ISSN 2078-6115. Вісник Львівського університету. Серія економічна. 2020. Випуск 58. С. 99-109
Visnyk of the Lviv University. Series Economics. 2020. Issue 58. P. 99-109

УДК 339.1-658.6 DOI: http://dx.doi.org/10.30970/ves.2020.58.0.5809
JEL C20, M30

АНАЛІЗ І МОДЕЛЮВАННЯ ПОВЕДІНКИ КЛІЄНТІВ З ВИКОРИСТАННЯМ БЕТА-РОЗПОДІЛІВ ВЕЙБУЛА

Тарас Панчишин, Оксана Марець, Руслана Гаєвська

Львівський національний університет імені Івана Франка
79008 м. Львів, проспект Свободи, 18
e-mail: taras.panchyshyn@lnu.edu.ua, oksana.marets@lnu.edu.ua,
gayevska.r@icloud.com

Анотація. *Ефективна маркетингова стратегія утримання клієнтів нині повинна базуватись на використанні різноманітних джерел даних та методів машинного навчання. Відповідно, маркетингові заходи зі збереження клієнтів сьогодні повинні включати інвестиції в аналітичні інструменти, точність таргетування реклами, оцінку ефективності дій та їх вартості. Таким чином, метою статті є узагальнення підходів до управлінських завдань, спрямованих на утримання клієнтів на різних етапах їх життєвого циклу, окреслення методів машинного навчання для прогнозування відтоку клієнтів та застосування різних методів для створення та тестування моделі прогнозу рівня утримання клієнтів. Для цього застосовані метод регресійного аналізу та ймовірнісний метод з використанням бета-розподілів. У статті продемонстровано, що прогнозування показника утримання клієнтів за допомогою лінійної регресії на основі перетворених даних дає змогу отримувати хороші результати навіть за невеликої кількості історичних даних. Однак, через відсутність єдиного підходу до трансформації даних у таких моделях для різних груп споживачів, обґрунтовано залучати до аналізу також ймовірнісні методи прогнозування показника утримання клієнтів. Використано бета-геометричний розподіл та бета-дискретний розподіл Вейбула. Моделювання здійснено на основі поділу даних на тренувальні та тестові для забезпечення адекватної перевірки точності моделі. За результатами дослідження сформульовано висновки, що навіть, володіючи достатньою кількістю історичних даних щодо поведінки клієнтів, ми не можемо створити універсальну модель, яка буде добре прогнозувати відтік клієнтів у всіх випадках. У кожному конкретному випадку треба враховувати демографічні характеристики групи користувачів, тривалість використання продукту та інші чинники.*

Ключові слова*: поведінка споживача, оцінювання, регресійний аналіз, показник утримання, бета-геометричний розподіл, бета-дискретний розподіл Вейбула.*

Постановка проблеми.

Враховуючи сучасні виклики, що стоять перед менеджерами по роботі з клієнтами чи не найважливішою є політика компанії щодо утримання клієнтів. І мова йде не лише про постійні традиційні маркетингові процеси утримання клієнта, а й про комплексну управлінську політику, побудовану на широких можливостях нових джерел даних та нових методологіях машинного навчання. Це і стало визначальним у формулюванні мети цього дослідження. Відтак, *метою дослідження* є узагальнення підходів до управлінських завдань, спрямованих на утримання клієнтів на різних етапах їх життєвого циклу, окреслення методів машинного навчання для прогнозування відтоку клієнтів та застосування різних методів для створення та тестування моделі прогнозу рівня утримання клієнтів.

Аналіз останніх досліджень і публікацій

Науковці та практики досліджували особливості поведінки клієнта протягом його життєвого циклу, використовуючи прикладний інструментарій середовища розробки мовою *R* [4], прогнозували метрики утримання клієнта на основі ймовірнісного підходу з використанням цього ж інструментарію *R* [7], інші автори порівнювали різні методи машинного навчання в цілях прогнозування відтоку клієнтів [8], вивчали динаміку індивідуальної поведінки клієнта та прогнозували рівень його утримання за допомогою бета-геометричного розподілу [5] або узагальнювали дослідження та праці числа науковців [1]. Однак зазначені автори спрямовували свої зусилля скоріш на прогнозування очікуваного відтоку клієнтів, а не на дослідження якісних метрик вимірювання та контролю утримання потенційних клієнтів. Існує багато підходів для моделювання цієї задачі: моделі RFM, data science, стохастичні моделі, економетричні моделі, дифузійні моделі, а також моделі рівня взаємозв'язку та моделі рівня обслуговування. Вибір того чи іншого алгоритму диктується здебільшого специфічними особливостями мети, для якої він буде реалізований.

