

УДК 659.1: 330.4
JEL C61, M37

МОДЕЛЮВАННЯ ПІДПИСКИ НА ЕМЕЙЛ РОЗСИЛКУ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Ольга Ключник¹, Мирослав Дацко²

¹ТОВ «МОВАДЕКС»

e-mail: olaklochnik14@gmail.com; ORCID: 0009-0009-7293-1748

²Львівський національний університет імені Івана Франка,
79008, м. Львів, просп. Свободи, 18

e-mail: myroslav.datsko@lnu.edu.ua; ORCID: 0000-0002-5181-463

Анотація. Цифрова сфера пропонує широкий спектр платформ і стратегій для реклами, що вимагає глибокого розуміння того, як можна оптимізувати онлайн-кампанії, щоб отримати максимальний вплив на споживачів. Метою статті є сегментування користувачів інтернет магазину та побудова моделей оцінки ймовірності підписки на e-mail розсилку для підвищення ефективності рекламних кампаній. У дослідженні використано методи кластерного аналізу, інтелектуального аналізу даних, логістичної регресії, які дають змогу аналізувати та прогнозувати поведінку споживачів, враховуючи різноманітні фактори, такі як цільова аудиторія, канали розміщення. Сегментування споживачів та побудова моделей оцінки ймовірності підписки на e-mail розсилку допомагає максимізувати результативність інвестицій у рекламні кампанії. Застосування математичного моделювання покращить розробку рекламних стратегій в інтернеті, оптимізує рекламний бюджет та збільшить конверсії.

Ключові слова: аналіз даних, інтернет-реклама, кластерний аналіз, логістична регресія, маркетингові дослідження, моделювання.

Постановка проблеми. У сучасному цифровому світі, де конкуренція в інтернеті надзвичайно висока, важливою є ефективність інтернет-реклами. Аналіз та прогнозування різноманітних аспектів цієї сфери дозволяє оптимізувати використання ресурсів та забезпечувати максимальну ефективність шляхом адаптування до швидкозмінного споживчого середовища та технологічних тенденцій. Методи прямого маркетингу, зокрема, директ мейл є тим інструментом, що потребує додаткового вивчення в контексті використання інтелектуальних методів аналізу даних.

Аналіз останніх досліджень і публікацій. Проблема ефективності маркетингових досліджень активно вивчається у наукових колах і знаходить своє відображення у працях вітчизняних та закордонних науковців. Зокрема Паламарчук О. М., Коркач І. В. звертають увагу на еволюцію маркетингових методів дослідження ринку зі зміною способів отримання споживачами інформації [13]. Процишин Ю. Т. [14] розглядає

теоретичні основи e-mail-маркетингу як ефективного інструменту комунікації із споживачем та виділяє його відмінності від спаму та разової розсилки, Янчук Т. В., Любінчак К. Р. [18] визначають основні переваги застосування методів прямого маркетингу та пропонують схему для створення ефективних поштових повідомлень під час проведення директ-маркетингових кампаній. Маказан Є. В. [12], Скригун Н.П., Семененко К.Ю. [15], стверджують що прямі поштові звернення мають значні переваги у формуванні лояльності споживачів, оскільки спрямовані на цільову аудиторію, використовують індивідуальні звернення до потенційних та існуючих клієнтів.

Разом з тим науковці активно досліджують необхідність широкого застосування методів інтелектуального аналізу даних для маркетингових досліджень, зокрема Ханін О. Г. [17] звертає увагу, що існує необхідність кластеризації даних в задачах маркетингу та констатує що групування в певному розумінні близьких об'єктів дозволяє підходити до вивчення кожної групи з єдиної позиції, наприклад, пропонуючи єдину маркетингову стратегію. Проблемі кластеризації даних присвячені праці О. І. Гарафонова [11], А.В. Федорченко, А.Б. Кулик, І.В. Пономаренко [16], А. Adolfsson, M. Ackerman, N.C.Brownstein [1].

У дослідженнях на схожу тематику розглядаються різні аспекти, зокрема ефективність рекламних стратегій в інтернеті, вплив соціальних мереж на просування брендів, поведінку споживачів у цифровому просторі та використання аналітики даних для таргетованого маркетингу. Серед них варто відзначити фундаментальні роботи Котлера про принципи онлайн-маркетингу [6], дослідження Ворамонтрі та Кліба про вплив соціальних мереж на поведінку споживачів [9], а також дослідження Десле та Клоарек про етичні наслідки реклами, керованої даними [2].

Незважаючи на значну кількість публікацій, додаткових наукових пошуків, на нашу думку, потребує проблема прикладного комплексного застосування методів інтелектуального аналізу даних у маркетингових дослідженнях з подальшою інтерпретацією отриманих результатів.

Постановка завдання. *Метою статті* є сегментування користувачів інтернет магазину та побудова моделей оцінки ймовірності підписки на e-mail розсилку з використанням методів інтелектуального аналізу даних та логістичної регресії.

