УДК 519.234

АНАЛИЗ И ПРОГНОЗИРОВАНИЕ КРЕДИТНЫХ ПОТЕРЬ С ПОМОЩЬЮ ДИСКРЕТНЫХ МОДЕЛЕЙ ВЫЖИВАЕМОСТИ

М.М. Васьковский
Доцент кафедры высшей математики БГУ, кандидат физико-математических наук, доцент; системный аналитик СООО "ХайКво Солюшенс"

А.О. Задорожнюк
Студентка БГУ

Кафедра высшей математики Белорусского государственного университета, Республика Беларусь
СООО «ХайКво Солюшенс», Республика Беларусь
E-mail: vaskovskii@bsu.by

М.М. Васьковский
Доцент кафедры высшей математики Белорусского государственного университета, кандидат физико-математических наук, доцент; системный аналитик СООО «ХайКво Солюшенс». Является специалистом в области стохастического анализа.

А.О. Задорожнюк
Студентка Белорусского государственного университета. Проводит научные исследования в рамках алгебраической теории графов.

Аннотация. В работе рассматривается применение дискретных моделей выживаемости для моделирования ожидаемых кредитных потерь на примере ипотечных кредитов США, выданных на фиксированный срок при фиксированной процентной ставке. В основе рассматриваемой модели лежат Age-Period-Cohort декомпозиции условных вероятностей, определяющих текущее состояние кредитного аккаунта. Важную роль при построении модели играет экономическое моделирование на основе использования реальных экономических данных по ключевых макроэкономическим факторам. Описываемый подход учитывает также индивидуальные характеристики кредитных аккаунтов посредством построения скоринговых моделей для условных вероятностей состояний кредитных аккаунтов.

Ключевые слова: Ожидаемые кредитные потери, логистическая регрессия, APC-декомпозиция, уравнение Орнштейна-Уленбека.

Современные модели для прогнозирования кредитных рисков, как правило, строятся с использованием статистических методов и методов теории случайных процессов.

Рассмотрим так называемую модель выживаемости дискретного времени (Discrete Time Survival Model) моделирования текущих ожидаемых кредитных потерь (Current Expected Credit Loss), описанную в [1, 2]. Не вдаваясь в технические детали, приведем основную идею модели DTSM на примере ипотечных кредитов США, выданных на фиксированный срок (как правило, до 30 лет). Нижеприведённый анализ данных осуществлялся на реальных данных по ипотечным кредитам США (предоставляемых агентствами Fannie Mae

и Freddi Mac), начиная с 2001 года. С целью построения более точной модели данные сегментировались по кредитному рейтингу FICO Score: сегмент Subprime соответствует значениям рейтинга ниже 660, сегмент Prime – значениям от 660 до 780 и сегмент Superprime – значениям свыше 780. В дальнейшем все численные результаты анализа будут приводиться на примере сегмента Subprime.

Введем обозначения и ключевые переменные. Через a, t, v будем обозначать соответственно возраст кредита (в месяцах), текущую дату наблюдения и дату открытия кредита, предполагая, что переменные a, t, v принимают конечное число значений.

Будем рассматривать вероятности дефолта (Probability of Default) и закрытия кредита (Probability of Attrition) как функции от a, t, а именно:

$$PD(a,t) = \frac{N_{def}(a,t)}{N_{act}(a,t-1)},$$

$$PA(a,t) = \frac{N_{attr}(a,t)}{N_{act}(a,t-1)},$$

где N_{act}, N_{def}, N_{attr} – числа активных, дефолтных и закрытых аккаунтов соответственно.

Также определим среднюю величину дефолтного баланса (Exposure at Default):

$$EAD(a,t) = \frac{B_{def}(a,t)}{N_{def}(a,t)},$$

где B_{def} – суммарный дефолтный баланс.

На первом этапе анализа для каждой из переменных PD, PA, EAD вычисляется декомпозиция (Age-Period-Cohort):

$$logit\left(PD(a,t)\right) = F_{PD}(a) + H_{PD}(t) + G_{PD}(v),$$

$$logit\left(PA(a,t)\right) = F_{PA}(a) + H_{PA}(t) + G_{PA}(v),$$

$$log\left(EAD(a,t)\right) = F_{EAD}(a) + H_{EAD}(t) + G_{EAD}(v),$$

где $logit(p) = \log\left(\frac{p}{1-p}\right)$ – обратная к логистической функции $g(z) = \frac{1}{1+e^{-z}}$, $z \in \mathrm{R}$.