Постановка завдання

Відповідно до досліджень попередників та власних спостережень можна дійти висновку, що рішення клієнт, приймає керуючись певними чинниками, або ж їх комбінацією. Одні клієнти можуть бути задоволені якістю продукту продовжувати співпрацю, натомість інших клієнтів можу відштовхувати активні маркетингові заходи. Одні клієнти приймають рішення про транзакцію швидко, іншим потрібно більше часу та більше інформації для завершення покупки. При розробці маркетингових кампаній фірми повинні шукати збалансовані рішення, щоб їх заходи не були надто поспішними чи, навпаки, запізнілими. Так, занадто повільна реакція на відтік клієнтів, може стати збитковою кампанією, оскільки повернення клієнта уже стане не рентабельним. Або ж занадто рання ініціативна кампанія у кращому випадку буде ще не актуальною для клієнта, а у гіршому випадку спонукатиме його до втрати лояльності та завершення співпраці.

Для оптимізації управлінських рішень та підбору необхідних стимулів для подовження життєвого циклу клієнта і збільшення показників конверсії нині маркетологам необхідно накопичувати дані та використовувати нові технології машинного навчання.

Одним із способів виявлення закономірностей поведінки клієнта є оцінка моделей ймовірності їх відтоку чи утримання як функції від часу на основі історичних даних попередніх рекламних кампаній. Наприклад, на рис. 1 продемонстровано, що ймовірність відтоку клієнтів із плином часу збільшується, тоді як ймовірність утримання потенційного клієнта зменшується.

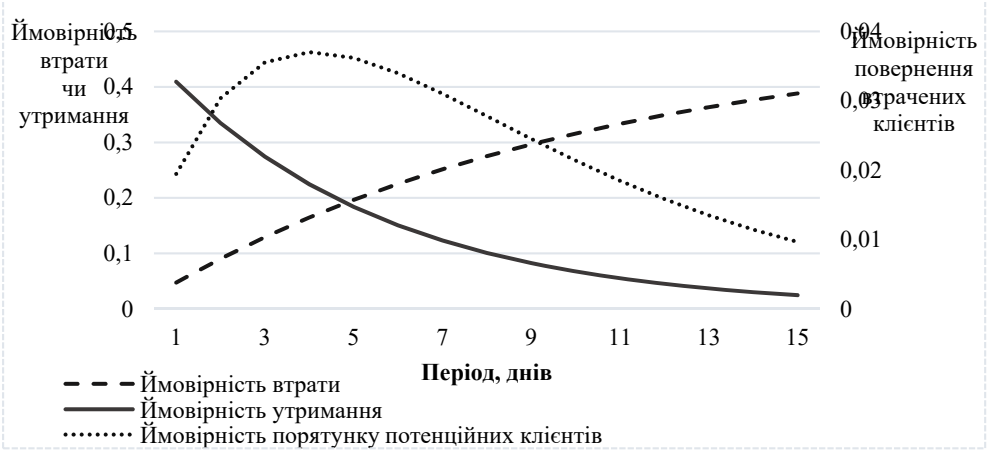

Рис. 1. Визначення періоду доцільності впровадження кампанії утримання

$P(Втрат) = \alpha_c(1-\exp^{-\beta_c t})$ та $P(Утримання) = \alpha_r \exp^{-\beta_r t}$ за умови, що: $\alpha_c = 0.5$; $\beta_c = 0.1$; $\alpha_r = 0.5$; $\beta_r = 0.2$; P(*Порятунку потенційних клієнтів*) = P(*Втрат*) × P(*Утримання*).

Джерело: [3, С.42]

Врівноваження цих тенденцій дає змогу дізнатися оптимальний час для досягнення цілі порятунку потенційного клієнта. Наявність таких даних у фірми дає їй змогу застосувати маркетингову політику у найбільш відповідний період, ще до появи ризику втрати клієнта, та одночасно не занадто рано, щоб відлякати його надмірною увагою.

На рис. 2 наведено приблизний план поведінки фірми протягом життєвого циклу клієнта. Бачимо, що не слід проводити цілеспрямованої політики відразу після залучення нового споживача. Однак навіть відразу після залучення фірма повинна подумати про його збереження.

