Інформаційні технології аналізу клієнтської бази абонентів та прогнозування їх поведінки

© Кузнєцова Н.В.
Інститут прикладного системного аналізу Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського»,
Київ, Україна
natalia-kpi@ukr.net

Анотація

У роботі показана можливість застосування інформаційних технологій для аналізу бази даних абонентів телекомунікаційної компанії з метою передбачення їх подальшої поведінки. Задача є актуальною не лише з точки зору прогнозування факту зміни абонентом телекомунікаційної компанії і відмовою від використання послуг, а й моменту, коли абонент лише почав над цим замислюватись. У статті вирішують дві задачі: задача класифікації (задача прогнозування можливого відтоку абонентів) та задача передбачення моменту часу, в який ця подія може відбутися. Для задачі класифікації можуть використовуватись різноманітні методи інтелектуального аналізу даних. У статті була побудована узагальнена лінійна модель, яка показала прийнятні предикативні властивості на основі індексу GINI, проте нижчі порівняно з логістичною регресією, нейронними мережами, градієнтним бустингом.

Автором було запропоновано розглядати задачу з точки зору виживання популяції – абонентів телекомунікаційної компанії. У роботі наведено основні теоретичні відомості з аналізу виживання та виконано їх формалізацію для вирішення задачі прогнозування відтоку клієнтів, зокрема з урахуванням їх типу (корпоративний чи приватний клієнт), а також часу настання події. Автором запропоновано розв'язувати задачу прогнозування поведінки клієнтів у часовому просторі для завчасного передбачення фінансових ризиків телекомунікаційної компанії, пов'язаних з недоотриманням прибутку через відтік клієнтів або зайвими витратами на додаткове обладнання, в якому немає потреби. Для цього пропонується прогнозувати період та обсяг можливих втрат і будувати функцію виживання та функцію можливих втрат для моделі пропорційних ризиків Кокса. Вони дозволяють визначати момент часу, в який відбувається перехід від критичного та катастрофічного фінансового ризику. Знання періоду настання ризику буде корисним для телекомунікаційної компанії з точки зору запобігання відтоку абонентів шляхом розробки персональних пропозицій та проведенням додаткових заохочень для існуючих клієнтів.

Ключові слова: інформаційні технології, фінансові ризики, моделі виживання, телекомунікаційна компанія, пропорційні ризики Кокса.

1 Вступ

Сучасний світ неможливо уявити без мобільних пристроїв, Інтернету, планшетів та комп'ютерів. Сьогодні клієнт, який не користується послугами мобільного зв'язку чи Інтернету, стає майже виключенням. Це скоріш за все клієнти, які змінюють оператора зв'язку або перебувають в Україні досить короткий час і потребують тимчасового тарифного пакету. Телекомунікаційні компанії зосереджують свої зусилля на розробці інформаційних технологій, що використовують сучасні методології інтелектуального аналізу даних та досліджують поведінку клієнтів-користувачів послуг телекомунікаційних компаній. Основною метою є виявлення уподобань клієнтів та утримання їх як абонентів, розробляючи та пропонуючи їм нові послуги та тарифні пакети згідно їх потребам.

2 Постановка задачі

Стандартні підходи та методи дозволяють побудувати математичні моделі, які будуть прогнозувати безпосередньо подію – можливий відтік клієнтів. Метою даного дослідження стало розроблення математичних моделей аналізу клієнтської бази та короткострокове та довгострокове прогнозування поведінки клієнтів за рахунок виявлення часового проміжку та групи абонентів з усієї бази клієнтів, які замислюються найближчим часом (від 1 місяця до 3 місяців) у зміні оператора.

3 Загальні припущення теорії аналізу виживання

Для аналізу даних використовується вибірка (популяція), яка характеризується певними ознаками: по кожному об'єкту відомий результат події (загибель чи виживання). Для цього здійснюється один з видів цензурування (відсікання). Спостереження називаються цензурованими, якщо спостережувана залежна

CEUR-WS.org/Vol-2067/paper8.pdf

змінна представляє момент настання термінальної події, а тривалість дослідження обмежена за часом. Можливі механізми цензурування змінних: фіксоване цензурування (спостереження відбувається протягом фіксованого проміжку часу) та випадкове цензурування (спостереження відбувається протягом проміжку часу, який настає після того часу, коли елементи вибірки пережили певну подію) [1-3].