В декомпозициях переменных PD, PA, EAD функции $F(a)$ являются жизненными циклами (Lifecycle), функции $H(t)$ (Environment) выражают влияние макроэкономических факторов (уровень безработицы, ВВП, доход на душу населения и др.), функции $G(v)$ (Credit quality) включают в себя влияние внутренних факторов неэкономического характера.

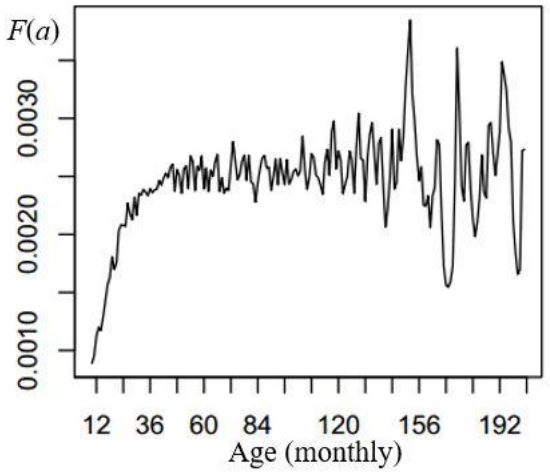

Рисунок 1a. Lifecycle, *PD*

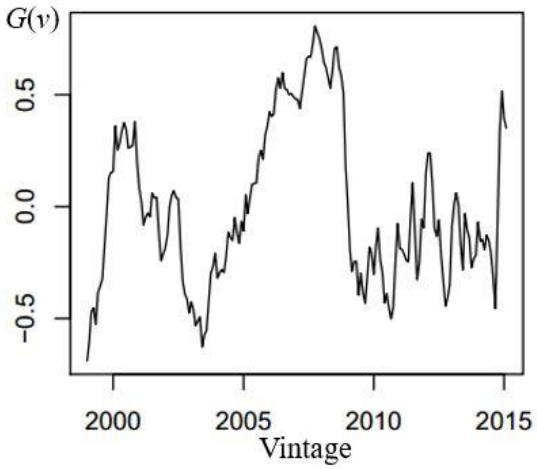

Рисунок 1b. Credit Quality, *PD*

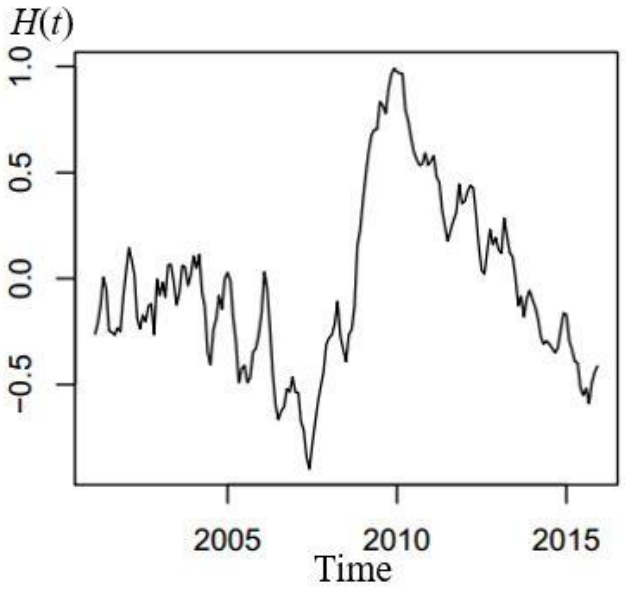

Рисунок 1c. Environment, *PD*

Второй этап анализа заключается в построении экономических моделей для функций $H(t)$. Экономические модели строятся на основе данных по основным макроэкономическим переменным:

$$H(t) = \beta_0 + \sum_{i=1}^{n} \beta_i f_i(t) + \varepsilon(t),$$

где $f_i(t)$ $(i = 1, \ldots, n)$ – некоторый набор макроэкономических факторов, β_i – искомые коэффициенты, $\varepsilon(t)$ – погрешность. Например, для PD получается следующий результат:

Таблица 1.

Лучшее приближение $H_{PD}(t)$.