На превентивному етапі слід таргетувати клієнта до того, як клієнт виявить будь-які ознаки зменшення зацікавленості. Наприклад, запропонувати клієнту ігрової компанії нову гру до того, як цікавість до поточної почне зменшуватися. На *проактивному* етапі варто розпочинати кампанію, орієнтовану на клієнтів, яких визначено як ризикованих, однак конкретних дій з їх боку ще не було. *Реактивний* етап — це коли фірма намагається перешкодити клієнту розривати співпрацю, вже зауваживши за ним відповідні дії. *Повернення* — це коли клієнт відписався і компанія намагається його повернути. Наприклад, проводить глибинне інтерв'ю, щоб з'ясувати причини відписки і сформувати спеціальну пропозицію. Дії *після повернення* відносяться до заходів, ініційованих після того, як клієнт відхилив пропозицію на попередньому етапі. Вони часто трапляються у галузях, де замовник повинен повернути пристрій компанії після дефекту [3.

Сьогодні компанії володіють великою кількістю структурованих та неструктурованих даних, аналіз яких дає можливість передбачати їх поведінку щодо товару/послуги, яку пропонує компанія. Результатом такого аналізу є цілеспрямовані маркетингові дії.

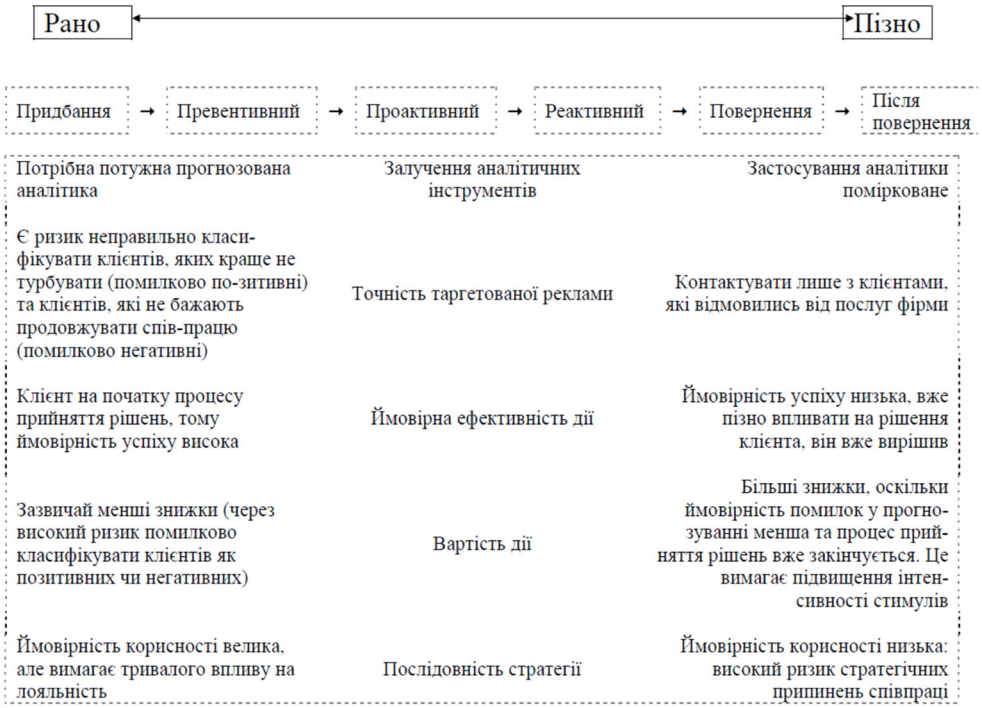

Рис. 2. Протилежні підходи до проведення кампаній утримання клієнтів

Джерело: [3, 8]

У таблиці 1 наведено дані та методи, які зараз застосовують для планування утримання клієнта, та те, де вони використовуються в процесі планування.

