Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
https://doi.org/10.36930/40330510
Article received 11.10.2023
Article accepted 26.10.2023
UDC 004.048
Correspondence author: B. M. Boyko b.boiko81@gmail.com
ISSN 1994-7836 (print)
ISSN 2519-2477 (online)

С. В. Олійник, Б. М. Бойко
Національний університет "Львівська політехніка", м. Львів, Україна

ПОРІВНЯННЯ МЕТОДІВ КЛАСТЕРИЗАЦІЇ ДЛЯ СТВОРЕННЯ ЦІЛЬОВИХ ГРУП КЛІЄНТІВ У НАБОРІ ДАНИХ

Розглянуто особливості використання методів кластеризації для створення цільових груп клієнтів. На основі аналізу літературних джерел визначено, що цільова маркетингова кампанія є ефективним способом досягнення конкретних цілей, таких як збільшення продажів, підвищення лояльності клієнтів або залучення нових споживачів. Проаналізовано методи кластеризації, які можна використовувати для створення цільових груп клієнтів. Досліджено ефективність різних методів кластеризації для створення цільових груп клієнтів. Проаналізовано дані про споживачів одного з українських підприємств. За результатами цього аналізу створено кілька кластерів клієнтів, які відрізнялися один від одного за такими параметрами, як вік, стать, місце проживання, рівень доходу та інші. Встановлено, що застосовані методи кластеризації доповнюють цільову маркетингову кампанію для досягнення конкретних маркетингових цілей. Доведено ефективність методу кластеризації, що залежить від специфіки досліджуваного бізнесу та наявних даних. Встановлено індивідуалізовані маркетингові стратегії для кожної цільової групи. Доведено, що моніторинг та аналіз результатів кампаній є ключовими для визначення ефективності і корегування їхньої діяльності. Досліджено ефективності методів кластеризації та цільових маркетингових кампаній, що допомагають оптимізувати витрати на маркетинг і підвищити результативність. Доведено сприяння у підвищенні задоволеності клієнтів. Встановлено важливість продовження дослідження та аналізу для адаптації стратегій до змін у споживацькому ринку та збереження конкурентних переваг. Використано кластеризацію для створення цільових груп клієнтів і застосовано потужний інструмент для розвитку ефективних маркетингових стратегій. За результатами дослідження з'ясовано, що застосування методів кластеризації для створення цільових груп клієнтів може підвищити ефективність маркетингових кампаній. Доведено ефективність застосування конкретних груп клієнтів для досягнення цілей кампанії, які спрямовані на широку аудиторію. Обґрунтовано дослідницький підхід щодо пристосування стратегій до змін у споживацькому ринку, що є важливим аспектом у динамічному бізнес-середовищі.

Ключові слова: кластеризація; маркетинг; цільова група; кластери; ефективність.

Вступ / Introduction

У сучасному світі, коли конкуренція між компаніями посилюється, бракує традиційних методів залучення та утримання клієнтів. Ринок переповнений схожими продуктами та послугами, і щоб виділитися, компаніям доводиться дізнаватися про своїх клієнтів більше, ніж їх конкуренти. Суть полягає в тому, щоб розуміти, що саме відзначає кожного клієнта та чому вони обирають один продукт перед іншим. У такому разі машинне навчання, зокрема методи кластеризації, можуть стати дієвим інструментом у пошуках відповідей [14, 15].

Сучасний маркетинг не може обійтися без важливого елемента – цільових кампаній. Ці стратегічні заходи спрямовані на точне направлення індивідуалізованих повідомлень до конкретних сегментів клієнтської аудиторії, цим самим максимізуючи результативність рекламних зусиль. Для досягнення цієї мети маркетологи активно використовують різні стратегії, які фокусуються на характеристиках і особливостях цільового аудиторного сегмента [1, 3].

Штучний інтелект набуває щораз більшої значущості у сфері цільового маркетингу. Алгоритми машинного навчання мають здатність аналізувати величезні обсяги клієнтських даних, що дає змогу маркетологам визначати закономірності та точно прогнозувати майбутню поведінку споживачів. Цей підхід, що ґрунтується на аналізі даних, дає маркетологам змогу створювати витончені цільові кампанії, які доставляють релевантні повідомлення вчасно [6, 10].

Об'єкт дослідження – кластеризація даних для створення таргет груп клієнтів.

Предмет дослідження – методи і засоби створення таргет груп клієнтів і порівняння ефективності використання цих груп у бізнесі.

Мета роботи – визначити найефективніший метод кластеризації даних для створення таргет груп клієнтів і порівняння ефективності їх використання в бізнесі.