Переменная	Преобразование, f_i	Оптимальное значение lag	Оптимальное значение win	Коэффициент, β_i
Real DPI	LogRatio	2	23	–1.734
Unemployment rate	Diff	5	24	0.078
House Price Index	LogRatio	12	17	–2.678
10-year Treasury yield	Diff	7	24	–0.106

Преобразования LogRatio и Diff задаются следующими формулами:

$$LogRatio\left(x,t,lag,win\right)=\log\frac{x(t-lag)}{x(t-lag-win)},$$

$$Diff\left(x,t,lag,win\right)=x\left(t-lag\right)-x\left(t-lag-win\right).$$

Проверялись комбинации от 1 до 4 макроэкономических факторов и преобразований из следующего списка: Real GDP (LogRatio), Real DPI (LogRatio), Unemployment rate (Diff, LogRatio), House Price Index (LogRatio), 5-year Treasury yield (Diff), 10-year Treasury yield (Diff), BBB corporate yield (Diff), Commercial Real Estate Price Index (LogRatio), CPI inflation rate (Diff), Market Volatility Index (LogRatio), Dow Jones Total Stock Market Index (LogRatio) со значениями lag от 0 до 12 и win от 1 до 24 [1].

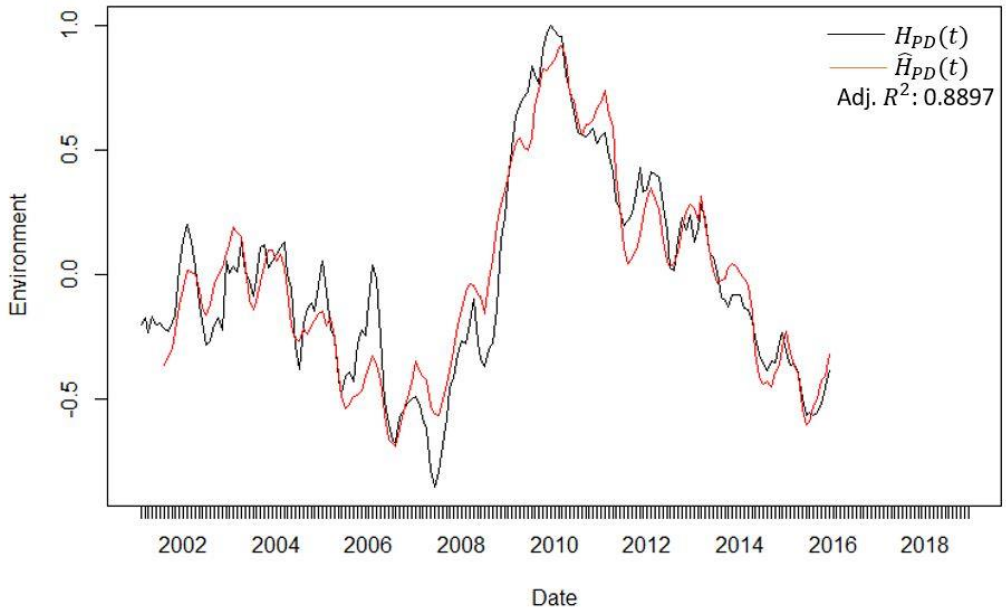

Рисунок 2. Сравнение $H_{PD}(t)$ и его приближения

Третий этап анализа заключается в построении так называемых скоринговых моделей для переменных PD, PA, EAD. Основная идея заключается в том, что при моделировании соответствующих переменных функция $H(t)$ заменяется ее экономическим приближением

$$\hat{H}(t) = \beta_0 + \sum_{i=1}^{n} \beta_i f_i(t), \tag{1}$$

а функция кредитного риска $G(v)$ заменяется индивидуальными характеристиками кредитных аккаунтов – скоринговыми факторами (персональный кредитный рейтинг –FICO Score, процентная ставка по кредиту, наличие текущей задолженности и др.). Продемонстрируем сказанное на примере переменной PD. Для каждого значения $h = 1, 2, \ldots$ получаем следующее представление:

$$\begin{gathered} \operatorname{logit}\left(PD_i(t,v,h)\right) = \\ = F_{PD}(t-v) + \hat{H}_{PD}(t) + c_0(h) + \sum_j c_j(h) X_{j,i}(t-h) + \varepsilon_i(h), \end{gathered} \tag{2}$$

где $PD_i(t,v,h)$ есть условная вероятность того, что аккаунт с номером i (открытый в момент v) станет дефолтом в момент времени t при условии, что он был активным в момент времени $t-1$, а скоринговый фактор X_j в момент времени $t-h$ принимал значение $X_{i,j}(t-h)$; коэффициенты $c_j(h)$ есть искомые скоринговые коэффициенты; $\varepsilon_i(h)$ – погрешность.