Методи дослідження. У статті використано різноманітні методи наукових досліджень. Системний – для визначення структури маркетингового дослідження. Методи моделювання: кластерний аналіз, логістична регресія, для побудови моделей оцінки ймовірності підписки на e-mail розсилку. Також застосовувалися методи емпіричного та теоретичного дослідження, такі як спостереження, порівняння, вимірювання та аналіз, для забезпечення комплексного підходу для обґрунтування висновків.

Виклад основного матеріалу дослідження. За останні роки цифровий маркетинг та інтернет-реклама зазнали значної трансформації, зумовленої стрімким розвитком інтернету та технологій. Очікується, що у 2024 році частка витрат на цифрові технології зросте до 59,1% та 60,3% у 2025 р. [7].

Цифровий маркетинг передбачає використання різних онлайн-каналів, таких як пошукові системи, соціальні мережі, електронна пошта та вебсайти, для просування товарів чи послуг, підвищення пізнаваності бренду та взаємодії зі споживачами [3].

З розвитком технологій методи онлайн-реклами еволюціонують, використовуючи інноваційні підходи, такі як програмована реклама та інтерактивний контент. Потужним інструментом для розуміння та сегментації моделей поведінки користувачів, що дозволяє розробляти ефективні та цілеспрямовані рекламні кампанії є алгоритми кластеризації [10].

Email-маркетинг – популярний спосіб для бізнесу залишатися на зв'язку зі своїми клієнтами. Однак не всі користувачі готові підписатися на електронну розсилку, тому використання логістичної регресії дозволить впевненіше прогнозувати, які користувачі з найбільшою ймовірністю підпишуться на отримання розсилки, що дасть змогу зосередити свої зусилля на найбільш сприйнятливій аудиторії, максимізуючи ймовірність перетворення потенційних клієнтів на підписників.

Логістична регресія – тип статистичної моделі (також відомий як логіт-модель) часто використовується для класифікації та прогнозовної аналітики. Логістична регресія оцінює ймовірність події на основі заданого набору даних незалежних змінних [4].

Для проведення дослідження у роботі було використано мову програмування Python [8] та такі основні бібліотеки: Scikit-learn, Numpy, Pandas, Matplotlib, а також набір даних про 5000 користувачів та 11 характеристик, зібрані інтернет-магазином у 2021 році, який було отримано із Kaggle [5].

Попередній аналіз покупців показав, що більшу частину з них складає чоловіча аудиторія, яка виявляє інтерес до продуктів, які пропонуються в цьому магазині. (Рис.1).

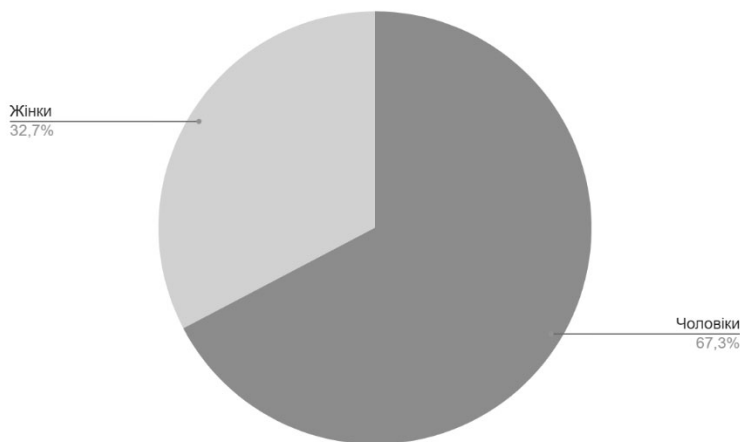


Рис. 1. Розподіл користувачів онлайн- магазину за статтю

Аналіз користувачів за типом пристрою, через який вони входять на платформу, показав, що найбільша кількість користується настільною версією інтернет-магазину,

що може бути пов'язано з тим, що комп'ютери та ноутбуки є більш комфортними при перегляді продуктів та оформленні замовлень, що полегшує взаємодію з інтерфейсом магазину (рис. 2).

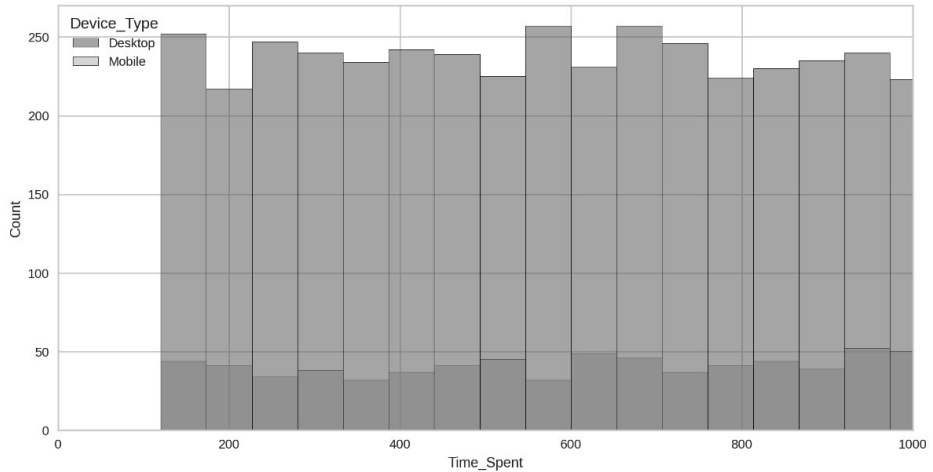


Рис. 2. Тип пристрою, який використав користувач для покупок в онлайн-магазині

Також користувачі частіше обирали можливість входу за допомогою своїх облікових записів Google або Facebook, що дозволяє уникнути необхідності створювати новий обліковий запис та швидко входити в систему (рис. 3).