Для досягнення зазначеної мети визначено такі основні *завдання дослідження*:

Інформація про авторів:
Олійник Святослав Васильович, студент, кафедра системи штучного інтелекту. **Email:** sviatoslav.oliinyk.knm.2020@lpnu.ua; https://orcid.org/0009-0002-0011-342X
Бойко Богдан Михайлович, аспірант, кафедра менеджменту і міжнародного підприємництва. **Email:** b.boiko81@gmail.com; https://orcid.org/0009-0002-9850-2764

Цитування за ДСТУ: Олійник С. В., Бойко Б. М. Порівняння методів кластеризації для створення цільових груп клієнтів у наборі даних. Науковий вісник НЛТУ України. 2023, т. 33, № 5. С. 77–83.
Citation APA: Oliinyk, S. V., & Boyko, B. M. (2023). Comparison of clustering methods for creating target customer groups in a set of data. *Scientific Bulletin of UNFU*, 33(5), 77–83. https://doi.org/10.36930/40330510

- дослідити різні методи кластеризації;
- зібрати та проаналізувати вхідні дані про клієнтів;
- використати різні методи кластеризації для створення таргет груп клієнтів;
- порівняти ефективність кожного методу та визначити найкращий.

Аналіз останніх досліджень та публікацій. Проаналізувавши різноманітні матеріали у сфері сегментування клієнтів на окремі цільові групи, можна зробити висновок, що найпопулярнішим і найефективнішим є K-Means метод кластеризації [2]. Він легко масштабується до оброблення великих обсягів даних, надзвичайно простий у реалізації та зрозумілий під час використання, що важливо для будь-якого бізнесу.

У роботі 2021 р. "Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation" із журналу "Science and Education Publishing" [1] розглянуто різні методи, однак за результатами експериментів із використанням ієрархічної кластеризації K-Means дає кращі результати.

Потрібно проаналізувати різноманітні матеріали, щоб дізнатися про методи вирішення описаної проблеми та визначити, що можна додати або вилучити для покращення кінцевого результату.

Зокрема, у роботі "Customer Segmentation using Machine Learning [14] обґрунтовано важливість управління взаєминами з клієнтами у наданні кращих послуг. Групування клієнтів, досягнуте за допомогою методів кластеризації, дає змогу визначити потенційно важливі групи клієнтів на ринку та отримати вигоду. Наведено різні підходи до кластеризації, такі як: K-Means, ієрархічна, на основі щільності та за спорідненістю – для того, щоб сегментувати клієнтів і відповідно до результатів застосовувати різні маркетингові стратегії. Зазначено також параметри кластеризації, такі як: географічні, демографічні, психографічні та поведінкові. Автори [14] обговорюють переваги та недоліки методів кластеризації. Зроблено висновок, що комбінований підхід до створення алгоритму може стати в нагоді в деяких ситуаціях.

У роботі [17] наведено набір даних, який отримано з облікових записів бази даних компанії, що здійснює роздрібну торгівлю спортивних речей у Туреччині. Запропонована методологія дослідження містить три основні кроки: попередні аналізи, такі як: очищення та трансформація даних, аналіз даних за допомогою RFM, двоетапний кластерний аналіз і кластеризація K-Means та подання результатів. У цьому дослідженні обґрунтовано важливість сегментації клієнтів для компаній, особливо для тих, хто працює в галузі роздрібної торгівлі, щоб краще зрозуміти характеристики, поведінку та демографічні характеристики своїх клієнтів. Автори зазначають, що простого групування клієнтів на підставі їхніх витрат недостатньо і рекомендують дві моделі сегментації клієнтів. Перша модель використовує двоетапну процедуру кластеризації, тоді як друга – кластеризацію K-Means.

У роботі [4] досліджено питання групування клієнтів для бізнесу за допомогою методів машинного навчання, зокрема алгоритму кластеризації K-Means. Групування клієнтів здійснюють за подібними характеристиками на ринку, таких як: зацікавленість у продукті, місце розташування та поведінка. Сегментація клієнтів охоплює індивідуальні маркетингові плани, підтримку бізнес-рішень, ідентифікацію окремих продуктів і способи управління попитом і потужністю пропозиції тощо. У цьому дослідженні подано моделювання та аналіз даних, коли варіації всередині кластера мінімізовані, а також візуалізацію набору даних за допомогою matplotlib і seaborn, щоб зрозуміти зв'язок між стовпцями. Автори дійшли висновку, що сегментація споживачів має вирішальне значення для будь-яких компаній, а алгоритм K-means може бути ефективним інструментом для її досягнення.

У дослідженні [15] наведено приклад використання аналізу кластеризації в сегментації ринку та споживачів. Автори застосували K-Means та алгоритми ієрархічної кластеризації до набору даних і з'ясували, що K-Means дають кращі результати. Вони також використовували аналіз головних компонент (PCA), щоб зменшити розмірність набору даних і покращити результати кластеризації. Автори виявили додатковий кластер клієнтів із PCA та оновили оптимальне значення K для K-Means. Також у цій роботі автори наголошують на важливості попереднього оброблення та масштабування даних, а також на використанні кількох алгоритмів кластеризації для порівняння отриманих результатів.