Ниже показана зависимость скоринговых коэффициентов для переменной Delinquency (времени задержки платежа в месяцах) от горизонта $h = 1, \ldots, 12$.

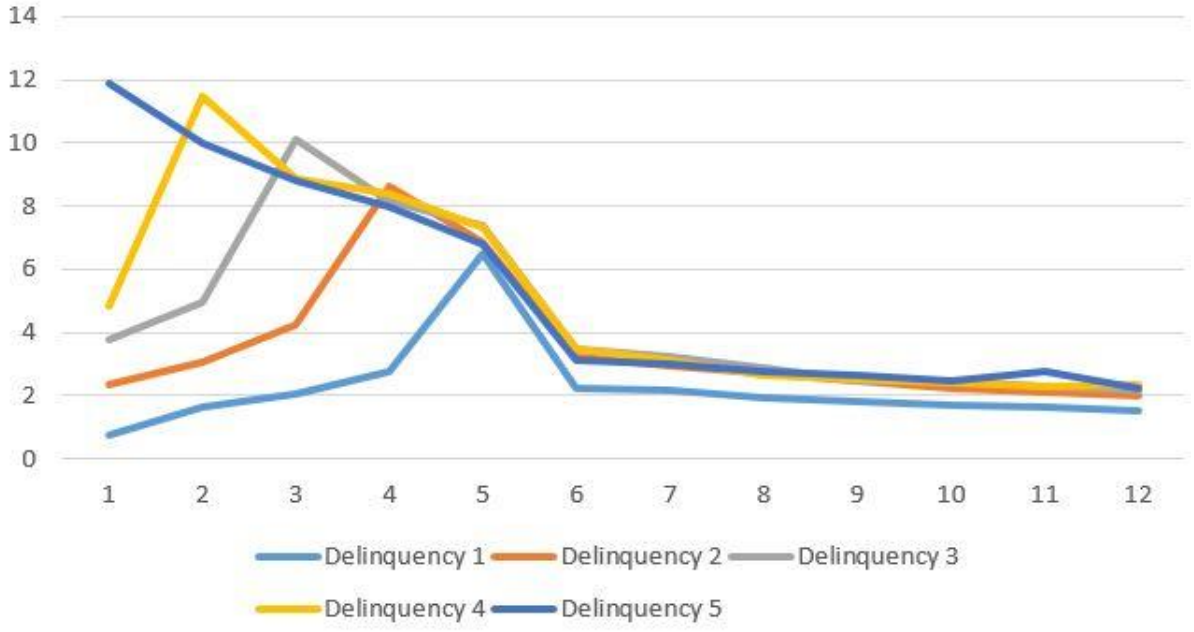

Рисунок 3. Скоринговые коэффициенты в зависимости от горизонта

Покажем, что соотношения (2) могут быть использованы для прогнозирования условной вероятности дефолта в будущем. Предположим, что нам известны значения скоринговых факторов X_j активного аккаунта с номером i в настоящий момент t_0. Тогда, согласно построенной модели, величина

$$PD_i\left(t,v,t-t_0\right)=g\left(F_{PD}\left(t-v\right)+\hat{H}_{PD}\left(t\right)+c_0\left(t-t_0\right)+\sum_j c_j\left(t-t_0\right)X_{j,i}\left(t_0\right)\right)$$

дает приближение к значению условной вероятности того, что аккаунт с номером i станет дефолтом в момент времени $t > t_0$.

Аналогично можно построить скоринговую модель и соответствующий прогноз для условной вероятности $PA_i(t,v,t-t_0)$ закрытия аккаунта с номером i в момент времени t.

Для переменной EAD скоринговая модель и соответствующий прогноз примут следующий вид:

$$\log\left(EAD_i\left(t,v,h\right)\right)=$$
$$F_{EAD}\left(t-v\right)+\hat{H}_{EAD}\left(t\right)+\tilde{c}_0\left(h\right)+\sum_j\tilde{c}_j\left(h\right)\tilde{X}_{i,j}\left(t-h\right)+\varepsilon_i\left(h\right),$$

$$EAD_i\left(t,v,t-t_0\right)=\exp\left(F_{EAD}\left(t-v\right)+\hat{H}_{EAD}\left(t\right)+\tilde{c}_0\left(t-t_0\right)+\sum_j\tilde{c}_j\left(t-t_0\right)\tilde{X}_{i,j}\left(t_0\right)\right),$$

где $\tilde{X}_j$, $\tilde{c}_j(h)$ – соответственно скоринговые факторы и скоринговые коэффициенты для переменной *EAD*.

