительности их жизни. Тогда получаем множество КК, которые будто бы начинаются в один момент. При этом, поскольку логистическая регрессия по своей природе является моделью статической, нужно сформировать вектор характеристик на определенный момент жизни каждого кредита. Для преодоления этой проблемы решено считать «периодом созревания» соглашения временной интервал семь месяцев.

При построении поведенческой скоринговой карты с использованием логистической регрессии нельзя обойтись без агрегирования исторических (по отношению к периоду созревания) данных, потому что иначе такую модель вообще нельзя считать поведенческой. В связи с этим в регрессоры были включены максимальные, минимальные и средние значения определенных параметров за период с 1-го по 7-й месяцы.

В результате такой предварительной обработки получен следующий список параметров: идентификатор сделки, номер месяца жизни кредита, данные о поведении клиента (остаток по телу кредита, процентам, просрочки по телу и проценту, количество дней просрочки по телу и процентам, сумма просрочки, лимит по КК, количество снятых и возвращенных денег, номер месяца первой просрочки более 100 грн.), аппликационные данные (тип клиента, возраст, год и период выдачи кредита, лимит на начало сделки, запрашиваемая сумма кредита, прописка, количество иждивенцев, доход, общий скоринговый балл, возраст клиента как контрагента, время последней прописки), агрегированные данные (максимальная задолженность по телу, максимальное количество просроченных месяцев, максимальная сумма, среднее значение ежемесячного снятия, платежа, максимальное значение отношения просрочки к установленному лимиту) и целевое поле (дефолт/не дефолт).

В процессе предварительного исследования данных установлено (табл. 1), что слишком мало записей попадает в определенные категории характеристики client type. Поэтому было принято решение объединить наименее репрезентативные из них: Employee с Insider и Salary с OldSalary. Такое решение объясняется также тем, что они отражают схожие категории клиентов. В первом случае — это группы связанных лиц и работников банка, во втором — участники зарплатных проектов и те, кто когда-то были их участниками.

Таблица 1

Название категории	Количество записей	Процент от общего количества, %
NULL	220	7,407
Client	125	4,209
Employee	100	3,367
Insider	9	0,303
Normal	1213	40,842
OldSalary	245	8,25
Salary	1058	35,622

После выполнения предварительной обработки была построена модель в соответствии с формулой

$$p = E(y \mid x) = \frac{e^{\beta^{\mathrm{T}} x}}{1 + e^{\beta^{T} x}} = \sigma(e^{\beta^{\mathrm{T}} x}) \tag{8}$$

средствами языка программирования *R* и оценены коэффициенты модели логистической регрессии [8]. В результате построена модель с оцененными коэффициентами $\ln \frac{p}{1-p} = \sum_{i=1}^{20} \beta_i x_i$.

```
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.863284   0.613071  -1.408 0.159093
age                0.004807   0.005337   0.901 0.367745
is_satisfiedt     -0.411450   0.119419  -3.445 0.000570 ***
liv_is_regt       -0.018226   0.129879  -0.140 0.888402
childcnt           0.030690   0.078934   0.389 0.697421
dependantcnt       0.011642   0.124057   0.094 0.925233
log(1 + income)    0.029583   0.022285   1.327 0.184346
has_u_scoret       0.879906   0.580642   1.515 0.129671
with_bank_mon     -0.018143   0.009391  -1.932 0.053368 .
reg_mon           -0.004135   0.003742  -1.105 0.269086
clienttypeClient   0.479826   0.319386   1.502 0.133009
clienttypeEmployee 0.275121   0.411532   0.669 0.503796
clienttypeNormal   0.690322   0.242271   2.849 0.004380 **
clienttypeSalary   0.050843   0.246264   0.206 0.836434
log(1 + max_outbody) 0.142916   0.140520   1.017 0.309129
log(1 + max_ovdbody) 0.380554   0.186531   2.040 0.041334 *
log(1 + max_ovd)  -0.023716   0.183434  -0.129 0.897131
log(1 + max_limit) -0.333863   0.087932  -3.797 0.000147 ***
log(1 + avg_montake) 0.367150   0.197322   1.861 0.062792 .
log(1 + avg_monpay) -0.556334   0.065413  -8.505  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Анализ модели по площади под ROC-кривой (AUC) на рис. 2 показывает, что она неплохо справляется с задачей распознавания дефолтных и недефолтных случаев: AUC = 0,804 — приемлемый по точности результат. Однако, несмотря на то, что для построения такой модели пришлось отбросить довольно много записей, а также то, что в процессе агрегации осталось значительное их количество, нельзя точно сказать, насколько этот показатель адекватен.

