КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій

Кафедра технологій управління

Спеціальність 122 – Комп'ютерні науки, освітня програма «Інформаційна аналітика та впливи»

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА на тему:

"Модель сегментації клієнтів як інструмент підвищення якості маркетингових кампаній"

Студента 2-го курсу групи ІАВ-21:	**Науковий керівник:**
Рудківського Євгеній Віталійовича (прізвище, ім'я, по батькові)	д-р. техн. наук, професор (науковий ступінь, вчене звання) Заріцький Олег Володимирович (прізвище, ім'я, по батькові)

(підпис студента) (дата) (підпис)

Попередній захист:

(Висновок: «До захисту в Екзаменаційній комісії»)

Завідувач кафедри
технологій управління _____ _____ _____

(підпис) (прізвище, ініціали) (дата)

Київ – 2025

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій

Кафедра технологій управління
Освітньо-кваліфікаційний рівень Магістр
Спеціальність 122-Комп'ютерні науки
Освітня програма Інформаційна аналітика та впливи

ЗАТВЕРДЖУЮ
Завідувач кафедри
професор Морозов В.В.
_____
«____» _____20___року

ЗАВДАННЯ
НА ВИКОНАННЯ КВАЛІФІКАЦІЙНОЇ РОБОТИ

Студент Рудківський Євгеній Віталійович

Група ІАВ-21

1. Тема кваліфікаційної роботи
Модель сегментації клієнтів як інструмент підвищення якості маркетингових кампаній

Затверджена наказом по від «___»_____20__ р. № ___.
2. Строк подання студентом готової роботи – "19" ___05__2025р.

3. Цільова установка та вихідні дані до роботи
Цільова установка: підвищення ефективності якості маркетингових кампаній за допомогою методів сегментації.
Вихідні дані: кластери клієнтів і ціна пропозиція для них .

4. Зміст роботи
1. Провести аналіз літературних джерел та існуючих технологічних рішень .
(перелік питань, що підлягають розробці)
2. Провести формалізацію методів аналізу даних для проєкту _
3. Розробити моделі Data Science для прогнозування кластерів клієнтів та оцінити їх якість.

4. Розробити інформаційну технологію для прогнозу клістерів клієнтів для підвищення якості маркетингових компаній.

5. Перелік графічного матеріалу (слайдів)

актуальність (2 слайди), об'єкт, предмет, мета та задачі дослідження (2 слайди), наукова новизна і практична цінність (1 слайд), побудова моделі та оцінка результатів (5 слайдів), інформаційна система (1 слайди), оцінка ефективності і перспективи впровадження (2 слайди), висновки (1 слайд).

6. Календарний план виконання роботи:

№ п/п	Назва частин роботи	%	Виконання роботи	
			За планом	Фактично
1.	Вибір теми дипломної роботи	3	01.10.24	01.10.24
2.	Протокол кафедри ТУ про затвердження тем дипломних робіт та призначення наукових керівників	2	27.12.24	27.12.24
3.	Складання розгорнутого плану кваліфікаційної роботи	5	18.01.25	19.01.25
4.	Ознайомлення з іноземною та вітчизняною літературою та основними поняттями за темою наукового дослідження	10	05.02.25	07.02.25
5.	Ознайомлення наукового керівника з розгорнутим планом кваліфікаційної роботи. Внесення змін	5	09.02.25	09.02.25
6.	Підготовка розділу 1 «аналіз основ використання сегментації для підвищення якості у маркетингових компаніях»	10	20.02.25	20.02.25
7.	Підготовка розділу 2 «Методи та методики сегментації, які підвищують якість маркетингових компаній»	15	12.03.25	14.03.25
8.	Підготовка розділу 3 «моделювання продукту сегментації для підвищення якості маркетингових кампаній»	15	01.04.25	01.04.25

9.	Підготовка розділу 4 «Практична реалізація запропонованого методу та оцінка якості в маркетингових компаніях»	14	17.04.25	17.04.25
10.	Оформлення кваліфікаційної роботи. Підготовка презентації	15	24.04.25	25.04.25
11.	Передача кваліфікаційної роботи науковому керівникові	2	26.04.25	26.04.25
12.	Передача кваліфікаційної роботи рецензенту для рецензування	2	01.05.25	01.05.25
13.	Попередній захист кваліфікаційної роботи	5	10.05.25	10.05.25

Дата видачі завдання «____» _____ 20__ р.