Таким образом, мы можем вычислить количество активных, дефолтных и закрытых аккаунтов на момент t по следующим рекуррентным формулам:

$$\tilde{N}_{def}\left(t,v\right)=\sum_i\left(PD_i(t,v,t-t_0)\prod_{h=1}^{t-t_0-1}\left(1-PD_i\left(t,v,h\right)-PA_i(t,v,h)\right)\right),$$

$$\tilde{N}_{attr}\left(t,v\right)=\sum_i\left(PA_i(t,v,t-t_0)\prod_{h=1}^{t-t_0-1}\left(1-PD_i\left(t,v,h\right)-PA_i(t,v,h)\right)\right),$$

$$\tilde{N}_{act}\left(t,v\right)=\tilde{N}_{act}\left(t-1,v\right)-\tilde{N}_{def}\left(t,v\right)-\tilde{N}_{attr}\left(t,v\right).$$

Величину ожидаемых кредитных потерь для аккаунтов, активных на данный момент t_0 можно определить следующим образом:

$$CECL\left(t_0\right)=\sum_{t>t_0}\sum_v\sum_i\left(EAD_i\left(t,v,t-t_0\right)PD_i\left(t,v,t-t_0\right)\prod_{h=1}^{t-t_0-1}\left(1-PD_i\left(t,v,h\right)-PA_i(t,v,h)\right)\right)$$

Таким образом, модель DTSM позволяет строить оценку текущих ожидаемых кредитных потерь в предположении, что известна экстраполяция функций $\hat{H}_{PD}(t)$, $\hat{H}_{PA}(t)$, $\hat{H}_{EAD}(t)$ на весь период существования аккаунтов, активных на текущую дату t_0.

Для нахождения экстраполяций функций $\hat{H}(t)$ общепринятым является следующий метод. Используя ежегодный краткосрочный прогноз по макроэкономическим показателям, выпускаемый Федеральной Резервной Системой, осуществляется экстраполяция функции $\hat{H}(t)$ на двухлетний период, начиная с момента t_0, используя известные коэффициенты β_i, входящие в правую часть соотношения (1). Для получения дальнейшей экстраполяции на период существования аккаунтов, активных на момент t_0, используется идея регрессии к среднему на основе модели Васичека: искомая экстраполяция функции $\hat{H}(t)$ задается на основе одномерного уравнения Орнштейна-Уленбека [3]

$$dx(t) = \theta\left(\mu - x(t)\right)dt + \sigma dW(t), \tag{3}$$

где μ – долгосрочное среднее, $\theta > 0$ – параметр, характеризующий скорость возврата к среднему значению, $\sigma > 0$ – параметр волатильности, $W(t)$ – винеровский процесс.

Решение уравнения (3) с начальным условием $x(\tau) = x_0$ задается соотношением

$$x(t) = x_0 e^{-\theta(t-\tau)} + \mu\left(1 - e^{-\theta(t-\tau)}\right) + \sigma\int_{\tau}^{t} e^{-\theta(t-s)} dW(s),$$

а в качестве экстраполирующей кривой выбирается математическое ожидание решения

$$E\left(x(t)\right) = x_0 e^{-\theta(t-\tau)} + \mu\left(1 - e^{-\theta(t-\tau)}\right),$$

где $\tau = t_0 + 24$ – момент начала долгосрочной экстраполяции.

Проблема идентификации компонент APC-декомпозиции

Так как дискретные переменные a, v, t линейно зависимы, то матрица X скоринговых факторов соответствующей регрессионной модели APC-декомпозиции

$$Y = XB + \varepsilon,\ Y \in \mathrm{R}^{m \times n},$$

является вырожденной, т. е. $\det\left(X^T X\right) = 0$. В этом случае метод наименьших квадратов оценки коэффициентов регрессии неприменим, так как задача минимизации

$$\arg\min_{B} |\, XB - Y\,|_2^2 \ ,$$

не является однозначно разрешимой.

Среди методов, связанных с регуляризацией некорректной задачи, можно выделить следующие:

1) гребневую регрессию (Ridge Regression), дающую оценку

$$\hat{B}_{RR} = \arg\min_B(|XB-Y|_2^2 + \gamma |B|_2^2) = (A+\gamma E)^{-1} X^T Y,$$

2) регрессию Лассо (Lasso Regression) с оценкой

$$\hat{B}_{LR} = \arg\min_B(|XB-Y|_2^2 + \gamma |B|_1),$$

где $\gamma > 0$ – параметр регуляризации.