У роботі [10] висвітлено питання використання кластеризації для аналізу клієнтів малих і середніх компаній. Алгоритм K-Means використовується для групування даних клієнтів у кластери на підставі подібності. Для визначення оптимальної кількості кластерів використовується Elbow метод. Автори зазначають, що групування клієнтів за допомогою кластеризації може допомогти малим і середнім підприємствам зрозуміти свою клієнтську базу й ухвалювати обґрунтовані рішення щодо своїх продуктів і маркетингових стратегій. У цій роботі наголошено на важливості профілювання клієнтів для визначення характеристик і поведінки клієнтів. Остаточна мета полягає в покращенні взаємин з клієнтами та забезпеченні обслуговування на підставі їхніх потреб і вподобань.

У роботі [7] наведено детальний огляд сегментації клієнтів на основі спільних характеристик, таких як: демографічні, географічні, психографічні та поведінкові. У роботі описано метод сегментації клієнтів на підставі збирання і попереднього оброблення даних про них, використання алгоритмів кластеризації, таких як K-Means, для групування клієнтів із подібними характеристиками та аналізу отриманих кластерів для розроблення гіпотез і розуміння для маркетингових кампаній. Тут також обґрунтовано переваги алгоритму кластеризації K-Means для сегментації клієнтів, зокрема масштабованість, ефективність та можливість інтерпретації. Окрім цього, наголошено на важливому сегментації клієнтів в ухваленні бізнес-рішень і підтримці ідентифікації та спрямуванні на потенційних клієнтів. Зазначено, що групування може допомогти підприємствам визначити можливості продукту та керувати пропозицією та попитом на них.

Результати дослідження та їх обговорення / Research results and their discussion

У сучасному світі щороку посилюється конкуренція між компаніями [2, 5]. Старі механізми для залучення та утримання клієнтів стають неефективними, оскільки ринок насичений подібними продуктами та послугами. Щоб вирішити цю проблему, компанії повинні знати своїх клієнтів краще, ніж конкуренти, а також розуміти,

що саме їм потрібно та чому вони обирають той чи інший продукт. Тому кластеризацію в машинному навчанні розглядають як один з механізмів для сегментації цільової аудиторії з подібними поведінкою та характеристиками [4, 11]. Цей аналіз спрямований на огляд поточного стану розвитку стратегій цільових кампаній на підприємствах з урахуванням кластеризації машинного навчання.

Сформулюємо математичну постановку завдання розподілу клієнтів на цільові групи: маючи набір даних X, що складається з n клієнтів, представлених у вигляді векторів ознак $x_i \in R^d$, де метою є групування клієнтів у K окремих кластерів. Кожен кластер повинен охоплювати клієнтів, які демонструють подібні характеристики, поведінку чи переваги. Це математичне формулювання дає змогу застосовувати алгоритми кластеризації для автоматичного визначення базових структур і шаблонів у наборі даних клієнта.

Розглянемо методи, які використані у наукових статтях для визначення таргет груп клієнтів. Вони містять такі алгоритми кластеризації, як K-Means, ієрархічну, на основі щільності та за спорідненістю.

Для аналізу використали публічний набір даних з ресурсу Kaggle [9], який містить інформацію про інтернет-торгівлю (рис. 1).

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	09-12-2011 12:50	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	09-12-2011 12:50	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	09-12-2011 12:50	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	09-12-2011 12:50	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	09-12-2011 12:50	4.95	12680.0	France

Рис. 1. Приклад виводу набору даних про інтернет-торгівлю / Example of the output of an internet sales dataset

Набір даних, який містить інформацію про понад 500 000 транзакцій з онлайн-магазину, є цінним ресурсом для проведення аналізу сегментації клієнтів. У ньому є інформація про: дату купівлі, придбані товари, ціну кожного товару та країну, де перебуває клієнт. Важливо зазначити, що цей набір даних може потребувати певного очищення та перетворення, перш ніж його можна буде використовувати для проведення експериментів. Це може бути видалення дублікатів, робота з відсутніми значеннями та стандартизація змінних, щоб забезпечити їх порівняння між клієнтами.

Вибір ознак. Результати експериментів дадуть змогу зрозуміти придатність кожного методу кластеризації для створення цільових груп клієнтів і допоможуть компаніям вивчити свою клієнтську базу та відповідно адаптувати свої маркетингові стратегії. Отримані кластери можуть надати цінну інформацію та розуміння базових закономірностей і характеристик у даних. Розглянемо функції параметрів, які будуть використані для кластеризації:

- *Revenue* – кластеризація на основі доходу може допомогти визначити різні групи клієнтів на основі їхніх моделей витрат. Це дасть змогу виявити сегменти клієнтів, які роблять значний внесок у дохід, а також тих, хто має нижчий рівень витрат. Цю інформацію можна використовувати для цільових маркетингових стратегій, зусиль щодо утримання клієнтів або персоналізованих рекомендацій.
- *UniqueItems* – кластеризація на основі кількості придбаних унікальних предметів може надати уявлення про різноманітність уподобань продуктів серед клієнтів. Це може допомогти ідентифікувати кластери клієнтів, які, зазвичай, купують широкий асортимент продуктів, і тих, хто має більш цілеспрямовані або спеціалізовані вподобання. Ця інформація може бути цінною для планування асортименту продукції.