Керівник роботи _д-р. техн. наук, професор_Заріцький Олег Володимирович
(посада, прізвище, ім’я, по батькові)

_____
(підпис)

Завдання прийняв до виконання студент групи ІАВ-21

__Рудківський Євгеній Віталійович_____
(прізвище, ім’я, по батькові)

_____
(підпис)

ЗМІСТ

АНОТАЦІЯ

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій

Кафедра технологій управління

Спеціальність 122 - Комп’ютерні науки,

Освітня програма “Інформаційна аналітика та впливи”

Дипломна робота магістра Рудківського Євгеній Віталійовича.

Тема роботи – «Модель сегментації клієнтів як інструмент підвищення якості маркетингових кампаній».

Мета дипломної роботи магістра – підвищення ефективності маркетингових компаній за допомогою збільшення якості сегментації.

Об’єкт дослідження – це процес сегментації клієнтів як інструмент підвищення ефективності маркетингових кампаній.

Предмет дослідження – моделі, методи та технології Data Science для підвищення ефективності маркетингових компаній.

Наукова новизна даної роботи полягає у створенні нової метрики, яка на відміну від інших метрик дозволяє оцінити якість моделі сегментації, навіть при невисоких результатах класичних метрик машинного навчання для кластеризації.

У роботі досліджуються існуючі підходи до використання аналітичних методів у задачах клієнтів для підвищення якості маркетингових компаній. Розробляється методика їх використання і нове оцінювання моделей, а також проводиться обґрунтування доцільності та необхідності впровадження запропонованої методики. Наводяться рекомендації щодо практичної імплементації методики.

Дипломна робота складається зі вступу, основної частини, яка включає чотири розділи, висновків та списку використаних джерел. Всього налічує 114 сторінок та перелік посилань з 45 джерел на 4 сторінках.

Ключові слова: сегментація, кластеризація, Data Science, маркетинг.

Перелік використаних скорочень та термінів

CRM – (Customer Relationship Management) – управління взаємовідносинами з клієнтами; програмне забезпечення або стратегія для покращення обслуговування та аналізу поведінки споживачів.

RFM-аналіз – (Recency, Frequency, Monetary analysis) – метод сегментації клієнтів за трьома параметрами: давністю останньої покупки, частотою покупок та сумою витрачених коштів.

K-Means – один із найпопулярніших алгоритмів кластерного аналізу, який використовується для групування схожих об’єктів (у цьому випадку – клієнтів) за певними характеристиками.

Сегментація ринку – процес поділу цільової аудиторії на відносно однорідні групи за спільними характеристиками, такими як демографія, поведінка, інтереси тощо.

Маркетингова кампанія – сукупність заходів, спрямованих на просування продукту чи послуги з використанням комунікаційних інструментів.

Таргетинг – процес спрямування маркетингових зусиль на конкретні сегменти клієнтів з урахуванням їхніх інтересів, потреб та поведінки.

Кластер – група об’єктів (наприклад, клієнтів), що мають схожі характеристики або поведінкові патерни.

Аналітика клієнтів – аналіз даних про клієнтів з метою виявлення закономірностей, що допомагають покращити прийняття бізнес-рішень.

Якість маркетингової кампанії – ефективність кампанії, яка вимірюється через досягнення цільових показників: рівень конверсії, продажів, залучення тощо.

Silhouette Score – показник, який оцінює, наскільки об’єкти схожі на власний кластер у порівнянні з іншими: значення в діапазоні від -1 до 1, де вищі значення означають кращу кластеризацію.

PCA (Principal Component Analysis) – метод зменшення розмірності даних, який дозволяє зберегти найважливішу інформацію при зниженні кількості ознак.

Agglomerative Clustering – ієрархічний метод кластеризації, що починається з індивідуальних точок і поступово об’єднує їх у групи (кластери).