Существуют методы, основанные на выделении наиболее "влиятельных" скоринговых факторов, среди которых стоит отметить следующие:

1) метод главных компонент (Principal Components Regression), основанный на нахождении сингулярного разложения матрицы $A = X^T X$, т.е. $A = VD^2S$, где матрицы V, S ортогональные и состоят соответственно из собственных векторов v_i матрицы A^T и собственных векторов s_i матрицы A, D – диагональная матрица, состоящая из положительных собственных значений $\lambda_1, \ldots, \lambda_r$ матрицы A. В качестве оценки выбирается сумма p первых главных компонент:

$$\hat{B}_{PCR} = \sum_{i=1}^{p} \hat{b}_i = \sum_{i=1}^{p} \frac{v_i^T Y}{\lambda_i} s_i,$$

где p – параметр регрессии. В частности, при $p = \mathrm{rank}\, A$ получаем

$$\hat{B}_{PCR} = A^+ X^T Y,$$

где A^+ – псевдообратная матрица Мура-Пенроуза, дающую решение задачи минимизации

$$\arg\min_B |XB-Y|_2^2 = A^+ X^T Y.$$

2) метод наименьших частных квадратов (Partial Least Square Regression), основанный на нахождении матрицы T, состоящей из p линейно независимых векторов t_i из числа линейных комбинаций исходных скоринговых факторов X, так, что вектора t_i имеют наибольшие возможные по модулю коэффициенты корреляции с левой частью Y и уже построенными векторами матрицы T. . На втором этапе применяется метод наименьших квадратов для решения задачи регрессии с невырожденной матрицей T (т.е. $\det(T^T T) \neq 0$.

Метод наименьших частных квадратов дает оценку

$$\hat{B}_{PLSR} = \arg\min_{B \in \mathrm{K}} |XB-Y|_2^2 ,$$

где K – пространство Крылова, т. е. линейное пространство, порожденное векторами z, Az, …, $A^{p-1}z$.

Также известны методы решения проблемы неоднозначной идентификации APC-декомпозиции с помощью симуляций Монте-Карло для цепей Маркова [4].

Замечание 1. Так как переменные PD, PA подчинены биномиальному распределению, то для нахождения декомпозиций для этих переменных необходимо использовать уравнение логистической регрессии.

Замечание 2. Приняты следующие ограничения в отношении функций $H(t), G(v)$:

1) функция $G(v)$ не содержит линейный тренд;

2) функция $H(t)$ не содержит линейный тренд и имеет нулевое среднее.

Литература

[1]. Breeden, J.L. Living with CECL: Mortgage modeling alternatives / J.L. Breeden. – Middletown, 2018.

[2]. Breeden, J.L. Incorporating lifecycle and environment in loan-level forecasts and stress tests / J.L. Breeden // European Journal of Operational Research. 2016. – Vol. 255, № 2. – P. 649 – 658.

[3]. Оксендаль, Б. Стохастические дифференциальные уравнения. Введение в теорию и приложения / Б. Оксендаль. – М.: Мир, 2000. – 400 с.

[4]. Schmid, V. Bayesian age-period-cohort modeling and prediction – BAMP/ V.J. Schmid, L. Held // Journal of Statistical Software. 2007. – Vol. 21, №8 – P. 1-15.

ANALYSIS AND FORECASTING OF CREDIT LOSS BASING ON DISCRETE SURVIVAL MODELS

M.M. VASKOUSKI
Associate professor of the Department of higher mathematics of the Belarusian State University, PhD;system analyst of HiQo Solutions, Ltd

A.O. ZADOROZHNYUK
Student of the Belarusian State University

Department of higher mathematics of the Belarusian State University, Republic of Belarus
HiQo Solutions, Ltd, Republic of Belarus
E-mail: vaskovskii@bsu.by

Abstract. We consider an application of discrete time survival models to the US mortgage credit loss modeling corresponding to credits with the fixed term and interest rate. The core of the model are the Age-Period-Cohort decompositions of the conditional probabilities of the accounts statuses. Economic modeling based on using of the real macroeconomic factors plays a significant role. The described approach takes into account individual account characteristics, which are involved into score models for the conditional probabilities of the accounts statuses.

Keywords: Current expected credit loss, logistic regression, APC-decomposition, Ornstein-Uhlenbeck equation.