Frequency – кластеризація на основі частоти покупок може окреслити різні рівні залучення або лояльності серед клієнтів. Це дасть змогу ідентифікувати кластери клієнтів, які часто роблять покупки, що свідчить про високу зацікавленість або лояльність до бренду, а також кластери клієнтів із меншою частотою покупок. Ця інформація може бути корисною для управління взаєминами з клієнтами, програм лояльності та цільових маркетингових кампаній.

Проведемо порівняльний аналіз таких методів:

- K-Means (англ. *K-Means Clustering*) – кластеризація методом *k*-середніх, означає впорядкування множини об'єктів у порівняно однорідні групи;
- Ієрархічна кластеризація (англ. *Hierarchical Cluster Analysis*, HCA), полягає в видобуванні даних і статистиці – метод кластерного аналізу, який намагається побудувати ієрархію кластерів;
- DBSCAN (англ. *Density-Based Spatial Clustering of Applications with Noise*) – одним з найпоширеніших алгоритмів кластеризації, а також найбільш цитованим у науковій літературі. Заснований на щільності, а саме: для заданої множини точок у деякому просторі він відносить в одну групу точки, які розташовані найбільш щільно (точки з багатьма сусідами) та розмічає точки, які знаходяться в областях з невеликою щільністю (чиї сусіди розташовані занадто далеко) як викиди;
- Affinity Propagation – у статистиці та інтелектуальному аналізі даних означає розповсюдження спорідненості – алгоритм кластеризації, заснований на концепції "передачі повідомлень" між точками даних.

Порівняльний аналіз різних методів кластеризації може бути корисним для вибору найкращого методу в конкретному випадку (таблиця) [8, 12].

Обираючи метод кластеризації, важливо враховувати особливості застосованого набору даних, конкретні вимоги і мету аналізу. Іноді комбінація різних методів або налаштування параметрів може дати кращі результати. Найефективніший метод буде залежати від вашого конкретного випадку.

Тому проведемо комплекс експериментів за допомогою обраного набору даних для отримання груп клієнтів і оцінення виконання кожного алгоритму.

Таблиця. Порівняльний аналіз методів кластеризації / Comparative analysis of clustering methods

Метод	Тип	Переваги	Недоліки
K-Means	Partitioning method	Простий і швидкий у реалізації. Добре працює для наборів даних з великою кількістю кластерів.	Потребує попереднього визначення кількості кластерів (K). Чутливий до початкового розташування центроїдів. Може давати різні результати за різних початкових припущень.
Ієрархічна кластеризація	Agglomerative (злиття) або Divisive (розділення)	Не потребує попереднього визначення кількості кластерів. дає змогу створювати ієрархію кластерів.	Має більший обчислювальний обсяг, особливо для великих наборів даних. Може бути менш ефективним для великих наборів даних.
DBSCAN	Density-based method	Автоматично визначає кількість кластерів. Добре розрізняє кластери великої та нерівномірної щільності. Може ігнорувати шумові точки.	Чутливий до параметрів, таких як радіус і мінімальна кількість точок в околі. Може бути менш ефективним для кластеризації великих даних з однорідними кластерами.
Affinity Propagation	Message Passing method	Автоматично визначає кількість кластерів і репрезентативні точки (exemplars). Добре працює для даних зі складною структурою і шумом.	Вимагає більше обчислювальних ресурсів. Вибір метрики схожості та параметрів може бути важливим завданням.

Перед початком експериментів потрібно підготувати дані для кластеризації. Видалимо всі NULL значення з набору даних та приведемо типи даних до тих, що відповідають значенням у колонках.

Експеримент № 1. Після очищення набору даних обчислюємо customer_features датафрейм, у якому будуть такі ознаки: Revenue, UniqueItems та Frequency. На їх основі будемо проводити кластеризацію за допомогою K-Means.

Після цього нормалізуємо датафрейм customer_features за допомогою функції StandardScaler() з бібліотеки Scikit-learn. Спочатку створюємо примірник масштабувальника, а потім підганяємо його до даних та трансформуємо за допомогою методу fit_transform() [13, 16].

Для того, щоб застосувати метод K-Means, спочатку потрібно вказати кількість кластерів. Для цього використовуємо Elbow метод, щоб визначити оптимальну кількість кластерів для вхідного набору даних. Метод ліктя передбачає побудову суми квадратів у межах кластера (WCSS) від кількості кластерів і пошук точки "ліктя", де швидкість зменшення WCSS значно сповільнюється. Побудуємо графік для цього методу (рис. 2).