Таблиця 1

Дані та методи управління утриманням клієнтів

Дані	Опис	Проблема
Звичайні клієнти	Транзакції, демографічні показники	Хто в зоні ризику; На кого таргетувати рекламу
Емоції / ставлення / риси	Висновки з інтелектуального аналізу тексту	
Деталі взаємодії	Взаємодія клієнтів з товаром, наприклад його використання або поведінка під час перегляду	
Соціальний вплив / зв'язок	Залученість	
Маркетингові дії спрямовані на утримання клієнтів	Зведені дані за регіонами, а також за періодом часу та / або окремо таргетованими кампаніями	Інтеграція стратегії
Неструктуровані дані про взаємодію користувача з фірмою	Текстові або голосові дані з кол-центру або обговорень в чаті	Чому в зоні ризику; На кого, коли та з яким стимулом таргетувати рекламу

Дані	Опис	Проблема
Статистичні методи	Регресія, логістична регресія, ПММ	Індивідуальний дизайн кампанії та багаторазове планування кампанії
Ймовірнісні моделі	BG/BB, sBG, BG/NBD, Pareto-NBD	Прогнозування сукупних моделей відтоку
Машинне навчання	Дерево ухвалення рішень, бутстреп-агрегування, підсилювання, випадковий ліс.	Моделі прогнозування відтоку Проактивне управління відтоком
Інтелектуальний аналіз тексту	Тематичне моделювання, модель «торба слів»	Кількісна оцінка емоцій, ставлення та неструктурованих даних
Динамічна оптимізація	Динамічне програмування	Планування кількох кампаній
Система підтримки рішень	Агентне моделювання, оптимізаційні моделі	Управління індивідуальною та багаторазовою кампаніями; інтеграція стратегії
Експерименти в полі Field Experiment	Відсоток врятованих, неоднорідність ефектів ставлення	Індивідуальна та багаторазова кампанія

Джерело: складено за [2, 3, 8]

Відповідно до даних таблиці 1 на початковому етапі аналізу дослідники використовують інформацію про діяльність споживачів та демографічні показники щоб зробити прогноз ризику відтоку. Для цього застосовують традиційні методи, такі як логістична регресія, ймовірнісні моделі ймовірностей та моделі пропорційних ризиків. До аналізу також залучають дані про кліки, активність у соціальних мережах, соціальні зв'язки та інформацію від Call-центрів / онлайн-чатів тощо.

Ці дані у поєднанні з розвитком методів машинного навчання (наприклад, методів класифікації та інтелектуального аналізу тексту) розширили можливості для розробки стратегій утримання клієнта. Також до аналізу можна задіяти дані про соціальні зв'язки клієнтів, зміст їх взаємодії з фірмою та їх емоційний стан. Використання методів машинного навчання дозволяють дослідникам нелінійно включати великий набір предикторів (часто сотні чи тисячі) та отримувати корисну інформацію з неструктурованих джерел даних, таких як аудіо- та відеорозмови та зображення. Використання цих даних та інструментів потенційно дозволить вирішити такі питання, як, наприклад, чому клієнт, приймає рішення про вихід та як його «врятувати» [2].

Виклад основного матеріалу дослідження

Однією із задач, які вирішують аналітики при прогнозуванні поведінки клієнтів, є аналіз нових когорт користувачів нещодавно запущених сервісів чи новостворених продуктів. Особливістю такого аналізу є наявність даних лише за перші декілька періодів життя нової когорти, які маркетолог повинен проаналізувати та розробити ефективну маркетингову політику на випередження.

Покажемо, що ми можемо робити достовірні прогнози на основі декількох точок емпіричних даних і використаємо для цього можливості прикладної програми Excel. Для цього скористаємося даними утримання клієнтів, представлені в таблиці 2.

Почнемо з даних когорти 1. Як бачимо з Таблиця 2, базовому періоду, де $t = 0$ відповідає показник утримання *rate* = 1, або 100% клієнтів. Нехай це когорта клієнтів, які підписались на послугу/продукт в один і той же день. В першому періоді ($t = 1$) з нами залишилось 63% клієнтів, в другому ($t = 2$) — 47,2 % від початкової кількості і так далі. З кожним часовим періодом залишається все менше і менше клієнтів, когорта «розмивається».

Таблиця 2

Вихідні дані для аналізу утримання клієнтів

Часовий період	Рівень утримання клієнтів (retention rate)				
	Когорта 1	Когорта 2	Когорта 3	Когорта 4	Когорта 5
0	1,0000	1,0000	1,0000	1,0000	1,0000
1	0,6300	0,5310	0,8694	0,6783	0,6311
2	0,4723	0,4520	0,7404	0,5554	0,4674
3	0,3895	0,4230	0,6553	0,4759	0,3815
4	0,3469	0,3940	0,5962	0,4183	0,3277
5	0,3213	0,3750	0,5523	0,3740	0,2903
6	0,3021	0,3560	0,5182	0,3385	0,2624
7	0,2883	0,3460	0,4906	0,3093	0,2408
8	0,2792	0,3363	0,4677	0,2848	0,2234
9	0,2745	0,3279	0,4483	0,2639	0,2089
10	0,2739	0,3205	0,4316	0,2458	0,1968
11	0,2735	0,3139	0,4169	0,2300	0,1864

Джерело: [1, 4], власні розрахунки

Для подальшого аналізу поділимо дані 1-ї когорти на дві частини:

- *тренувальні*, або дані, на яких ми будемо навчати модель — дані за 1-й та 2-й періоди;
- *тестові*, або дані, на яких ми будемо перевіряти модель — дані за періоди 3—11.