При розробці математичних моделей враховувались коваріанти, тобто параметри, що характеризують поведінку клієнтів, як статичні параметри – характеристики клієнта, так і динамічні параметри його поведінки (обсяг трафіку, кількість хвилин дзвінків тощо).

Функція виживання визначається як $S(t) = P(T > t)$, а функція ризику $h(t) = \lim_{\Delta \to 0} \frac{P(t < T < t + \Delta \mid T > t)}{\Delta}$, $h(t) = -\frac{\frac{dS(t)}{dt}}{S(t)}$.

Найпростіша функція, яка визначає, що ризик є константою в часі: $h(t) = \lambda$, або що еквівалентно $\log h(t) = \mu$.

Оскільки $S(t) = \exp[-\int_0^t h(u)du]$, то після підстановки та інтегрування отримуємо: $S(t) = e^{-\lambda t}$, а $f(t) = \lambda e^{-\lambda t}$. Це функція щільності ймовірності з відомим експоненційним розподілом з параметром λ. Таким чином, сталий ризик передбачає експоненційний розподіл для часу, поки не наступить подія (або час між подіями) [4-5].

Модель виживання може будуватись з горизонтом часу і ймовірність відтоку клієнта в наступний період PO (probability of outlet) може бути обчислена таким чином [2]:

$$PO(t \mid x) = P(t \le T < t + b \mid T \ge t, X = x) = \frac{P(T < t + b \mid X = x) - P(T \le t \mid X = x)}{P(T \ge t \mid X = x)} = \quad , \qquad (1)$$

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

де t – час спостереження обслуговування клієнту, а x - значення коваріаційного вектору X для цього клієнта, тобто параметри самого клієнта, його тарифного плану та його поведінки.

Для розподілу часу життя можна прийняти узагальнену лінійну модель [7]:

$$P(T \le t \mid X = x) = F_\theta(t \mid x) = g(\theta_0 + \theta_1 t + \theta^{\mathrm{T}} x),$$

де $\theta = (\theta_2, \theta_3, \ldots, \theta_{p+1})^{\mathrm{T}}$ p-вимірний вектор, g - відома функція зв'язку, така як логістична чи пробіт-функція. Таким чином, ця модель характеризує умовний розподіл часу обслуговування абоненту телекомунікаційною компанією T в термінах невідомих параметрів. Як тільки ці параметри будуть оцінені, отримаємо оцінку функції умовного розподілу, $F_{\hat{\theta}}$ і, нарешті, оцінка відтоку клієнта (PO) може бути обчислена шляхом включення цієї оцінки у рівняння (1), тобто

$$\hat{PO}^{GLM}(t \mid x) = \frac{F_{\hat{\theta}}(t + b \mid x) - F_{\hat{\theta}}(t \mid x)}{1 - F_{\hat{\theta}}(t \mid x)} = 1 - \frac{S_{\hat{\theta}}(t + b \mid x)}{S_{\hat{\theta}}(t \mid x)},$$

де $\hat{\theta} = \hat{\theta}^{GML}$ є оцінкою максимальної правдоподібності вектору параметрів.

Розглянемо одновимірний випадок. У такому випадку $\theta = \theta_2$ і умовний розподіл задається моделлю $F(t \mid x) = g(\theta_0 + \theta_1 t + \theta_2 x)$, зі щільністю $f(t \mid x) = \theta_1 g'(\theta_0 + \theta_1 t + \theta_2 x)$. Оскільки зазвичай задана випадкова цензурована справа вибірка, то умовна функція правдоподібності представляє собою добуток членів, що включають умовну щільність, для нецензурованих даних та умовної функції виживання для цензурованих даних:

$$L(Y, X, \theta) = \prod_{i=1}^{n} f(Y_i \mid X_i)^{\delta_i} (1 - F(Y_i \mid X_i))^{1-\delta_i},$$

де Y_i – строк обслуговування і-го клієнту в телекомунікаційній компанії і δ^i є індикатором відтоку для і-го клієнту.