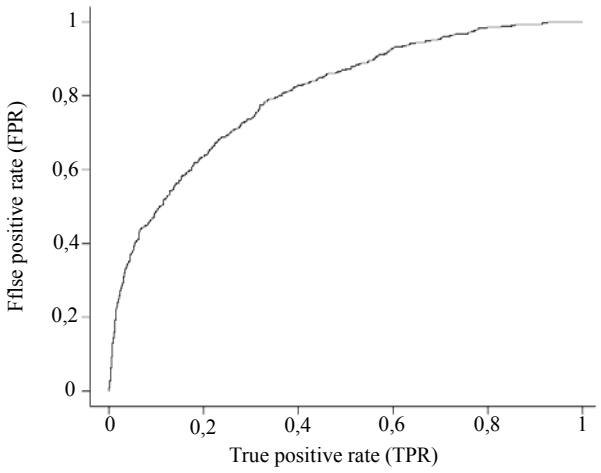

Рис. 2

Пропорциональные риски Кокса

Предварительный анализ данных для построения динамического поведенческого скоринга с помощью модели PH свидетельствует о том, что они согласованны и не содержат противоречий. Приведем несколько описательных характеристик основных переменных, используемых при построении модели:

- средний возраст КК — 9,113 месяцев, максимальный — 31 месяц;
- максимальная задолженность по телу кредита — 100 000 грн.; среднее значение такой задолженности — 4 105,5 грн.;
- среднее значение просроченной задолженности — 171 грн.; максимальное — 51 474 грн.;
- наибольшая величина лимита — 250 000 грн.; среднее значение — 6 884 грн.;
- в среднем ежемесячно клиенты пользовались лимитом в 942,9 грн.; при этом с их стороны в среднем поступало меньше — 740,2 грн., что свидетельствует о тенденции клиентов к просрочке в данной выборке;
- средний возраст владельцев КК — 38,67 лет, медиана — 37 лет; максимальный — 66 лет; минимальный — 20 лет;
- средний доход — 3 827 грн.; максимальный — 120 000 грн.; минимальный — 0 грн.

Построение модели базируется на формуле (2) PH. Оценивание коэффициентов модели происходило путем максимизации частичной функции правдоподобия (6). Для этого применялась функция coxph библиотеки survival.

Отметим, поскольку было принято решение не использовать категоризацию переменных, а считать непрерывные величины непрерывными, для сглаживания влияния больших значений проводилось логарифмирование соответствующих полей, значения которых достигали третьего порядка и выше.

Из-за чрезмерной корреляции многих показателей (например, общая просрочка состоит из просрочки по процентам, по комиссии и по телу; по своей сути она является линейной комбинацией этих трех величин) для построения модели выбраны наиболее значимые показатели. Таким образом, в модель включен 21 регрессор:

$$\ln\lambda(t, x(t), \beta) = \sum_{i=1}^{21} \beta_i x_i . \qquad (9)$$

Оцененные коэффициенты модели, т.е. вектор в формуле (2), экспонента от него, оценки стандартных отклонений, *Z*-статистика и *p*-значение приведены в модели логистической регрессии.

Предварительно можно сказать, что наибольшее влияние имеет параметр pdd_new, что соответствует месяцам просрочки, и log (1 + ovd), что является логарифмом от величины просроченной задолженности. Такой результат прогнозируем, поскольку целевое поле формировалось именно на основе этих двух значений.

Однако стоит отметить довольно неадекватные значения в последнем столбце табл. 3, что указывает на необходимость более детального исследования модели. В связи с этим выбран метод, описанный в работе [6], который заключается в отборе определенного количества наибольших оценок функции риска и соответствует количеству фактических случаев дефолта. На основе такого отбора строится ROC-кривая и оценки AUC (рис. 3).