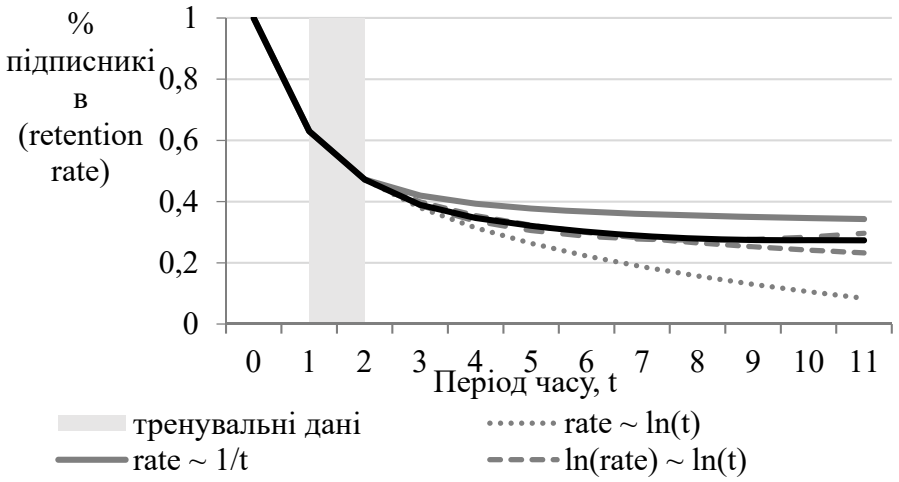

Рис. 3. Фактичні та прогнозні значення показників утримання підписників когорти 1

Джерело: власні розрахунки

Крива з даними представлена на рис. 3. З її форми, бачимо, що лінійну функцію тут застосовувати немає сенсу. Ми будемо використовувати такі перетворення даних:

1) логарифм змінної *t*: *rate* ~ ln(*t*);
2) трансформація змінної *t*: *rate* ~ 1/*t*;
3) логарифм змінних *t* та *rate*: ln(*rate*) ~ ln(*t*);

4) логарифм змінної *t* та прогноз значення *rate* як відсоток відносно попереднього періоду: $rate_{ланцюги} \sim \ln(t)$.

Для розрахунку параметрів рівняння використаємо функції MS Excel SLOPE() та INTERCEPT(). У результаті побудови лінійної регресії за двома історичними рівнями ряду і з відповідними трансформаціями ми отримали такі форми рівнянь трендів (Рис. 3).

Проаналізуємо точність прогнозування знайдених рівнянь за допомогою абсолютних похибок у відсотках (APE — absolute percentage error), що характеризують відносні розбіжності між емпіричними даними та теоретичними для кожного із рівнянь тренду для кожної точки (Рис. 4).