Dendrogram – графічне зображення результатів ієрархічної кластеризації, що показує, як об'єкти були об'єднані на кожному етапі.

Elbow Method – метод визначення оптимальної кількості кластерів, заснований на аналізі зміни внутрішньої дисперсії в залежності від кількості кластерів.

Z-нормалізація – стандартизація даних шляхом приведення їх до середнього 0 і стандартного відхилення 1.

One-hot encoding – метод кодування категоріальних змінних у вигляді бінарних векторів.

Python – мова програмування, що часто використовується для аналізу даних і побудови моделей кластеризації.

scikit-learn – популярна бібліотека Python для машинного навчання, що включає алгоритми кластеризації, метрики, і засоби попередньої обробки.

ВСТУП

У сучасному світі бізнесу, де інформаційні потоки стали надзвичайно насиченими, а увага споживача — обмеженим ресурсом, компаніям доводиться шукати нові, більш точні та персоналізовані підходи до комунікації зі своєю аудиторією. Раніше ефективною могла бути масова реклама, орієнтована на всіх одразу. Проте сьогодні цей підхід дедалі більше втрачає свою дієвість. Клієнти очікують, що бренди розумітимуть їхні потреби, цінності, стиль життя та емоційні запити. Відтак у центрі уваги сучасного маркетингу постає поняття сегментації клієнтів — процесу поділу споживачів на групи за спільними характеристиками з метою формування ефективних і цільових маркетингових кампаній.

Сегментація — це не просто поділ на вікові чи географічні категорії. Сучасні моделі клієнтської сегментації враховують поведінкові особливості, частоту покупок, рівень витрат, реакцію на знижки, стиль комунікації та навіть цінності, якими керується споживач. Завдяки цьому бізнес може не лише оптимізувати свої рекламні витрати, але й значно підвищити якість взаємодії з клієнтом. Адже правильно підібране повідомлення у правильний час і через зручний канал має набагато більше шансів привернути увагу, викликати інтерес і спонукати до дії.

Крім того, моделі сегментації відкривають нові можливості для прогнозування поведінки клієнтів, визначення рівня їхньої лояльності та ймовірності відтоку. Наприклад, аналізуючи дані про частоту та обсяг покупок, компанія може визначити, які клієнти є «найціннішими» — тобто приносять найбільший прибуток. Саме на таких споживачів доцільно спрямовувати основні маркетингові зусилля. У той час як іншим — пропонувати спеціальні умови, аби стимулювати активність або повернути інтерес до бренду.

Таким чином, модель сегментації клієнтів — це не лише інструмент глибокого аналітичного аналізу, а й практичний ключ до побудови ефективного, персоналізованого маркетингу. В умовах динамічного ринку та зростаючих вимог споживачів, саме здатність розуміти свою аудиторію стає вирішальною перевагою

бізнесу. І сегментація тут відіграє роль не другорядного етапу, а стратегічної основи, яка може суттєво підвищити ефективність усіх маркетингових зусиль.

У даній кваліфікаційній роботі магістра буде розглянуто розробку технології Data Science для прогнозування сегменту клієнтів, включаючи аналіз даних, використання статистичних методів та методів машинного навчання, створення моделі оцінки на основі отриманих результатів, створення системи для надсилання листів на основі сегментів.

Об'єктом даного дослідження є процес сегментації клієнтів як інструмент підвищення ефективності маркетингових кампаній.

Предметом дослідження є моделі, методи та технології Data Science для підвищення ефективності маркетингових компаній.

Метою даної роботи є підвищення ефективності маркетингових компаній за допомогою збільшення якості сегментації.

Методи дослідження. Інформаційною базою дослідження стали наукові праці вітчизняних та зарубіжних фахівців, присвячені проблематиці сегментації клієнтів, застосуванню методів кластерного аналізу, моделей поведінкової аналітики та інформаційних технологій у різних сферах. У процесі дослідження було використано результати фундаментальних і прикладних праць, аналітичні звіти, статистичні дані та публікації провідних науковців, маркетингової аналітики та машинного навчання. Додатково було враховано практичний досвід автора, отриманий під час реалізації проєкту з сегментації клієнтів у АТ КБ «ПРИВАТБАНК».