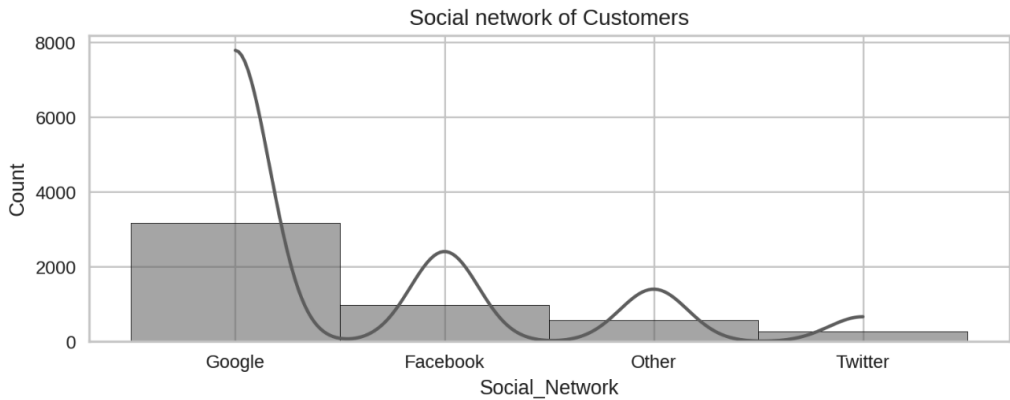


Рис. 3. Соціальна мережа, через яку користувач увійшов в систему

Вхідні змінні складаються з числових, категорійних і бінарних змінних. Детальніший опис вхідних змінних показано в таблиці 1.

Таблиця 1

Опис вхідних змінних для сегментації користувачів онлайн-магазину

Вхідна змінна	Опис змінної	Тип даних
Customer_id	Ідентифікатор користувача	Числові
Age	Вік користувача	Числові
Gender	Стать користувача	Бінарні (0: чоловік, 1: жінка)
Revenue_Total	Загальна сума продажу по користувачу	Числові
N_Purchases	Кількість покупок за день	Числові
Purchase_VALUE	Вартість останньої покупки в євро	Числові
Pay_Method	Спосіб оплати	Категорійні (0: Digital Wallets, 1: Card, 2: PayPal, 3: Other)
Time_Spent	Час проведений на сайті в секундах	Числові
Social_Network	Соціальна мережа, через яку користувач увійшов в систему	Категорійні (0: Google, 1: Facebook, 2: Twitter, 3: Other)
Device_Type	Тип пристрою, який використав користувач	Бінарні (0: десктоп-версія 1: мобільна версія)
PromoCode	Використання промокоду	Бінарні (0: не використано, 1: використано)

Для сегментації користувачів цього інтернет-магазину використано методи кластеризації, які є алгоритмами неконтрольованого машинного навчання, такі як K-means, ієрархічна кластеризація. Кожен алгоритм має свої переваги та недоліки. Тому відповідно до отриманих результатів кількох моделей і порівняння кластерів, було побудовано остаточну модель.

Для кластеризації методом K-means першочергово було перетворено числові ознаки для використання однакового масштабу за допомогою класу StandardScaler, який здійснює стандартизацію показників таким чином, щоб вони мали середнє значення 0 і стандартне відхилення 1.

На наступному етапі відібрано відповідну кількість кластерів за допомогою методу ліктя. Оптимальна кількість кластерів становила 3. Щоб оцінити продуктивність побудованої моделі обраховано силуетний показник. Коефіцієнт силуету цієї моделі склав 0.5644, що вказує на допустиме розділення кластерів. Кількість клієнтів у кожному кластері для даної моделі показано на рисунку 4.

Оскільки кластеризація K-means не є очевидним чи оптимальним рішенням, тому додатково здійснено розподіл користувачів за допомогою ієрархічної кластеризації з використанням методу Варда. Очевидною перевагою цього методу є відсутність необхідності попереднього визначення кількості кластерів, на відміну від кластеризації методом K-means.

За допомогою побудови дендрограми, можна візуалізувати етапи проведеної ієрархічної кластеризації. Важливо враховувати, що відстані між вертикальними лініями на дендрограмі вказують на рівень поділу між кластерами.