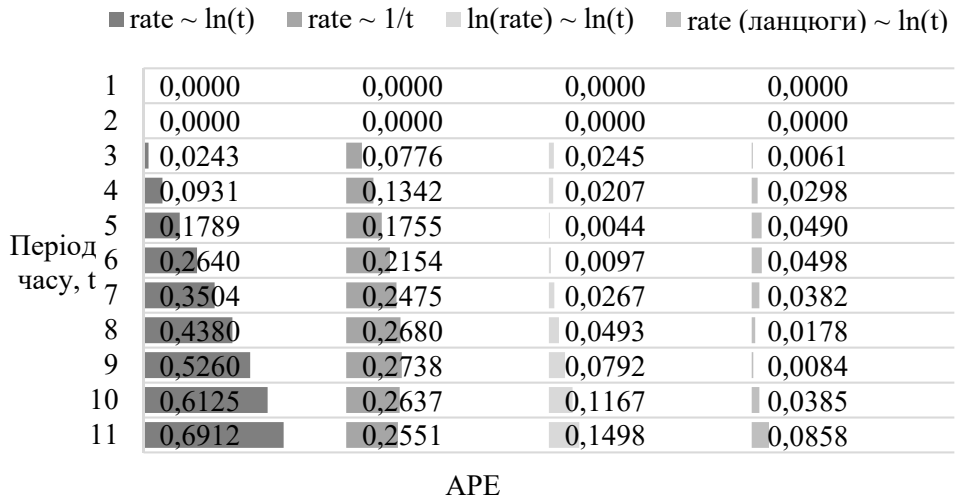

Рис. 4. Абсолютні похибки у відсотках для 4-х рівнянь для кожної точки когорти 1

Джерело: власні розрахунки

Значення APE для першого рівняння, де застосоване логарифмічне перетворення змінної *t*, вказують, що прогноз для 3-ї точки даних на 2,43% відрізняється від фактичних даних, для 4-ї точки — на 9,31%, далі похибка збільшується, і в 11-ій точці прогноз відрізняється від фактичних даних на 69,12%. Робимо висновок, що це рівняння тут не підходить, а кращими для аналізу є третя та четверта криві.

Порахуємо середні абсолютні похибки у відсотках (*MAPE*) за формулою:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i},$$

де y_i, $\hat{y}_i$ — це відповідно емпіричні та теоретичні рівні ряду, *n* — кількість тестових рівнів ряду.

З даних таблиці 3 можемо дійти висновку, що для цієї когорти найкращий прогноз дає перетворення № 4, за яким отримана залежність.

Таблиця 3

Похибки прогнозу 4-х рівнянь лінійної регресії для точок 1 та 2 когорти №1

Спосіб трансформації даних	Рівняння тренду	MAPE
rate ~ ln(t)	rate = -0,2275 * ln(t) + 0,63	0,3532
rate ~ 1 / t	rate = 0,3154/t + 0,3146	0,2123
ln(rate) ~ ln(t)	ln(rate) = -0,4156 * ln(t) – 0,462	0,0534
rateланцюги ~ ln(t)	rateланцюги = 0,1727 * ln(t) + 0,63	0,0359

Джерело: власні розрахунки

Застосуємо цей підхід на даних когорт 2 - 5 і подивимось, які перетворення для побудови рівнянь тренду будуть показувати кращі результати (Рис. 5).

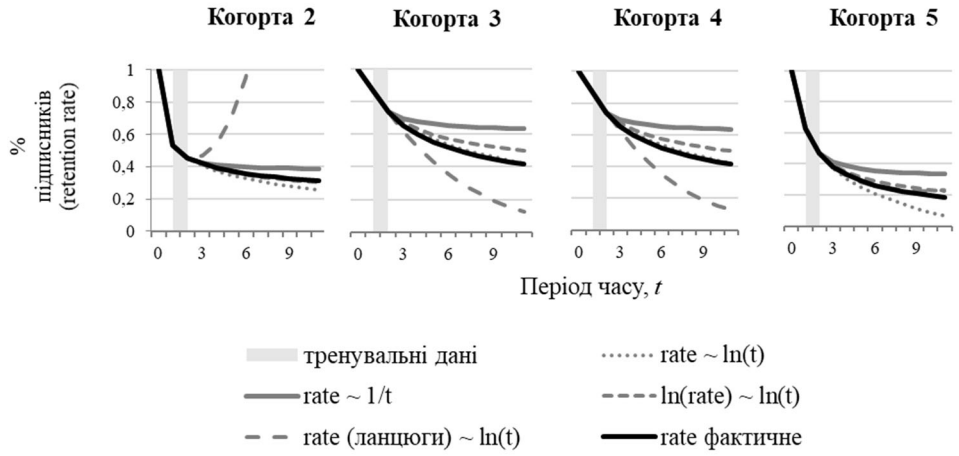

Рис. 5. Фактичні та прогнозні значення показників утримання підписників когорт 2 - 5

Джерело: власні розрахунки

Ми застосували цей підхід для когорт 2 - 5 і порахували похибки прогнозу рівнянь на тестових даних (таблиця 4).

Таблиця 4

Похибки прогнозу рівнянь (MAPE) лінійної регресії на точках 1 - 2 для когорт № 2 - 5

Спосіб трансформації даних	MAPE на тестових даних для когорт			
	2	3	4	5
rate ~ ln(t)	0,1074	0,0259	0,0691	0,3075
rate ~ 1/t	0,1337	0,3094	0,5184	0,4471
ln(rate) ~ ln(t)	0,0258	0,1219	0,2533	0,1222
rate ланцюги ~ ln(t)	6,5976	0,3921	1,0300	0,0684

Джерело: власні розрахунки

Тренування моделей для 5-х когорт показало хороші результати використання для побудови рівняння трендів 1-го, 3-го та 4-го способів трансформації вихідних даних моделі відповідно.