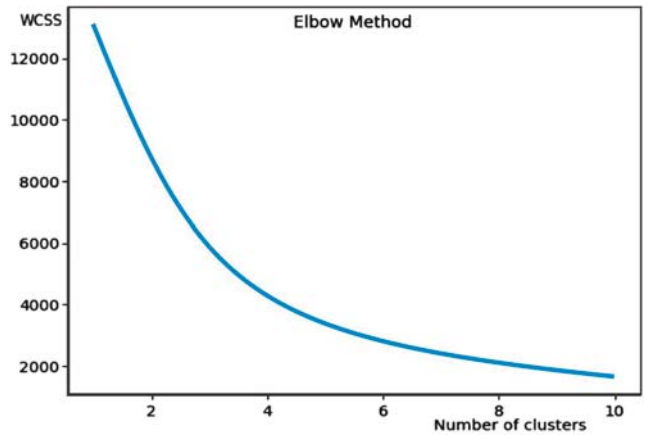

Рис. 2. Графік методу ліктя / Elbow method graph

Для цього вирішили використовувати 10 кластерів на основі результатів аналізу, щоб отримати групи з різноманітними характеристиками.

Створюємо примірник класу K-Means з n_clusters параметром, де передаємо значення з 10 кластерів. Далі застосовуємо метод fit(), куди передаємо нормалізовані дані [16].

Створимо діаграму розсіювання, щоб візуалізувати результати кластеризації у двох вимірах. Кожна точка на діаграмі розсіювання позначає клієнта, а колір – до якого кластера він належить. Використовуємо функцію scatter() із Matplotlib для створення діаграми розсіювання (рис. 3) [13].

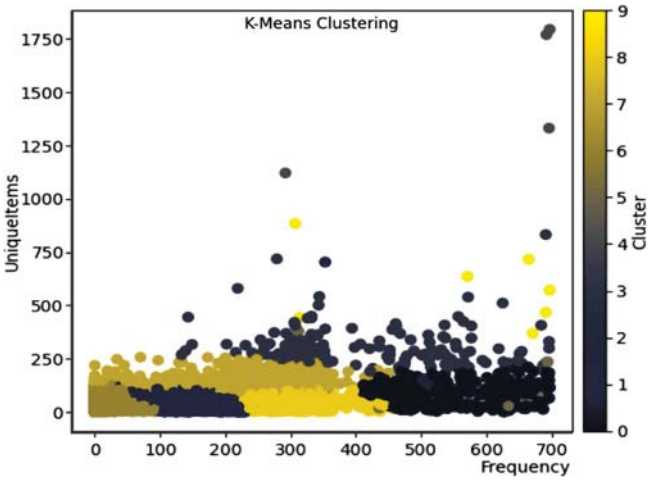

Рис. 3. Діаграма розсіювання / Scatter diagram

На графіку можна побачити точки, пофарбовані різними кольорами, які відповідають за різні кластери. Це може допомогти виявленню будь-яких закономірностей або тенденцій.

Експеримент № 2. Використаємо клас AgglomerativeClustering від Scikit-learn для виконання ієрархічної кластеризації набору даних клієнта. В агломеративній ієрархічній кластеризації метод починається з кожної точки даних як окремого кластера, а потім об'єднує найближчі кластери разом, утворюючи ієрархію кластерів. Виконаємо аналогічні кроки, як для методу K-Means, та візьмемо 10 кластерів для початкового значення.

Створимо примірник класу AgglomerativeClustering з euclidean метрикою відстані та критерієм зв'язку ward до нормалізованих ознак. Критерій зв'язування ward мінімізує дисперсію кластерів, що об'єднуються.

Створимо діаграму розсіювання, щоб візуалізувати результати кластеризації у двох вимірах. Кожна точка на діаграмі розсіювання позначає клієнта, а колір – до якого кластера він належить. Використовуємо функцію scatter() із Matplotlib для створення діаграми розсіювання (рис. 4).

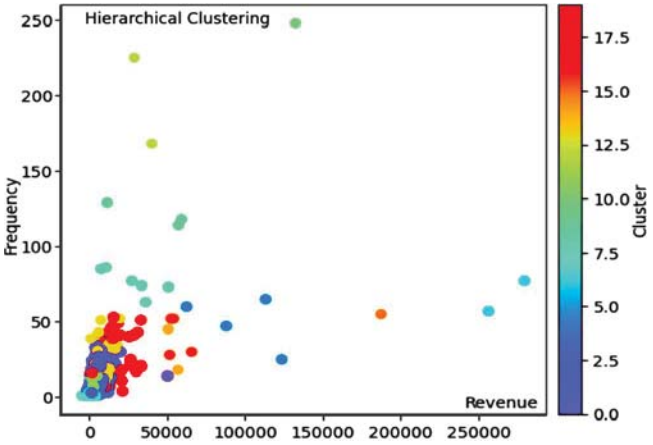

Рис. 4. Діаграма розсіювання / Scatter diagram

На графіку можна побачити кластери відносно двох ознак: Revenue та Frequency. Можемо проаналізувати кластери, обчисливши середні значення ознак для кожного. Це дасть змогу визначити шаблони та характеристики груп клієнтів.

Експеримент № 3. Після оброблення даних та приведення окремих полів до потрібних типів, було отримано df датафрейм. Далі обчислюємо customer_features датафрейм, у якому будуть такі ознаки: Revenue, UniqueItems та Frequency. На основі цих ознак будемо проводити кластеризацію за допомогою алгоритму Affinity Propagation.