При дальнейшем исследовании подобное поведение очевидно. Поскольку в регрессоры включены такие поля, как количество дней просрочки и ее величина (значительно коррелируют с целевым полем, ведь оно формируется на основе значений именно этих двух полей), то полученный результат прогнозируемый. Несмотря на то, что построенную модель пока нельзя применить, такое поведение модели свидетельствует о корректности подхода.

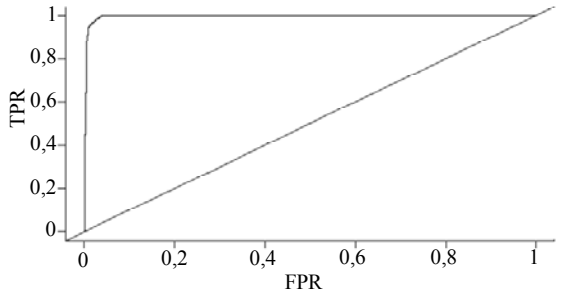

Рис. 3

Для того чтобы сделать модель более адекватной, а также одновременно обеспечить ее предикативность, логичным решением является применение лагов, т.е. значений регрессоров, смещенных во времени [9]. Построение таких моделей рассмотрим ниже.

Прогнозирующие модели со смещенными во времени значениями

В результате анализа модели решено применять смещенные значения. Отметим, что сравнение по тесту Вальда (WT) [9] и тесту множителей Лагранжа (LM) указывает на то, что параметр pdd_new, который соответствует «количеству месяцев просрочки», малозначительный, его можно исключить. Поэтому входные данные остались теми же, за исключением нескольких изменений:

— параметр «количество месяцев просрочки» было решено не применять из-за чрезмерной корреляции с другими показателями и низкой значимостью;

— добавились значения переменных: остаток по кредиту, просрочка, количество снятых и возвращенных на карточный счет денег с лагами 1–3.

После получения обновленных данных построено три модели в соответствии с применением регрессоров с лагами 1–3:

$$h_k(t, x(t-k), \beta) = \sum_{i=1}^{5} \beta_i x_i(t-k) + \sum_{i=6}^{21} \beta_i x_i \, . \qquad (10)$$

В правой части уравнения первое слагаемое — динамические параметры, а второе слагаемое — статические параметры.

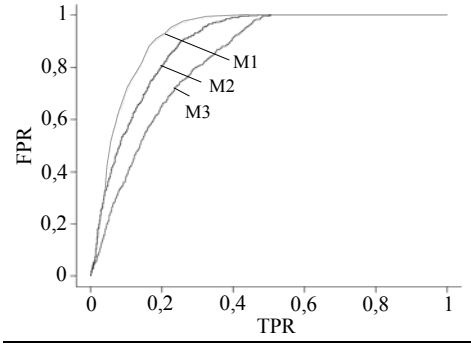

Рис. 4

Для сравнения моделей между собой рассчитано значение AUC и построены ROC-кривые (рис. 4, M1–M3). Такие методы сравнения общеприняты и обеспечивают оценку возможностей классификации по каждой модели.

Как и следовало ожидать, возможности моделей распознавать «плохих» ухудшаются с увеличением лаговости регрессоров. Однако при этом улучшаются их предикативные свойства (табл. 2).

Таблица 2

Модель с лагом	M1	M2	M3
AUC	0,918881	0,887751	0,828898
GINI	0,8378	0,7755	0,6578
AIC	7495,905	7819,972	8946,005
WT	1410,135	1384,005	1308,561

Сравнение различных статистик моделей показывает, что несмотря на лучшее качество модели M1, использование большего количества регрессоров может быть более оправданным, поскольку они более значимые.

Непараметрическая регрессия и оценка Каплан–Мейера

Оценка Каплан–Мейера (KM) является в определенной степени обобщением эмпирической функции выживания и учитывает цензурированные наблюдения. Формула KM для вероятности выживания в определенное время ограничивается произведением характеристик, соответствующих лицам, которые остались в живых после времени t_i. Поэтому часто такую оценку также называют Product-limit estimator [10].

Для расчета оценки KM все наблюдения сортируются в порядке возрастания времени их жизни. Первое вхождение начинается в нуле. Вероятность выживания к этому времени равна 1. Дальнейшие наблюдения исключаются в момент времени их гибели (возможно, в результате цензуры). Множество под риском, которое обозначается $R(t_i)$, — это количество всех индивидов, доживших хотя бы до времени t_i.