Для прогнозування показника утримання клієнтів використаємо ще підхід, запропонований П.Фадером і Б.Гарді [5, 6]. Цей підхід базується на використанні бета-геометричного розподілу та бета-дискретного розподілу Вейбула. За свідченням практиків, цей підхід дає досить точні результати [4, 5]. Для розрахунку прогнозних значень скористаємось можливостями бібліотеки foretell для мови програмування *R* [7].

Ми застосували цей підхід, щоб знайти параметри функції для двох точок всіх когорт. Ми спрогнозували значення коефіцієнта утримання для точок 3 - 11, та порахували середні абсолютні похибки у відсотках (*MAPE*) на тестових даних (таблиця 5).

Таблиця 5

Похибки прогнозу показника утримання підписників із застосуванням ймовірнісного підходу до точок 0 - 2 когорт № 1 - 5

Розподіл та загальний вигляд функції	MAPE на тестових даних для когорт				
	1	2	3	4	5
бета-геометричний розподіл Beta(a, b + t)/Beta(a, b)	0,2054	0,0178	0,2715	0,2286	0,1012
бета-дискретний розподіл Вейбула Beta(a, b + tc)/Beta(a, b)	0,2486	0,0804	0,2352	0,0885	0,1400

Джерело: власні розрахунки на основі бібліотеки foretell для мови програмування R [7]

Порівняння розрахованих похибок для цих рівнянь (таблиця 4, таблиця 5) вказує на те, що немає одного підходу для прогнозування, яке буде показувати хороші результати на всіх даних. Треба висувати гіпотези і тестувати їх. Крім того, для підвищення точності прогнозів є сенс створювати більш деталізовані когорти. Наприклад, групувати дані не лише за датою придбання, а за ціною, типом придбаного продукту, регіоном замовника, типом пристрою.

Висновки та перспективи подальших досліджень

Отже, прогнозування даних на основі лише двох історичних даних дає достовірні результати. Однак цей підхід не дає можливості натренувати модель для прогнозування поведінки нових когорт. Ми можемо добитися високої точності прогнозу, але для різних когорт ми отримували різні способи оптимізації моделі прогнозування клієнтів. Крива відтоку може суттєво відрізнятися між новими та поточними клієнтами через постійний вплив різних чинників. Це поява конкурентних продуктів, інфляційні очікування, активніша таргетована реклама можуть по-різному вплинути на різні когорти. Тому у такому випадку необхідно тестувати модель та підбирати свій спосіб тренування моделі. Крім того, додавання нових історичних даних в модель може змінити параметри рівняння.

Перспективи подальших досліджень передбачають залучення до аналізу поведінки клієнта дослідження неоднорідності досліджуваних сукупностей, тривалості користування продуктом, форм кривих утримання та рівня схильності до відтоку на індивідуальному рівні.

Список використаних джерел

1. Левчук П. Прогнозирование оттока клиентов в Excel. URL: https://ecommerce-in-ukraine.blogspot.com/2020/05/churn-prediction-in-excel.html. (дата звернення: 25.12.2020)
2. Ahn J., Hwang J., Kim D., Choi H., Kang, S. A survey on churn analysis in various business domains. IEEE Access. 2020, №8 URL: https://ieeexplore.ieee.org/ document/ 9281029
3. Ascarza, E., Neslin, S.A., Netzer, O. *et al.* In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions. *Cust. Need. and Solut.* 2018 №5**,** P. 65–81. URL: https://doi.org/10.1007/s40547-017-0080-0
4. Bryl' S. LTV prediction for a recurring subscription with R. 2018. URL: https://www.analyzecore.com/2018/09/19/ltv-prediction-for-a-recurring-subscription-with-r/ (дата звернення: 24.12.2020)
5. Fader P. S., Hardie B. G. S., Liu Y., Davin J., Steenburgh T. How to project customer retention revisited: The role of duration dependence. *Journal of Interactive Marketing.* 2018. №43. P. 1-16. URL: https://doi.org/10.1016/j.intmar.2018.01.002
6. Fader, P. S., Hardie, B. G. S. How to project customer retention. *Journal of Interactive Marketing*. 2007. №21 (1), P.76-90. URL: https://doi.org/10.1002/dir.20074
7. Srihari J. Projecting Customer Retention based on Fader and Hardie Probability Models. 2019. URL: https://CRAN.R-project.org/package=foretell (дата звернення: 24.12.2020)
8. Vafeiadis T., Diamantaras K. I., Sarigiannidis G., Chatzisavvas K. C. A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*. 2015. №55, P. 1-9. URL: https://doi.org/10.1016/j.simpat. 2015.03.003