Таким чином, логарифмічна функція правдоподібності визначається [2]:

$$l(\theta) = \ln(L(Y, X, \theta)) = \sum_{i=1}^{n} [\delta_i \ln(f(Y_i \mid X_i)) + (1 - \delta_i) \ln(1 - F(Y_i \mid X_i))] =$$

$$= \sum_{i=1}^{n} [\delta_i \ln(\theta_i g'(\theta_0 + \theta_1 Y_i + \theta_2 X_i)) + (1 - \delta_i) \ln(1 - g(\theta_0 + \theta_1 Y_i + \theta_2 X_i))] =$$

$$= \sum_{i=1}^{n} \delta_i [\ln(\theta_1) + \ln(g'(\theta_0 + \theta_1 Y_i + \theta_2 X_i))] + \sum_{i=1}^{n} (1 - \delta_i) \ln(1 - g(\theta_0 + \theta_1 Y_i + \theta_2 X_i))$$

І, нарешті, оцінка знаходиться як максимізація функції логарифмічної правдоподібності:

$$\hat{\theta}^{GML} = \arg\max_{\theta} l(\theta).$$

4 Моделі пропорційних ризиків Кокса

Відома модель Кокса, запропонована в 1972 році, інтенсивно використовується в самих різних областях, особливо в медицині і страхування, для оцінки умовного ризику захворювання при заданих значеннях вихідних ознак [1-3]. Модель Кокса заснована на припущенні, що функцію ризику можна факторизувати, тобто представити у вигляді добутку двох функцій:

$$h_i(t) = h_0(t) \cdot \psi(X_{i1}, ..., X_{ik}),$$

де $h_0(t)$ – базова функція інтенсивності, що включає фактор часу, але не включає коваріанти, а $\psi(X_{i1}, ...X_{ik})$ – лінійна функція досліджуваних ознак, яка не включає фактор часу.

Досить часто модель записують у наступному вигляді :

$$h_i(t) = h_o(t) \cdot e^{\{\beta_1 X_{i1} + ... + \beta_k X_{ik}\}},$$

$$\ln h_i(t) = \ln h_o(t) + \beta_1 X_{i1} + ... + \beta_k X_{ik},$$

де $\beta_1, ...\beta_k$ – невідомі параметри.

5 Аналіз поведінки клієнтів за допомогою SAS-технологій

Для поставленої задачі моделювання використовували реальні статистичні дані клієнтської бази телекомунікаційної компанії за 2014-2016 роки [6]. Вхідна вибірка складалась з 150 тисячі абонентів та відповідно інформації про їх активність в мережі протягом 15 місяців (увесь 2014 рік та початок 2015 року). Кожний місяць активності абонента описується наступними 10 показниками:

- кількість хвилин вхідних дзвінків (INCOMING);
- кількість хвилин вихідних дзвінків на стаціонарні номери (PSTN) ;
- кількість хвилин вихідних дзвінків на мобільні номери інших операторів (ALIEN);
- кількість хвилин вихідних дзвінків на мобільні номери цього ж оператора в одному регіоні (REGION);
- кількість хвилин вихідних дзвінків на мобільні номери цього ж оператора в інший регіон (AREA);
- кількість хвилин вихідних дзвінків на мобільні номери інших мобільних операторів (OMO_MINS);
- кількість хвилин вихідних дзвінків на мобільні номери всередині мережі (ONNET_MINS);
- кількість хвилин вихідних дзвінків на міжнародні номери (INTERN_MINS);
- кількість мегабайт спожитого інтернет трафіку (GPRS_USG_MB);
- кількість надісланих СМС (SMS);

Також відома інформація щодо дати активації абонента (Oblast_Activated), його статі (SEX), віку (AGE) та індикатору, чи є він корпоративним клієнтом (COMPANY), і моделі пристрою зв'язку (мобільний телефон, планшет тощо).