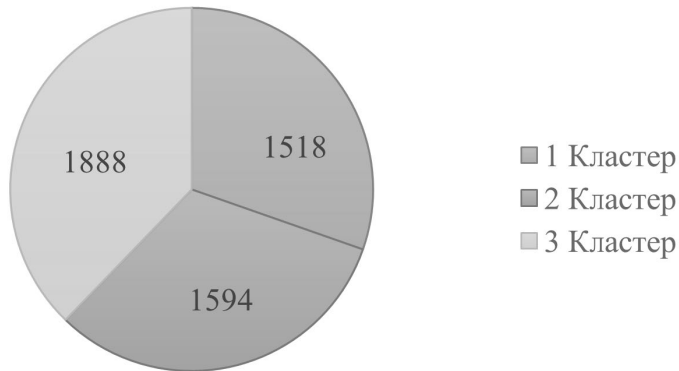


Рис. 4. Кількість користувачів у кожному кластері за методом K-means

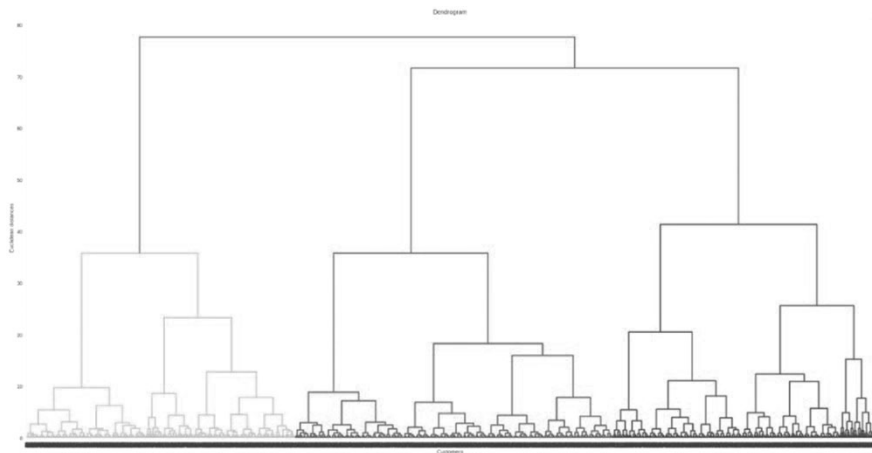


Рис. 5. Ієрархічна дендрограма користувачів інтернет-магазину

За допомогою отриманої дендрограми, зображеної на рисунку 5, можна встановити порогове значення рівним 50, і на основі початкових даних очевидно, що горизонтальна лінія перетинає 3 вертикальні лінії дендрограми.

Можна побачити, що як і в попередньому методі так і тут було виокремлено 3 кластери. Тому оптимально сегментувати споживачів онлайн-магазину на три групи (сегменти). Розподіл користувачів за кластерами відображено на рисунку 6.

Проаналізувавши користувачів за кожним кластером отримано наступні результати.

До першого кластеру увійшли користувачі середнього віку, що купують на великі суми. Це може вказувати на високий рівень доходу користувачів цього кластеру. При цьому кількість покупок у цьому сегменті є порівняно меншою проте їх вартість вища, що може означати купівлю дорожчих товарів. Користувачі з цього кластеру також

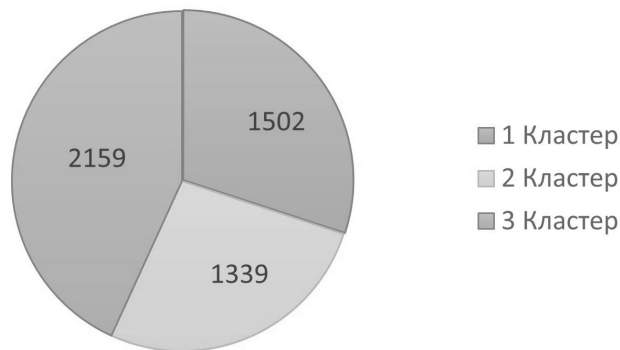


Рис. 6. Кількість користувачів у кожному кластері за ієрархічною кластеризацією

проводять менше часу на сайті порівняно з іншими кластерами. Крім цього, в цей сегмент входить найбільше користувачів жінок і вони надають перевагу оплаті іншими способами.

Другий кластер характеризується користувачами з найнижчим середнім віком, найнижчою загальною сумою продажу і середньою кількістю покупок. Вартість товарів, які купують користувачі в цьому кластері також є найнижчою. Аналізуючи тривалість перебування на сайті, можна побачити, що клієнти проводять менше часу на сайті порівняно з третім кластером. Більшість користувачів чоловіки. Покупці використовують електронні гаманці під час оплати та надають перевагу входу в систему за допомогою Google акаунта.

Третій кластер характеризується користувачами з найвищим середнім віком та більшим загальним чеком за день, що може свідчити про вищий рівень доходу серед всіх користувачів. Цей кластер також відрізняється найбільшою кількістю покупок, але середня вартість куплених товарів в ньому є нижчою ніж у першому кластері. Також користувачі, які увійшли до цього сегменту проводять найбільше часу на сайті та надають перевагу входу в систему за допомогою електронної пошти та пароллю – без використання соцмереж, а також надають перевагу оплаті іншими платіжними методами.