Основная идея расчета представляется формулой

$$\hat{S}(t_j) = \prod_{i=1}^{j} \hat{P}(T > t_i \mid T \geq t_i) \tag{11}$$

с учетом того, что

$$\hat{P}(T > t_i \mid T \geq t_i) = \frac{n_i - d_i}{n_i}, \tag{12}$$

где n_i — количество наблюдений во множестве риска $(n_i = \mid R(t_i) \mid)$; d_i — количество субъектов, которые погибли в момент t_i $(d_i = \mid D_i \mid)$.

ISSN 0572-2691

Статистика КМ позволяет исследовать популяцию в общем и проанализировать, какая ее часть остается живой до определенного момента времени, поскольку позволяет оценить вероятность такого выживания.

На основе отобранной выборки построена оценка КМ (рис. 5). Однако такая модель, не позволяет оценивать отдельно каждого клиента, поскольку рассматривает в качестве субъекта всю популяцию. Тем не менее она позволяет качественно оценивать кредитный портфель по разным типам кредитов.

Поскольку в выборке были задействованы кредиты разных лет, целесообразно выполнить срез в соответствии с датами выдачи, сравнивая модели за разные годы на основании оценок КМ.

Как видно из рис. 6–8, поток кредитов в 2015 г. в чем-то повторяет 2013 г. При этом добросовестное поведение было у клиентов, которым выданы КК в 2014 г. После уточнения особенностей, связанных с внутренними настройками политики банка, определено, что в 2014 г. был поднят скоринговый балл для клиентов, которым выдавали кредит, ужесточена скоринговая карта и соответственно КК выдавались самым добросовестным (в соответствии со скоринговой моделью) клиентам банка. Эти условия были ослаблены в 2015 г., что и подтвердила оценка КМ. Неоднородность выборки свидетельствует о необходимости дальнейшей стратификации выборки и построении моделей.

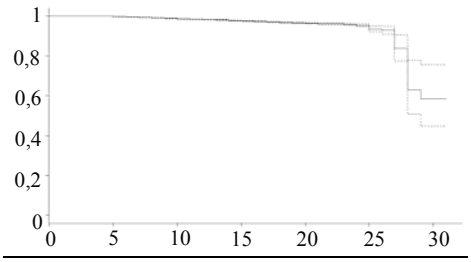

Рис. 5

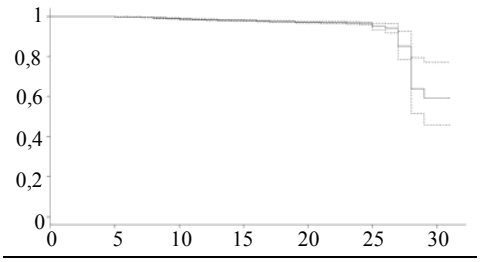

Рис. 6

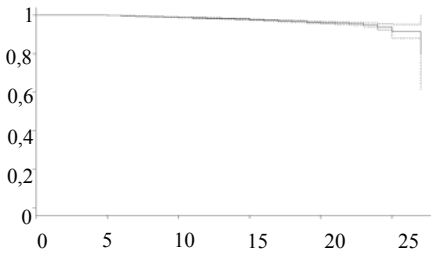

Рис. 7

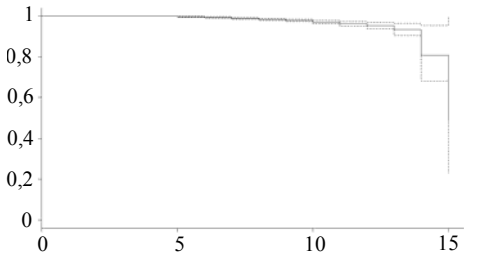

Рис. 8

Анализ полученных результатов

Выполнено сравнение всех описанных моделей, построенных методами анализа выживания и с помощью логистической регрессии (рис. 9, M1–M3 и Mlog). Из рис. 9 видно, что модели с лагами 1 и 3 значительно лучше классифицируют клиентов. При этом модель M3 хуже справляется с задачей в нижних сегментах, но показывает значительно лучшие результаты в верхних сегментах. Это хорошо для банковского сектора, поскольку нужно выбрать оптимальный порог отсечения, правее которого будут находиться клиенты, которым можно выдавать КК. Цель банков — минимизация количества «плохих» клиентов в выборке, а это достигается благодаря выпуклости в верхнем сегменте модели M3, которая позволяет

отобрать больше «хороших» клиентов. В свою очередь, логит-модель лучше классифицирует «плохих» клиентов, т.е. ее целесообразно применять для скоринга мошенничества.