Після нормалізації ознак отримуємо normalized_features датафрейм, який передаємо параметром у метод fit_predict створеного примірник класу DBSCAN зі значенням параметрів епсилон 0,5 і мінімальною кількістю вибірок 5. Після кластеризації проаналізуємо кластери, обчисливши середні значення ознак для кожного визначеного кластеру.

Щоб візуалізувати результати кластеризації, збудуємо силуетний графік, щоб їх оцінити. Силуетний графік – це метод візуалізації, який використовується для оцінювання результатів кластеризації. Графік допомагає зрозуміти, наскільки добре точки даних згруповані разом і наскільки кластери відрізняються один від одного (рис. 5).

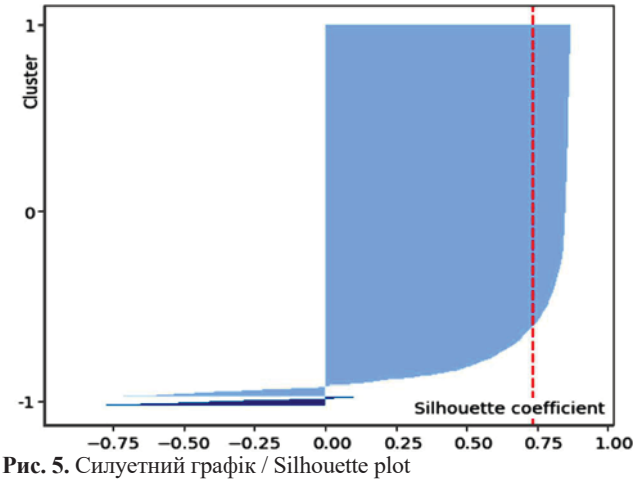

Рис. 5. Силуетний графік / Silhouette plot

Із наведеної вище візуалізації даних можна легко побачити, наскільки точка даних подібна на власний кластер порівняно з іншими кластерами.

Експеримент № 4. Проведемо аналогічні дії, проте для Affinity Propagation. Цей алгоритм кластеризації здатний ідентифікувати примірники або прототипи в наборі даних. Приклади – це точки даних, які є найбільш репрезентативними для відповідних кластерів. У цьому випадку демонструється, як застосувати Affinity Propagation для кластеризації підмножини аналізованих даних клієнтів.

Після нормалізації ознак отримуємо normalized_features датафрейм, який передаємо параметром у метод fit() примірника класу AffinityPropagation для застосування цього алгоритму. Попередньо потрібно задати параметр damping, який контролює ступінь впливу примірників на призначення точок даних кластерам. Більше значення демпфування може привести до стабільніших результатів кластеризації, але також до меншої кількості кластерів. У цьому випадку було встановлено значення демпфування на 0,95.

Після кластеризації створюємо новий cluster_data датафрейм і додаємо нову колонку Cluster, куди подаємо значення атрибуту labels_ із створеного примірник класу AffinityPropagation. Для візуалізації створюємо діаграму розсіювання (рис. 6).

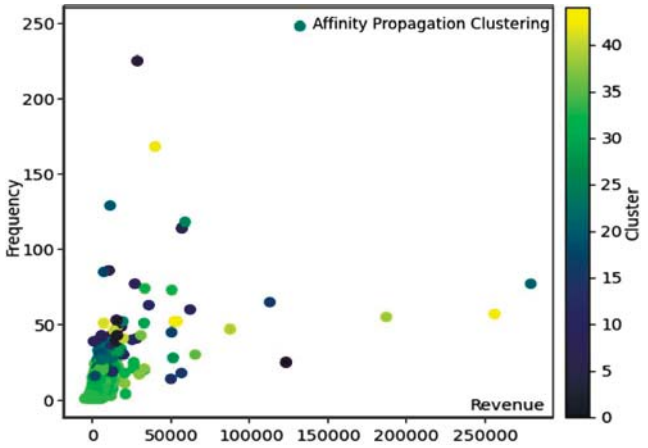

Рис. 6. Діаграма розсіювання / Scatter diagram

На графіку зображено колір кожної точки відповідно до свого кластера. Можна побачити, наскільки добре кластери розділяють точки даних на цьому графіку.

Загалом було проведено дослідження із використанням чотирьох типів кластеризації, які застосовувалися на вхідних датах, що містять інформацію про транзакції онлайн-магазину. Провівши експерименти можна порівняти ці різні підходи.

Найчастіше використовують алгоритм кластеризації для сегментації клієнтів K-Means. Для цього типу потрібна початкова кількість кластерів, що важко передбачити одразу, і це може вплинути на результат кластеризації. Проте з використанням методу ліктя можливо знайти оптимальне число кластерів для вхідного набору даних, що і було зроблено у дослідженні.

Унаслідок роботи ієрархічної кластеризації отримали кластери. Початкова кількість цих кластерів була визначена також методом ліктям. Зазначимо, що часова складність ієрархічної кластеризації є більшою, ніж за методом K-Means, тому він придатний для набору даних малого та середнього розміру.