Проведений аналіз даних дозволив сформулювати характеристику користувачів для кожного сегменту, яка представлена в таблиці 2.

Таблиця 2

Характеристика користувачів за сегментами

	1 Кластер	2 Кластер	3 Кластер
Середній вік	30–40	18–30	40–60
Стать	Жінки	Чоловіки	Чоловіки
Тривалість перебування на сайті, хв	15–25	30–40	60–80
Загальна сума продажу, €	25 000–45 000	30 000–60 000	5 000–8 000
Кількість покупок на день	1–3	4–5	6–7
Тип пристрою	Мобільна версія	Настільна версія	Настільна версія
Спосіб входу в систему	Google Account	Google Account	Created Account

Отже, кластеризація дозволяє таргетувати рекламу на конкретні сегменти користувачів, які найбільше зацікавлені в продукції, створювати релевантні рекламні повідомлення, що відповідають інтересам і потребам цільової аудиторії та використовувати ефективні ключові слова у рекламі для користувачів, що зацікавлені товарами інтернет-магазину.

Е-mail-маркетинг – популярний спосіб для бізнесу залишатися на зв'язку зі своїми клієнтами. Однак не всі готові підписатися на електронну розсилку, тому використання логістичної регресії дозволяє прогнозувати, які користувачі з найбільшою ймовірністю погодяться на використання такого каналу маркетингової комунікації, що дасть змогу зосередити свої зусилля на сприйнятливій аудиторії та дозволить оптимізувати розсилку.

Для побудови моделей логістичної регресії було використано показники (табл. 3), які є важливими для прогнозування належності до конкретного сегмента, а також додано бінарну змінну «Newsletter» (NL), яка вказує на те, чи підпишеться користувач на розсилку.

Таблиця 3

Опис показників для побудови моделей логістичної регресії

Показник	Умовне позначення
NL	Підписка на розсилку (1 – так, 0 – ні)
Age	Вік користувача
Gen	Стать користувача
RevT	Загальна сума продаж, здійснених користувачем
NPur	Кількість покупок за день
PurV	Вартість останньої покупки
PayM_C	Спосіб оплати – Картою
PayM_PP	Спосіб оплати – PayPal
PayM_DW	Спосіб оплати – Цифровий гаманець
PayM_O	Спосіб оплати – Інший спосіб
TSp	Час проведений на сайті, сек
SN_G	Соціальна мережа входу в систему – Google
SN_F	Соціальна мережа входу в систему – Facebook
SN_O	Соціальна мережа входу в систему – логін та пароль
SN_T	Соціальна мережа входу в систему – Вхід через логін та пароль
DevT	Тип пристрою (1 – десктопна, 0 – мобільна)
ProC	Використання промокоду (1 – так, 0 – ні)

Оскільки у результаті проведеної кластеризації, оптимальною кількістю було визначено три сегменти цільової аудиторії, то, на нашу думку, доцільно побудувати три окремі моделі логістичної регресії, по одній моделі для кожного кластеру, щоб виявити патерни та важливі змінні, що впливають на ймовірність підписки на розсилку. Надалі можна використати цю інформацію для розробки кращих маркетингових стратегій, персоналізованих підходів та оптимізації рекламних кампаній для кожної групи споживачів.

На рисунку 7 можна побачити, що у першому та другому кластері кількість підписок є досить малою, а у третьому кластері більша частина покупців погодилась на підписку, тому потрібно проаналізувати, які фактори на це вплинули у кожному випадку.

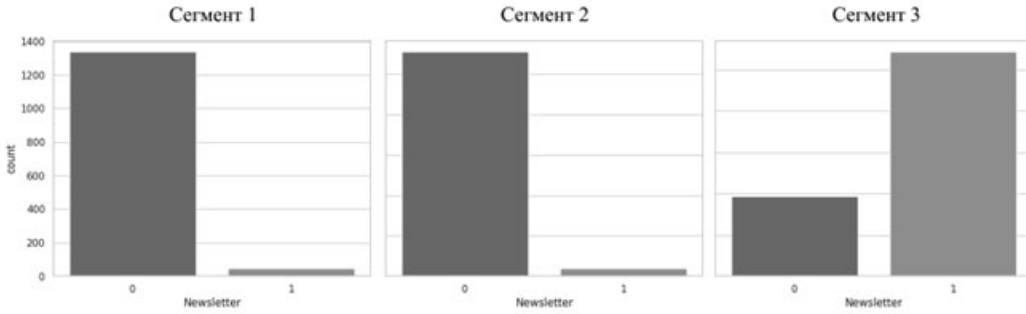


Рис. 7. Співвідношення підписки на розсилку за сегментами споживачів

Логістична модель для першого сегменту має вигляд:

$$\ln \left(\frac{p}{1-p} \right)_1 = 0,25 - 1,6857SN_o - 0,7695SN_G + 0,0727RevT + 0,0028TSp - 2,9159SN_F - 5,0177PayM_o - 4,3455PayM_{pp} - 5,1101PayM_c - 5,6014PayM_{DW} \quad (1)$$