Практична частина дослідження була реалізована в межах внутрішнього аналітичного проєкту в АТ КБ «ПРИВАТБАНК». На основі побудованої сегментаційної моделі було проведено A/B-тестування маркетингових кампаній. Результати продемонстрували підвищення конверсії порівняно з аналогічними кампаніями, у яких не використовувалася сегментація. Це дозволило обґрунтувати доцільність впровадження розробленої моделі у постійну операційну діяльність

банку навіть при низьких показниках метрик кластеризації (нова метрика показала, що створену модель кластеризації можна використовувати).

Для досягнення даної мети потрібно виконати наступні *завдання*:

1. Провести аналіз моделей, методів та існуючих технологічних рішень сегментації клієнтів засобами Data Science, вивчити теоретичні класичні джерела за обраною науковою проблемою, визначити стан розробки питань обраної наукової проблеми у вітчизняній та іноземній літературі;
2. Сформувати датасет для прогнозування сегментів клієнтів;
3. Провести аналіз даних про клієнти, включаючи параметри, які характеризують його - демографічні, поведінкові;
4. Розробити модель сегментації клієнтів із використанням статистичних методів та методів машинного навчання;
5. Провести аналіз отриманих результатів та якості побудованої моделі;
6. Розробити інформаційну технологію використання моделі та визначити перспективи її використання.

Наукова новизна даної роботи полягає у створенні нової метрики, яка на відміну від інших метрик дозволяє оцінити якість моделі сегментації, навіть при невисоких результатах класичних метрик машинного навчання для кластеризації.

Практичною цінністю є інформаційної технології, яка складається з блоків з можливістю їх редагування під різні потреби бізнесів. Розроблена інформаційна технологія може стати потужним інструментом для оптимізації витрат і пошук росту, надаючі більш таргетовані пропозиції клієнту, що в результаті підвищує ROI.

РОЗДІЛ 1
АНАЛІЗ ОСНОВ ВИКОРИСТАННЯ СЕГМЕНТАЦІЇ ДЛЯ ПІДВИЩЕННЯ ЯКОСТІ У МАРКЕТИНГОВИХ КАМПАНІЯХ

1.1 Сутність та значення сегментації клієнтів

У сучасних умовах високої конкуренції та динамічного розвитку ринкових відносин ефективність маркетингових кампаній значною мірою залежить від здатності компаній ідентифікувати та розуміти свою цільову аудиторію. Одним із ключових інструментів для досягнення цієї мети є сегментація клієнтів – процес поділу споживачів на групи за певними характеристиками, що дозволяє адаптувати маркетингові стратегії відповідно до специфіки кожного сегмента.

Сегментація клієнтів є фундаментальним елементом стратегічного маркетингового управління, оскільки вона сприяє оптимізації комунікаційних процесів, підвищенню рівня задоволеності споживачів та ефективному використанню ресурсів компанії. Завдяки сегментації бізнес має змогу розробляти персоналізовані маркетингові компанії, що забезпечують більш високий і сучасний розвиток цифрових технологій та зростання обсягів доступних даних сприяють впровадженню нових методів сегментації, зокрема аналітичних підходів на основі машинного навчання. Використання алгоритмів кластерного аналізу, нейронних мереж та інших моделей штучного інтелекту дозволяє більш точно ідентифікувати приховані закономірності у поведінці споживачів та підвищувати ефективність маркетингових комунікацій.

Значення сегментації клієнтів виходить за межі суто маркетингових завдань і охоплює ширший спектр управлінських рішень. Вона є основою для розробки цінових стратегій, продуктового асортименту, каналів дистрибуції та інших елементів маркетинг-міксу. Крім того, ефективне використання сегментації сприяє

підвищенню конкурентоспроможності компанії, оскільки дозволяє їй адаптуватися до змін у споживчих уподобаннях та швидше реагувати на виклики ринку.

Таким чином, сегментація клієнтів є невід'ємним елементом сучасного маркетингового управління, що забезпечує персоналізацію маркетингових кампаній та підвищує ефективність взаємодії з аудиторією. Подальший розвиток методів сегментації, зокрема із застосуванням великих даних та штучного інтелекту, відкриває нові можливості для бізнесу в умовах цифрової трансформації економіки. конверсію та зміцнення лояльності клієнтів.