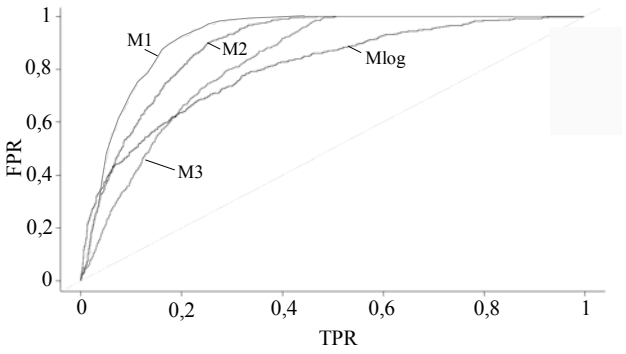

Рис. 9

Итак, можно сделать вывод, что методы теории выживания целесообразно применять для построения моделей определения и классификации «хороших» клиентов. При этом такие модели естественным образом обеспечивают прогнозирование благодаря использованию при построении смещенных во времени значений переменных (лагов). Также нужно отметить возможность скоринга портфелей КК на основе статистики КМ, что дает возможность сравнивать поведение популяций в общем, как это показано в статье. В результате такого анализа выявлено, что выборки 2013 и 2014 гг. значительно отличаются, что наталкивает на мысль о возможности стратификации данных и рассмотрения их по отдельности.

Таблица 3

МОДЕЛИ	AUC	AIC
M1	0,919	8946,005
M2	0,888	7819,972
M3	0,829	7495,905
Logit	0,804	2334,007

Заключение

Анализ поведения клиентов банка — важный аспект управления рисками. Кроме того, что своевременное предвидение дефолта заемщика может сохранить прибыль и устойчивое развитие финансового учреждения, такой анализ также необходим для расчета резервов, соответсвует требованиям международных стандартов и является обязательным на уровне действующего законодательства. Фактически подход к формированию капиталовложений основывается на оценке вероятности потери платежеспособности клиента. Это еще раз подтверждает актуальность скоринговых моделей, которые позволяют сравнивать клиентов между собой.

Особое внимание привлекает такой вид кредита, как кредитная карта. В связи с ее динамичным характером возникает проблема выявления каких-то закономерностей и своевременного реагирования на изменения в поведении владельца этого платежного средства. Наиболее подходящий инструмент для решения этой задачи — поведенческий скоринг, а самая распространенная модель для построения скоринговой карты — логистическая регрессия. Однако, как показывает практика, этот подход не дает желаемых результатов. Во-первых, такая

модель статическая, во-вторых, ее трудно применять для прогнозирования. Поэтому предложено рассмотреть альтернативную методологию, основанную на приемах анализа выживания.

В данной работе описаны основные принципы теории выживания. Введены такие понятия, как функция риска, модель PH и статистика KM, т.е. предложен математический аппарат для построения модели. Оказывается, что пропорциональные риски Кокса позволяют включать во множество регрессоров переменные, зависящие от времени. Использование такого функционала способствовало применению значения переменных, смещенных во времени, естественным образом обеспечивая прогнозирование. Для проведения вычислительного эксперимента использована выборка, состоящая из 376789 записей по 30000 КК, выданным в 2013–2016 гг. Однако для построения моделей использованы аппликационные данные. В связи с этим количество доступных записей уменьшилось. Для модели PH выборка состоит из 55286 наблюдений по 4037 КК с ежемесячной детализацией. При этом следует отметить, что для построения модели на основе логистической регрессии происходило агрегирование и отбор так называемых «зрелых» карт, что привело к уменьшению количества строк в выборке и к определенной потере информации.

Из сравнения полученных результатов следует, что возможность классификации моделей PH уменьшается при увеличении лаговости ковариант, однако даже при применении третьего лага такая модель показывает лучшие результаты, чем обычная логистическая регрессия.