Були використані можливості інформаційних технологій SAS Enterprise Miner для побудови і оцінювання параметрів та якості узагальненої лінійної моделі, написані власні коди на SAS Base для побудови моделей виживання з відповідними розподілами (моделі пропорційних ризиків Кокса та напівпараметричної моделі). Отримані коефіцієнти узагальненої лінійної моделі наведені у таблиці 1.

Після цього модель була оцінена на основі ROC-кривої та індексу GINI (рис.1).

За площею під ROC-кривою (AUC=0,8246) обраховуємо значення індексу GINI: $GINI = 2 * AUC - 1 = 0{,}6492$. Це говорить про прийнятні предикативні якості моделі, тобто дозволяє спрогнозувати саму подію відтоку клієнту (чи буде відтік, чи ні).

Таблиця 1. Коефіцієнти лінійної узагальненої моделі

Source	DF	Type I SS	Mean Square	F Value	Pr > F
ONNET_MINS	1	2668.403703	2668.403703	18684.3	<.0001
OMO_MINS	1	250.054005	250.054005	1750.89	<.0001
PSTN_MINS	1	20.909361	20.909361	146.41	<.0001
INTERN_MINS	1	59.120719	59.120719	413.97	<.0001
GPRS_USG_MB	1	460.615988	460.615988	3225.26	<.0001
SUBSCRIPTION_TYPE_CO	2	25.330932	12.665466	88.68	<.0001
time	1	6285.412857	6285.412857	44010.8	<.0001

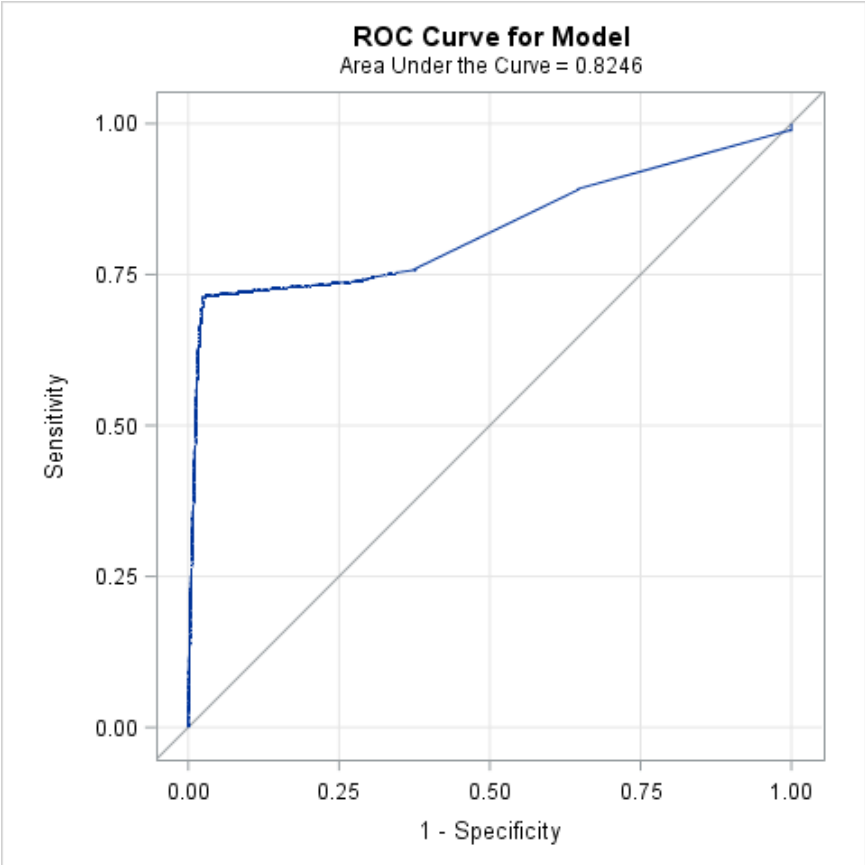

Рис. 1. ROC-крива для узагальненої лінійної моделі

Далі було здійснено моделювання та прогнозування самого факту відтоку за допомогою градієнтного бустингу, випадкового лісу, нейронних мереж та логістичної регресії. Результати для індексу GINI були на рівні 0,65 - 0,684, що вище, порівняно з узагальненою моделлю. Однак ці моделі не можуть бути використані для прогнозування самого періоду можливого відтоку.