DBSCAN, який можна використовувати для пошуку кластерів довільної форми, але він не придатний для більшої різниці щільності в точках даних та має нижчу ефективність кластерного результату.

Affinity Propagation також не потребує подання початкової кількості кластерів, а ефективність результатів кластеризації є високою. Проте час кластеризації трива-

лий, що гарантує її високу ефективність. Тому цей метод краще застосовувати до набору даних малого та середнього розміру.

Отже, остаточної відповіді на те, який алгоритм найкращий немає. Кожен з них має свої переваги та недоліки стосовно якоїсь конкретної ситуації.

Обговорення результатів дослідження. Питання сегментації клієнтів для встановлення індивідуалізованих маркетингових стратегії для кожної цільової групи висвітлено у низці наукових робіт. Зокрема, у роботі [14] розглядають набір даних, отриманих з облікових записів бази даних компанії, що здійснює роздрібну торгівлю спортивними товарами. Також зазначають важливість сегментації клієнтів для компаній, особливо тих, що працюють у сфері роздрібної торгівлі, щоб краще зрозуміти характеристики, поведінку та демографічні показники своїх клієнтів.

У дослідженні [17] обґрунтовують важливість кластеризації клієнтів для бізнесу та те, як цього можна досягти за допомогою методів машинного навчання, зокрема алгоритму кластеризації K-Means. Також подано моделювання та аналіз даних, коли варіації всередині кластера мінімізовані.

У роботі [9] досліджують використання аналізу кластеризації в сегментації ринку та споживачів. Автори [9] застосували K-Means та алгоритми ієрархічної кластеризації до набору даних і виявили, що K-Means дає кращі результати.

У роботі [11] розглядають приклад використання кластеризації для аналізу клієнтів малих і середніх компаній. Алгоритм K-Means використовується для групування даних клієнтів у кластери на основі подібності. Для визначення оптимальної кількості кластерів використовується метод Elbow.

У роботі [1] подано детальний огляд сегментації клієнтів, яка є процесом поділу клієнтської бази на різні групи на основі загальних характеристик, таких як: демографічні, географічні, психографічні та поведінкові чинники. Також описано метод сегментації клієнтів, який передбачає збирання та попереднє оброблення даних про клієнтів, використання алгоритмів кластеризації.

Завдяки поглибленому аналізу літературних джерел визначено важливість цільових кампаній для охоплення конкретних сегментів споживачів і максимізації маркетингової ефективності. Звертаючись до постановки проблеми, на меті було застосувати обраний метод кластеризації для створення таргетованих груп клієнтів.

Отже, за результатами виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – удосконалено процес вибору методу кластеризації, який найкраще підходить для розроблення цільових маркетингових кампаній та сегментації клієнтів.

Практична значущість результатів дослідження – удосконалено підходи до дослідження й аналізу набору даних, а також вивчено нові методи кластеризації для забезпечення актуальності та конкурентоспроможності компаній у сучасному бізнес-середовищі.

Роботу, в перспективі, можна розширити для вивчення методів кластеризації, які можуть обробляти складні структури даних, такі як нелінійні зв'язки або дані великої розмірності.

Висновки / Conclusions

Отже, досліджено створення цільових груп клієнтів за допомогою методів кластеризації. Завдяки поглибленому аналізу літературних джерел визначено важливість цільових кампаній для охоплення конкретних сегментів споживачів і максимізації маркетингової ефективності. Звертаючись до постановки проблеми, на меті було застосувати обраний метод кластеризації для створення таргетованих груп клієнтів.

Під час аналізу матеріалів і методів обґрунтовано основні стратегії, які використовують в цільових кампаніях, і розглянуто поточний стан досліджень у цій галузі. Наголошено на важливості методів кластеризації як потужного інструменту для ідентифікації груп клієнтів на основі подібності та шаблонів у їхніх характеристиках і поведінці. Окрім цього, досліджено допоміжні методи аналізу даних, які можуть підвищити точність та надійність результатів кластеризації. Здійснено порівняння різних методів кластеризації для створення таргет груп клієнтів і їх оцінювання. Експериментально оцінено методи з використанням на заданому наборі даних. Досліджено ефективність і придатність різних алгоритмів кластеризації для застосування у сценаріях реального світу.

За результатами дослідження зроблено такі висновки:

- Встановлено, що кожен метод кластеризації має свої переваги та недоліки стосовно якоїсь конкретної ситуації. Однак вибір правильного методу залежить від багатьох чинників і потребує ретельного аналізу і вивчення кожної ситуації окремо.
- Застосовано експерименти з використанням різних методів, внаслідок виконання яких створено групи клієнтів із початкових даних. З'ясовано, що найдоцільніше використовувати алгоритми Affinity Propagation або ж алгоритм K-Means, які дали найкращі результати.