1.2 Аналіз об'єкта дослідження

Об'єктом даного дослідження є процес сегментації клієнтів як інструмент підвищення ефективності маркетингових кампаній. У сучасних умовах високої конкуренції, зростаючої кількості даних та розвитку цифрових технологій компанії змушені застосовувати більш точні методи аналізу споживчої поведінки для досягнення бізнес-цілей. Сегментація клієнтів є ключовим елементом такої аналітики, оскільки дозволяє виявити глибші закономірності в поведінці споживачів та адаптувати маркетингові стратегії до потреб різних категорій клієнтів.

Сегментація клієнтів передбачає поділ споживачів на однорідні групи на основі певних характеристик, таких як демографічні, географічні, психографічні та поведінкові фактори. Головна мета цього процесу – підвищення релевантності маркетингових комунікацій, оптимізація витрат на просування товарів і послуг, а також формування персоналізованих пропозицій, що сприяють зростанню рівня задоволеності клієнтів.

Сучасні технологічні можливості дозволяють здійснювати сегментацію на основі аналізу великих даних (Big Data), машинного навчання та алгоритмів штучного інтелекту. Це значно розширює можливості маркетологів у точному

визначенні споживчих переваг, прогнозуванні поведінки клієнтів та автоматизації маркетингових рішень.

Ефективне застосування сегментації клієнтів дозволяє бізнесу оптимізувати маркетингові стратегії, підвищити ефективність рекламних кампаній та підсилити персоналізацію взаємодії з клієнтами. Наприклад, персоналізовані пропозиції та таргетована реклама на основі поведінкових даних значно підвищують конверсію, тоді як загальні маркетингові кампанії часто мають нижчу ефективність через відсутність індивідуального підходу.

Таким чином, сегментація клієнтів є потужним інструментом маркетингової аналітики, що дозволяє не лише підвищити якість комунікацій між бізнесом та споживачами, а й забезпечити ефективне управління ресурсами компанії. Використання сучасних підходів до сегментації, включаючи аналітичні методи та алгоритми штучного інтелекту, відкриває нові можливості для вдосконалення маркетингових стратегій та досягнення конкурентних переваг на ринку.

1.3 Використання сегментації у маркетингових кампаніях

Сегментація клієнтів дозволяє компаніям оптимізувати маркетингові стратегії, роблячи їх більш персоналізованими та ефективними. Завдяки правильному поділу аудиторії можна пропонувати продукти та послуги, які найкраще відповідають потребам різних груп клієнтів [1].

Для прикладу візьмемо просту сегментація - демографічну, яка дозволяє розбити клієнтів на сегменти:

1. Молодь (18–25 років).
2. Середній вік (26–54 років).
3. Старший вік (55+ років).

Молодь - для цієї вікової категорії характерні інші потреби та фінансові звички. У рамках демографічної сегментації, маркетингові кампанії можуть бути

орієнтовані на специфічні інтереси молоді. Такі як, продукти для активного способу життя. Наприклад, компанія може пропонувати мобільні платіжні сервіси або кредитні картки з бонусами за покупки на розважальних платформах (наприклад, кіно, ресторани, кафе). Інноваційні технології та гаджети. Кампанії можуть бути спрямовані на стимулювання покупки нових гаджетів, аксесуарів або мобільних додатків з персоналізованими налаштуваннями для цієї вікової категорії. Знижки та пільги. Введення програм лояльності або знижок на освітні послуги, тренінги, курси для молоді - все це може бути корисним і залучити їх до постійного користування продуктами або послугами.