Для прогнозування часу можливої зміни абонентом оператора зв'язку були побудовані моделі виживання та фукція втрат (збитковості) для моделі Кокса (рис. 2 та рис. 3) [6,7]. Групи абонентів розділялись на корпоративних та індивідуальних клієнтів (враховувалась також стать клієнтів). Моделювалась поведінка клієнтів кожної групи окремо для прогнозування можливого відтоку клієнтів та періоду, в який це може відбутись.

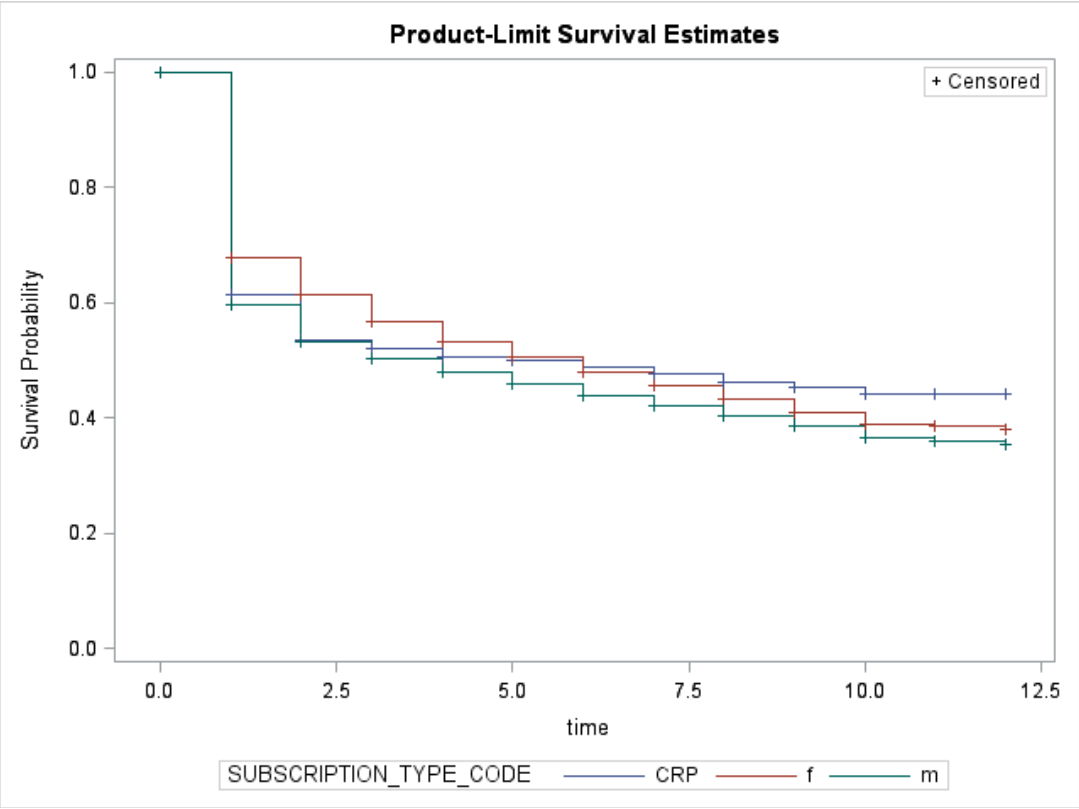

Рис. 2 Графіки функцій виживання для згрупованих даних

Визначення рівня небезпеки і ключових моментів часу, які характеризують допустимий, критичний та катастрофічний рівень ризику є задачею системного аналізу, яку необхідно вирішувати в рамках кожного виду ризику незалежно від типу ризику та галузі, в якій він спостерігається. Автором пропонується підхід, що базується на визначенні втрат компанії як допустимих $\lambda(t_1 \mid x) = c_1$, критичних $\lambda(t_2 \mid x) = c_2$ та катастрофічних $\lambda(t_3 \mid x) = c_3$, де c_1, c_2, c_3 - певні константи, які визначаються компанією в залежності від її фінансових оборотів, потужностей, тощо (наприклад, обсяг власного капіталу).