References

1. Abdulhafedh, A. (2021). Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation. *Journal of City and Development*, 3(1), 12–30. https://doi.org/10.12691/jcd-3-1-3
2. Agrawal, S. (2021). Introduction to Matplotlib and Seaborn. URL: https://medium.com/analytics-vidhya/introduction-to-matplotlib-and-seaborn-e2dd04bfc821
3. Ajay, M. S. (2020). Data-Preprocessing with Python. URL: https://medium.com/dataseries/data-preprocessing-with-python-3914d3e9dd30
4. Alzahrani, L. A. (2021). Customer Segmentation: Unsupervised Machine Learning Algorithms in Python. URL: https://towardsdatascience.com/customer-segmentation-unsupervised-machine-learning-algorithms-in-python-3ae4d6cfd 41d
5. Archana, K., & Saranya, K. G. (2019). Mall Customer Segmentation Using Clustering Algorithm. *International Journal of Multidisciplinary Educational Research* (IJMER), 8(6), 94–99. https://doi.org/10.22214/ijraset.2023.48368
6. Christopher, A. (2021). Hierarchical Clustering and Density-Based Spatial Clustering of Applications with Noise. URL: https://medium.com/mlearning-ai/hierarchical-clustering-and-density-based-spatial-clustering-of-applications-with-noise-dbscan-b8d903095532
7. Hua, N., & Leu, B. A. (2022). Machine learning-based customer segmentation. International Journal of Information Security, 9(7).
8. John, J. M., Shobayo, O., & Ogunleye, B. (2023). An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. Analytics, 2, 809–823. https://doi.org/10.3390/analytics2040042
9. Kaggle. Online Retail. (2023). URL: https://www.kaggle.com/datasets/hellbuoy/online-retail-customer-clustering?resource=download

10. Khyati, Mahendru. (2019). How to Determine the Optimal K for K-Means? International Journal of Advance Research in Computer Science and Management Studies, 1(6), 90–95. URL: https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb
11. Kulkarni, S. (2022). Advance Customer Segmentation. International Journal of Creative Research Thoughts (IJCRT), 10(1), 114–119. URL: https://www.academia.edu/100683793/Advance_Customer_Segmentation
12. Lorbeer, B., Kosareva, A., Deva, B., Softić, D., Ruppel, P., & Küpper, A. (2018). Variations on the clustering algorithm BIRCH. Big Data Res, 11, 44–53. https://doi.org/10.1016/j.bdr.2017.09.002
13. Pandas Python library. (2023). URL: https://pandas.pydata.org/docs/getting_started/index.html
14. Pyla, S. D., & Seshashayee, Dr. M. (2022). Customer segmentation using machine learning. *International Research Journal of Modernization in Engineering Technology and Science*, 4(5). URL: https://www.irjmets.com/uploadedfiles/paper/issue_5_may_2022/23752/final/fin_irjmets1653303840.pdf
15. Sagar, A. (2019). Customer Segmentation Using K Means Clustering. URL: https://towardsdatascience.com/customer-segmentation-using-k-means-clustering-d33964f238c3
16. Scikit Learn Python library. (2023). URL: https://scikit-learn.org/stable/modules/clustering.html#clustering
17. Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. IOP Conf. Series: Materials Science and Engineering, l(336). https://doi.org/10.1088/1757-899X/336/1/012017

S. V. Oliinyk, B. M. Boyko

Lviv Polytechnic National University, Lviv, Ukraine

COMPARISON OF CLUSTERING METHODS FOR CREATING TARGET CUSTOMER GROUPS IN A SET OF DATA

The study examines the use of clustering techniques to create target customer groups. Based on the analysis of literary sources, a targeted marketing campaign is determined to be an effective way to achieve specific goals, such as increasing sales, increasing customer loyalty, or attracting new consumers. Clustering is supposed to be the method that can be used to create target customer groups. Therefore, the purpose of the work is to investigate the effectiveness of various clustering methods for creating target customer groups. In the course of the study, the analysis of data on consumers of one of the Ukrainian enterprises was carried out. Based on this analysis, several clusters of customers were created, which differed from each other by such parameters as age, gender, place of residence, income level, and others. It has been established that the applied clustering methods complement the target marketing campaign to achieve specific marketing goals. The effectiveness of the clustering method has been demonstrated, which depends on the specificity of the studied business and available data. Individualized marketing strategies have been established for each target group. It has been proven that monitoring and analysing campaign results are crucial for assessing effectiveness and making adjustments to their activities. The effectiveness of clustering methods and target marketing campaigns has been examined, which enable optimizing marketing expenses and enhance performance. Promotion in increasing customer satisfaction is demonstrated. The importance of continuing research and analysis to adapt strategies to changes in the consumer market and maintain competitive advantages has been established. Clustering has been utilized to create target customer groups and applied as a powerful tool for the development of effective marketing strategies. The results of the study showed that the use of clustering methods to create target customer groups can lead to an increase in the effectiveness of marketing campaigns. This is due to targeted campaigns that target specific groups of customers are more effective than campaigns that target a broad audience. The research approach has established the significance of adapting strategies to changes in the consumer market, which is a vital aspect in a dynamic business environment.

Keywords: clustering; marketing; target group; clusters; efficiency.