Модель для другого сегменту має вигляд:

$$\ln \left(\frac{p}{1-p} \right)_2 = 0,5003 - 0,5001ProC - 0,7358SN_G - 2,4932PayM_{DW} - 2,4381SN_T - 0,0007TSp - 0,1359NPur - 2,5199PayM_{pp} - 52,5199PayM_c + 2,8059Gen \quad (2)$$

Для третього сегмента модель виглядатиме так:

$$\ln \left(\frac{p}{1-p} \right)_3 = 0,509 + 2,8534PayM_{DW} + 3,0053PayM_c + 0,8283SN_G + 3,2567PayM_{pp} + 1,2166SN_F - 0,0013Sp - 0,0335Age - 0,0261RevT + 2,3633PayM_o \quad (3)$$

Важливо розуміти, що наведені коефіцієнти для моделей логістичної регресії не можна трактувати аналогічно до коефіцієнтів лінійної регресії, оскільки залежною змінною рівняння виступає не безпосередньо ймовірність чи здійснить користувач підписку на розсилку, а натуральний логарифм від «log-odds ratio». Тому для одержання ймовірності підписки слід здійснити зворотній перехід.

Для оцінювання статистичної значущості зв'язків, що описуються моделями, проаналізовано такі показники як стандартна похибка, z -scores та p -value.

Оцінка p -значення, яке вказує на статистичну значущість кожного коефіцієнта за нульовою гіпотезою «коефіцієнт дорівнює нулю» показала що всі коефіцієнти мають значущість ($P > |z| < 0.05$), це означає, що всі вони відрізняються від нуля зі статистичною значущістю на рівні довіри 95%, для трьох моделей.

Також інші метрики, такі як AIC (інформаційний критерій Акаїке) та BIC (бассівський інформаційний критерій), які використовуються для порівняння моделей, показують низькі значення, що вказує на високу придатність моделей.

Оцінка точності моделей рівна 0,8910 для першої моделі та 0,878 та 0,7158 відповідно для другої та третьої, дає підстави говорити про хорошу якість розроблених моделей.

Висновки та перспективи подальших досліджень. Логістична регресія може бути ефективним інструментом для оцінки ймовірності підписки користувачів на e-mail розсилку. З отриманих результатів можна зробити висновок, що побудовані моделі логістичної регресії є придатними для прогнозування підписки на розсилку на основі показників, які було розглянуто.

Факторами, які позитивно впливають на ймовірність підписки користувачів першого сегменту, є загальна сума продажу по користувачу та час, проведений на сайті, тобто, чим більше часу перебуває користувач на сайті та чим більшу суму він витратив за увесь час, тим більша ймовірність, що користувач підпишеться на розсилку. У той же час можливість входу через Google акаунт або за допомогою входу в обліковий запис, створений саме в цьому інтернет-магазині, негативно впливають на ймовірність підписки на розсилку електронною поштою. Такі фактори, як вхід за допомогою Facebook акаунту та оплата PayPal картою за допомогою цифрового гаманця також мають негативний вплив.

Для другого сегменту – фактори, такі як використання промокоду, вхід за допомогою Google та Twitter, спосіб оплати (цифрові гаманці, PayPal, картки) та час, проведений на платформі, негативно впливають на ймовірність. Натомість такі фактори як кількість покупок, спосіб оплати картою та стать користувача мають позитивний вплив на ймовірність підписки.

Аналізуючи коефіцієнти моделі логістичної регресії для останнього сегменту можна зробити висновки, що користувачі які використовують цифрові гаманці, картки, Google та PayPal, з більшою ймовірністю підпишуться на e-mail розсилку.

При цьому, чим старші користувачі, довше провели часу на сайті та чим більше витратили на покупки і увійшли в акаунт через Facebook, то з меншою ймовірністю підпишуться.

Результати побудови логістичної регресії можуть допомогти в оптимізації рекламних кампаній, шляхом ефективного розподілу ресурсів, надаючи пріоритет сегментам з вищою прогнозованою ймовірністю підписки. Такий цілеспрямований підхід може призвести до зростання рентабельності використання маркетингових бюджетів та більшої віддачі від інвестицій.