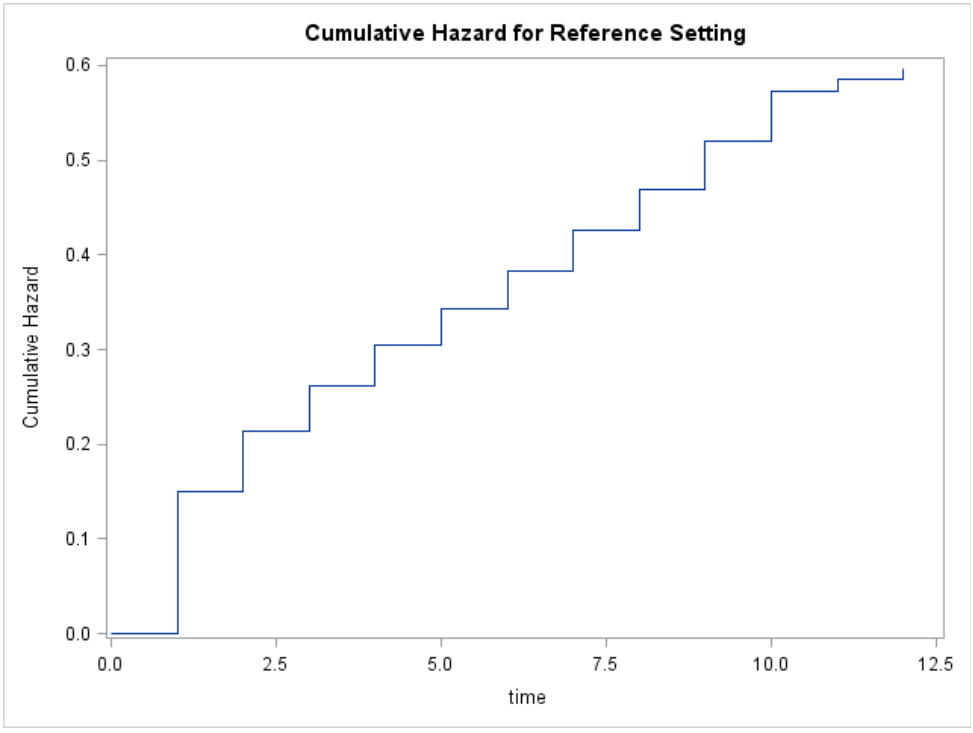

Рис. 3 Графік функції втрат для моделі Кокса

Далі, постає питання визначення допустимого, критичного та катастрофічного часу t_1, t_2, t_3. Якщо рівні втрат компанії задані на рівні 20%, 40% та 50% відповідно, то побудувавши графік функції втрат для моделі Кокса, ми отримуємо, що $t_1 = 2$ *місяці*, $t_2 = 7$ *місяців*, $t_3 = 9$ *місяців*. Таким чином, при встановлених припущеннях за нашою моделлю клієнти телекомунікаційної компанії переходять з критичного до катастрофічного рівня ризику з 7 по 9 місяць. Тому, в саме цей період телекомунікаційній компанії доцільно здійснювати додаткові дії для утримання абонентів шляхом розробки персональних пропозицій та проведенням додаткових заохочень для існуючих клієнтів.

Отже, за розглянутими моделями ми можемо визначити прийнятний ступінь ризику та час, в який момент ризик переходить в катастрофічний, а також рівень втрат, які в цей момент буде нести телекомунікаційна компанія з точки зору недоотримання доходу через відтік клієнтів (абонентів).