В статье описана последовательность построения моделей оценки клиентов методами логистической регрессии и анализа выживаемости, отбор параметров и сравнение промежуточных результатов. Предложены рекомендации по улучшению предсказуемых качеств моделей на основе методов теории выживания и перспективы дальнейшего их развития для других видов финансовых рисков. Также стоит обратить внимание на потенциальные возможности улучшения построенных моделей путем более детального анализа независимых переменных и формирования различных целевых полей.

Применение такого динамического и поведенческого оценивания клиентов и кредитов с помощью моделей анализа выживаемости позволит банкам своевременно реагировать и существенно снижать потери из-за дефолтов.

Н.В. Кузнєцова, П.І. Бідюк

МОДЕЛЮВАННЯ КРЕДИТНИХ РИЗИКІВ НА ОСНОВІ ТЕОРІЇ ВИЖИВАННЯ

Описано основні принципи теорії аналізу виживання, покроково розписано послідовність побудови моделі оцінки клієнтів методами логістичної регресії і аналізу виживання. Введено такі поняття, як функція ризику, модель пропорційних ризиків Кокса і статистика Каплан–Мейера. Проведено експериментальні дослідження, які показали доцільність використання запропонованих моделей для вирішення завдань поведінкового скорингу, оскільки пропорційні ризики Кокса дозволяють включати до множини регресорів змінні, що залежать від часу. Дано рекомендації щодо поліпшення якостей моделей, а також окреслено перспективи подальшого застосування моделей пропорційних ризиків для інших видів фінансових ризиків, де також необхідно оцінювати цілу групу (популяцію) в часі.

N.V. Kuznetsova, P.I. Bidyuk

MODELING OF CREDIT RISKS ON THE BASIS OF THE THEORY OF SURVIVAL

The basic principles of the theory of survival analysis are described, step by step the construction of models of assessment of the clients by the methods of logistic regression and survival analysis are shown. The following concepts as a function of risk, the Cox proportional hazard model and the Kaplan-Meier statistics are introduced. Experimental studies have been carried out. They have shown the expediency of using the proposed models for solving the problems of behavioural scoring, since Cox's proportional risks allow the inclusion of a set of regressors with variables that depend on time. Suggested recommendations for improving the predictive qualities of models to overcome the heterogeneity of the sample, in particular the further stratification of the sample, and outlined the prospects for further development of proportional risk models for other financial risks, where it is also necessary to estimate the whole group (population) in time.

1. *Siddiqi N.* Credit risk scorecards: developing and implementing intelligent credit scoring. — Cary, North Carolina, USA. — 2005. — 196 p.
2. *Cox D.R., Society S.B.* Regression models and life-tables // Methodological. — 2007. — **34**, N 2. — P. 187–220.
3. *Cao R., Vilar J.M., Devia A.* Modelling consumer credit risk via survival analysis // SORT. — 2009. — **33, N 1. —** P. 3–30.
4. *Marimo M.* Survival analysis of bank loans and credit risk prognosis master of science mathematical statistics. — http://wiredspace.wits.ac.za/jspui/bitstream/10539/18597/1/Mercy%20Marimo%20Thesis_Survival%20Analysis_28.03.%202015_v1.pdf.
5. *Basel II*: International convergence of capital measurement and capital standards: a Revised Framework. — http://www.bis.org/publ/bcbs54.htm .
6. *Stepanova M., Thomas L.C.* Survival analysis methods for personal loan data // Operations Research. — 2002. — **50**, N 2. — P. 277–289.
7. *Fleming, T.R., Harrington D.P.* Counting processes and survival analysis. — New York. : John Wiley & Sons — 1991.
8. *Фомін О.В., Кузнєцова Н.В.* Скорингові моделі поведінки клієнтів-власників кредитних карток для оцінки їх платоспроможності // Системні науки та кібернетика. — 2016. — № 5. — С. 56–67. — http://mmsa.kpi.ua/sites/default/files/ssc/issues/ssc_5_2016.pdf.
9. *Бидюк П. И., Романенко В. Д., Тимощук О. Л.* Анализ временных рядов. — Киев: Политехника, 2013. — 600 с.
10. *Dabrowska D.* Non-parametric regression with censored survival time data // Scandinavian Journal of Statistics. — 1987. — **14**, N 3. — P. 181–197.

Получено 31.05.2017