References

1. Levchuk P. (2020) Prohnozyrovanye ottoka klyentov v Excel. [Forecasting customer churn in Excel]. Retrieved from: https://ecommerce-in-ukraine.blogspot.com/2020/05/churn-prediction-in-excel.html. [in Russian]
2. Ahn J., Hwang J., Kim D., Choi H., Kang, S. (2020) A survey on churn analysis in various business domains. *IEEE Access.* 8. Retrieved from: https://ieeexplore.ieee.org/document/9281029
3. Ascarza, E., Neslin, S.A., Netzer, O. *et al.* (2018) In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions. *Cust. Need. and Solut.* 5**,** 65-81. doi.org/10.1007/s40547-017-0080-0
4. Bryl' S. (2018) LTV prediction for a recurring subscription with R. Retrieved from: https://www.analyzecore.com/2018/09/19/ltv-prediction-for-a-recurring-subscription-with-r/
5. Fader P. S., Hardie B. G. S., Liu Y., Davin J., Steenburgh T. (2018) How to project customer retention" revisited: The role of duration dependence. *Journal of Interactive Marketing.* 43. 1-16. doi.org/10.1016/j.intmar.2018.01.002
6. Fader, P. S., Hardie, B. G. S. (2007) How to project customer retention. *Journal of Interactive Marketing*. 21 (1), 76-90. doi.org/10.1002/dir.20074
7. Srihari J. (2019) Projecting Customer Retention based on Fader and Hardie Probability Models. Retrieved from: https://CRAN.R-project.org/package=foretell
8. Vafeiadis T., Diamantaras K. I., Sarigiannidis G., Chatzisavvas K. C. (2015) A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*. 55, 1-9. doi.org/10.1016/ j.simpat.2015.03.003

ANALYSIS AND SIMULATION OF CUSTOMER BEHAVIOR WITH THE IMPLEMENTATION OF BETA-WEIBUL DISTRIBUTIONS

Taras Panchyshyn, Oksana Marets, Gayevska Ruslana

Ivan Franko National University of Lviv, 79008 Lviv, Prospekt Svobody, 18
e-mail: taras.panchyshyn@lnu.edu.ua, oksana.marets@lnu.edu.ua, gayevska.r@icloud.com

Abstract. Key issue of every business with customers who periodically pay for subscription is retention rate. It is important to differentiate between different types of customers from the point of view of churning, so manager can apply effective strategy to prevent it. And it is not only about the constant traditional marketing process of customer retention, but also about a comprehensive management policy based on the wide range of new data sources and new methodologies, such as machine learning. According to this approach, retention campaigns nowadays must include: initial investment in analytics capabilities, accuracy of targeting, estimation of action effectiveness and cost. Thus the purpose of the article is to summarize approaches to managerial tasks to prevent customer churn according to time factor. Also we outline wide range of machine learning techniques to predict customer churn and apply different methods to create and test a model to project retention rate.

We used such methods as analysis, synthesis, data transformation, curve-fitting regression analysis, beta-geometric distribution and beta-discrete Weibul distribution.

The article demonstrates that predicting customer retention using linear regression and training a regression model allows to obtain good results even with a small amount of historical data. However, due to the lack of a unified approach to data transformation in such models for different groups of customers, we justified the feasibility of using probabilistic method for predicting the customer retention rate using the beta-geometric distribution and beta-discrete Weibul distribution with the help of package foretell build for R. We also highlight the importance of splitting data for training and for testing models. The study concludes that even with sufficient historical data on customer behavior, we are not able to create a model that can predict the outflow of customers in all possible cases. Apart from data transformations and methods applied we should consider demographic characteristics of customer group, duration of product usage and other factors. Therefore, it is necessary to test the model and choose your way of training the model.

Keywords: Customer behavior, estimation, regression analysis, retention rate, beta geometric distribution, beta discrete Weibull distribution.

Стаття надійшла до редколегії 06.04.2020
Прийнята до друку 03.07.2020