Список використаних джерел

1. Adolfsson A., Ackerman M., Brownstein N. C. To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*. 2019. Vol. 88. P. 13–26. URL: <https://doi.org/10.1016/j.patcog.2018.10.026>
2. Deslée, A., & Cloarec, J. Safeguarding Privacy: Ethical Considerations in Data-Driven Marketing. In *The Impact of Digitalization on Current Marketing Strategies*. Emerald Publishing Limited. 2024. P. 147–161.
3. Chaffey D., Ellis-Chadwick F. *Digital Marketing: Strategy, Implementation and Practice*. Pearson Education. Limited. 2016. 728 p.
4. IBM. What is logistic regression? URL: <https://www.ibm.com/topics/logistic-regression>.
5. Kaggle. URL: <https://www.kaggle.com>.
6. Kotler P., Keller K. *Marketing Management: 15th Edition*. Pearson. 2015. 832 p.
7. Latest 2023 dentsu ad spend re-forecast shows: Digital reaches maturity in the marketing mix. URL: <https://www.dentsu.com/news-releases/latest-2023-dentsu-ad-spend-re-forecast>
8. Python documentation. URL: <https://www.python.org>.
9. Voramontri D., Klieb L. Impact of social media on consumer behaviour. *International Journal of Information and Decision Sciences*. 2019. Vol. 11(3). P. 209–233.
10. Unlocking Insights: Cluster Analysis, Segmentation, and Customer Segmentation for Effective Marketing. URL: <https://www.wizaly.com/blog/analysis/cluster-analysis-segmentation>.
11. Гарафонов О. І. Використання cgm-системи та кластеризації у маркетинговій діяльності сучасного підприємства. *Науковий вісник Полісся*. 2015. № 3 (3). С. 85–89.
12. Маказан Є. В. Direct-mail як ефективний інструмент прямого маркетингу. Вісник Приазовського державного технічного університету. 2016. Вип. 32. С. 197–203
13. Паламарчук О. М., Коркач І. В. Тенденції розвитку цифрового маркетингу в сучасних умовах господарювання. *Економічний вісник університету*. 2023. Вип. № 57. С. 63–68. doi: <https://doi.org/10.31470/2306-546X-2023-57-63-68>
14. Процишин Ю. Т. Ефективність використання e-mail маркетингу. [The effectiveness of using e-mail marketing] *Міжнародний мультидисциплінарний науковий журнал «ЛОГОС. Мистецтво наукової думки»*. 2019. № 5. С. 13–15. URL: <https://ojs.ukrlogos.in.ua/index.php/2617-7064/article/view/20>.
15. Скригун Н. П., Семененко К. Ю., Авраменко І. М. Е-mail як ефективний канал маркетингових комунікацій. *Причорноморські економічні студії*. 2016. Вип. 6. С. 106–110. URL: http://nbuv.gov.ua/UJRN/bses_2016_6_24
16. Федорченко А. В., Кулик А. Б., Пономаренко І. В. Особливості застосування методу кластеризації в маркетингових дослідженнях фармацевтичного ринку України. *Маркетинг і цифрові технології*. 2023. Т. 7. № 1. С. 7–28.
17. Ханін О. Г. Метод χ^2 - кластеризації в задачах маркетингу. *Економічний аналіз: зб. наук. праць*. Тернопіль : Видавничо-поліграфічний центр Тернопільського національного економічного університету «Економічна думка», 2016. Т. 26. № 1. С. 38–42.
18. Янчук Т. В., Любінчак К. Р. Прямий маркетинг в Україні: основні проблеми та перспективи розвитку. *Економіка та суспільство*. 2022. Вип. 43. doi: <https://doi.org/10.32782/2524-0072/2022-43-80>.