.

6 Висновки

Проведене моделювання підтвердило доцільність використання методів з теорії виживання, оскільки враховуються навіть спостереження з невідомим результатом, тобто ті, по яких не встановлений факт відтоку і вони досі обслуговуються оператором, що значно розширює і наближає вибірку до реальних статистичних даних. Окрім цього, побудовані моделі дозволяють включати прогнози описаних факторів, динаміку поведінки, що дозволяє будувати динамічні моделі, які є більш точними та функціональними. І, нарешті, побудовані моделі дозволяють здійснювати прогнозування фактору ризику та можливих втрат з урахуванням часу, тобто на певний період вперед.

Література

1. Cox D. R. Regression models and life-tables / D. R. Cox, S. Society, S. B. Methodological // 2007. — Vol. 34, No. 2. — P. 187–220.
2. Cao R., Vilar J.M., Devia A. Modelling consumer credit risk via survival analysis / SORT 33 (1) January-June 2009, p.3-30.
3. Marimo M. Survival analysis of bank loans and credit risk prognosis master of science mathematical statistics / M. Marimo // [Електронний ресурс]. — Режим доступу : http://wiredspace.wits.ac.za/jspui/bitstream/10539/18597/1/Mercy%20Marimo%20Thesis_Survival%20Analysis_28.03.%202015_v1.pdf.
4. Stepanova M. Survival analysis methods for personal loan data / M. Stepanova, L. C. Thomas // Operations Research. — 2002. — Vol. 50, No. 2. — P. 277–289.
5. Fleming, T.R., Harrington, D. P. Counting Processes and Survival Analysis. - John Wiley & Sons - New York. 1991.
6. Кузнєцова Н.В. Моделювання фінансового ризику в телекомунікаційній сфері / Н.В. Кузнєцова, П.І. Бідюк // Наукові вісті НТУУ “КПІ”. – 2017. – №5. – С. 51–58.
7. Бидюк П. И. Анализ временных рядов / П.И. Бидюк, В. Д. Романенко, О. Л. Тимощук. – Киев: Политехника, 2013. – 600 с.

Information Technologies for Clients’ Database Analysis and Behaviour Forecasting

© Nataliia V. Kuznietsova
Institute for Applied System Analysis of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute",
Kyiv, Ukraine
natalia-kpi@ukr.net

Abstract

In the paper the possibility of applying the information technologies for analyzing the customer database of telecommunication company subscribers with the purpose of predicting their further behaviour is shown. The task is relevant not only in terms of forecasting the fact of the change of the subscriber of the telecommunications company and the refusal of using the services, but also the moment when the subscriber only began to think about it. In the

article two problems: the task of classification (the task of forecasting the possible outflow of subscribers) and the task of predicting the time at which this event may occur are solved. Different methods of data mining could be used for the classification problem. In the article a generalized linear model was built and showed acceptable predicative properties based on the GINI index, but lower compared to logistic regression, neural networks and gradient boosting.

The author proposed to consider the problem in terms of survival of the population - subscribers of the telecommunication company. The paper presents the main theoretical information of the survival analysis and formalizes them for solving the problem of forecasting outflow of clients, in particular, taking into account their type (corporate or private client), as well as the time of occurrence of the event. The author proposes to solve the problem of clients' behaviour forecasting in time space for the early prediction of financial risks of a telecommunication company. The financial risks are caused with a lack of profit through the outflow of customers or excess costs for additional equipment, which is not needed. In the paper it is proposed to predict the period and amount of possible losses and build survival probability function and cumulative hazard function for the Cox proportional risks model. These functions allow us to determine the time at which the transition from critical to catastrophic financial risk occurs. Knowledge of the risk period will be useful for the telecommunication company in terms of preventing the outflow of subscribers by developing personalized offers and providing additional incentives for existing customers.

Keywords: information technology, financial risks, survival models, Telecommunication Company, Cox proportional risks.