References

1. Adolfsson, A., Ackerman, M., Brownstein, N. C. (2019) To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*. Vol. 88. 13–26. URL: <https://doi.org/10.1016/j.patcog.2018.10.026>
2. Deslée, A., & Cloarec, J. (2024). Safeguarding Privacy: Ethical Considerations in Data-Driven Marketing. In *The Impact of Digitalization on Current Marketing Strategies* (pp. 147–161). Emerald Publishing Limited.
3. Chaffey, D., Ellis-Chadwick, F. (2016) *Digital Marketing: Strategy, Implementation and Practice*. Pearson Education Limited. 728 p.
4. IBM. What is logistic regression? URL: <https://www.ibm.com/topics/logistic-regression>.
5. Kaggle. URL: <https://www.kaggle.com>.
6. Kotler P., Keller K. L. (2015) *Marketing Management* 15th Edition. Pearson, 832 p.
7. Latest 2023 dentsu ad spend re-forecast shows: Digital reaches maturity in the marketing mix. URL: <https://www.dentsu.com/news-releases/latest-2023-dentsu-ad-spend-re-forecast>.
8. Python documentation. URL: <https://www.python.org>.
9. Voramontri, D., & Klieb, L. (2019). Impact of social media on consumer behaviour. *International Journal of Information and Decision Sciences*, 11(3), 209–233.
10. Unlocking Insights: Cluster Analysis, Segmentation, and Customer Segmentation for Effective Marketing. URL: <https://www.wizaly.com/blog/analysis/cluster-analysis-segmentation>.
11. Harafonova, O. I. (2015) Vykorystannia crm-systemy ta klasteryzatsii u marketynhovii diialnosti suchasnoho pidpriemstva. [Using CRM systems and clustering in the marketing activities of a modern enterprise]. *Naukovyi visnyk Polissia*. № 3 (3). 85–89. [in Ukrainian].
12. Makazan, Ye. V. (2016) Direct-mail yak efektyvnyi instrument priamoho Markentyntu. [Direct mail as an effective direct marketing tool]. *Reporter of the priazovskyi state technical university*. Section: Economic sciences. 2016. Is. 32. Vol 1. 197–203 [in Ukrainian].
13. Palamarchuk, O. M., Korkach, I. V. (2023) Tendentsii rozvytku tsyfrovoho marketynhu v suchasnykh umovakh hospodariuvannia. [Trends in the development of digital marketing in modern business conditions]. *Ekonomichnyi visnyk universytetu*. Vypusk №57. 63–68 doi:<https://doi.org/10.31470/2306-546X-2023-57-63-68> [in Ukrainian].
14. Protsyshyn, Yu.T. (2019) Efektyvnist vykorystannia e-mail marketynhu. *Mizhnarodnyi multydystrylinarnyi naukovyi zhurnal «ΛΟΗΟΣ. Mystetstvo naukovoï dumky»* №5, 13–15. URL:<https://ojs.ukrlogos.in.ua/index.php/2617-7064/article/view/20> [in Ukrainian].
15. Skryhun, N.P., Semenenko, K.Iu., Avramenko, I.M. (2016). E-mail yak efektyvnyi kanal marketynhovykh komunikatsii. [E-mail as an effective marketing communications channel.] *Prychornomorski ekonomichni studii*. Vypusk 6. 106–110. URL: http://nbuv.gov.ua/UJRN/bzes_2016_6_24 [in Ukrainian].
16. Fedorchenko, A. V., Kulyk, A. B., Ponomarenko, I. V. (2023). Osoblyvosti zastosuvannia metodu klasteryzatsii v marketynhovykh doslidzhenniakh farmatsevtichnoho rynku Ukrainy. [Peculiarities of applying the clustering method in marketing research of the pharmaceutical market of Ukraine] *Marketynh i tsyfrovi tekhnolohii*. Tom 7. № 1. 7–28 [in Ukrainian].
17. Khanin, O. H. (2016) Metod χ^2 - klasteryzatsii v zadachakh marketynhu [Tekst] [χ^2 -clustering method in marketing problems] *Ekonomichnyi analiz: zb. nauk. prats. Ternopil: Vydavnycho-polihrafichnyi tsentr Ternopil'skoho natsionalnoho ekonomichnoho universytetu «Ekonomichna dumka»*, Tom 26. № 1. 38–42. ISSN 1993-0259. [in Ukrainian].
18. Yanchuk, T.V., Liubinchak, K. R. (2022) Priamyi marketynh v Ukraini: osnovni problemy ta perspektyvy rozvytku. [Direct marketing in Ukraine: main problems and development

prospects] *Ekonomika ta suspilstvo*. Vypusk 43. 2022. doi: <https://doi.org/10.32782/2524-0072/2022-43-80>. [in Ukrainian].

EMAIL SUBSCRIPTION MODELING USING MACHINE LEARNING METHODS

Olha Klochnyk¹, Myroslav Datsko²

¹*Quality Assurance Engineer, MOVADDEX LLC,
e-mail: olaklochnik14@gmail.com; ORCID: 0009-0009-7293-174*

²*Ivan Franko National University of Lviv,
18 Svobody Ave., Lviv, 79008,
e-mail: myroslav.datsko@lnu.edu.ua; ORCID: 0000-0002-5181-4632*

Abstract. The digital sphere offers many advertising platforms and strategies, which necessitates a profound comprehension of how online campaigns can be optimised to maximise consumer impact.

The relevance of the study is determined by the necessity to apply methods and models that will allow the analysis and prediction of the effectiveness of advertising campaigns, taking into account various factors, such as the target audience and placement channels, to maximise the effectiveness of investments in this segment of activity. In this regard, the objective of the study is to utilise data mining methods to segment online store users and construct models for estimating the probability of subscribing to an e-mail using logistic regression.

In the segmentation of online store users, clustering methods were employed, namely unsupervised machine learning algorithms, including K-means and Ward's method. Each algorithm possesses its own advantages and disadvantages. The construction of a unified model was undertaken to ensure the segmentation of customers into three groups was appropriate.

The subsequent phase of the study involved the development of three logistic regression models, with each model dedicated to a specific customer segment. The objective of this phase was to identify patterns and significant variables that influence the probability of subscribing to the e-newsletter. The low values of Akaike's information criterion and Bayesian information criterion used to compare the models indicate that the models are highly fit.

The estimates of model accuracy, which are 0.8910 for the first model and 0.878 and 0.7158 for the second and third models, respectively, suggest that the developed models are of good quality.

The results of the logistic regression will allow to optimise the advertising campaign by efficiently allocating resources, giving priority to customer segments with a higher predicted probability of subscription.

Keywords: data analysis, online advertising, cluster analysis, logistic regression, marketing research, modeling.

Стаття надійшла до редакції 06.11.2024

Прийнята до друку 29.01.2025