Середній вік - ця група має більш стабільний рівень доходів і часто перебуває на піку своєї кар'єри. У цьому віці люди зазвичай роблять значні фінансові кроки, такі як купівля житла, автомобілів, планування сімейного бюджету. Для цієї аудиторії можуть бути розроблені наступні маркетингові стратегії:

1. Іпотечні та автокредити - компанії можуть пропонувати спеціальні умови для придбання нерухомості або транспортних засобів, зниження ставок за умовами кредитів для цієї вікової групи, враховуючи їхні стабільні доходи та зростаючі фінансові можливості.
2. Інвестиційні продукти - пропозиції щодо інвестицій у фінансові інструменти, такі як акції, облігації чи пенсійні фонди, які дозволяють накопичувати на пенсію чи майбутнє дітей.
3. Кредити для бізнесу та самозайнятих осіб - пропозиція про кредити для малих та середніх підприємців, на фінансування власних проектів або стартапів.

Старший вік - для цієї вікової групи важливими є безпека та стабільність. У маркетингових кампаніях для цієї аудиторії акцент робиться на надійність, довіру та довгострокові гарантії. Можна застосовувати такі стратегії:

1. Депозитні програми та накопичувальні рахунки - люди цієї вікової групи часто шукають способи безпечного зберігання своїх коштів, банки можуть пропонувати їм вигідні депозитні програми з фіксованими ставками, а також послуги з накопичення на пенсію.
2. Офлайн та онлайн консультації - можна організувати персоналізовану допомогу в плануванні фінансів через консультації в офісах або в телефонному режимі, оскільки ця група може відчувати певні труднощі з онлайн-сервісами.
3. Програми підтримки здоров'я та благополуччя - спеціальні акції, пов’язані з медичними послугами, страхуванням здоров’я або пропозиції на покупку товарів для здоров'я та комфортного життя.

Сегментація клієнтів є невід’ємною складовою сучасного маркетингу, що надає компаніям значні переваги. Вона дозволяє більш точно налаштувати маркетингові стратегії, відповідаючи на потреби та інтереси різних груп споживачів. Завдяки сегментації, компанії здобувають змогу створювати персоналізовані пропозиції, які не тільки підвищують ефективність маркетингових кампаній, а й покращують взаємодію з клієнтами.

Основною перевагою сегментації є її здатність підвищувати точність орієнтації реклами та продуктів на відповідні групи клієнтів. Це дозволяє знижувати витрати на рекламні кампанії, адже маркетингові зусилля фокусується на найбільш потенційно вигідних сегментах. В результаті зростає зацікавленість у пропозиціях, що веде до збільшення продажів та покращення загальної рентабельності.

Крім того, сегментація надає компаніям можливість поглибленого розуміння потреб своїх клієнтів, що сприяє створенню більш ефективних стратегій лояльності та покращенню сервісу. Це не лише дозволяє підтримувати постійну взаємодію з клієнтами, але й формує довгострокові відносини, що надають значну конкурентну перевагу на ринку. Розглянемо, як деякі з найбільших компаній світу,

таких як Amazon, Nike та Spotify, використовують сегментацію для досягнення своїх бізнес-цілей.

Amazon є одним із найяскравіших прикладів того, як сегментація може допомогти в персоналізації покупок і підвищенні ефективності маркетингових кампаній. Всі ми знайомі з тим, як після відвідування інтернет-магазину Amazon на нас з'являються рекомендації товарів. Це не випадковість, а результат поведінкової сегментації, яка аналізує попередні покупки, пошукові запити та інші дії користувачів на сайті. Такий підхід дозволяє компанії створювати індивідуальні пропозиції для кожного користувача. Amazon розглядає користувачів через призму таких параметрів, як історія покупок, бажані товари, переглянуті категорії та інші взаємодії з платформою. Наприклад, якщо користувач часто купує книги про здоров'я, система буде пропонувати йому нові надходження в цій категорії або супутні продукти, такі як фітнес-обладнання чи органічні добавки. Завдяки цьому покупці отримують релевантні рекомендації, що збільшує ймовірність покупки та покращує їхній досвід.

Nike, одна з провідних світових компаній, що виробляє спортивний одяг та взуття, демонструє приклад використання сегментації на основі фізичної активності та інтересів своїх клієнтів. Використовуючи психографічну сегментацію, Nike пропонує спеціальні продукти та послуги для людей, які активно займаються спортом, а також для тих, хто просто шукає комфортне та стильне спортивне взуття та одяг для повсякденного носіння. Завдяки своїй платформі Nike Training Club та Nike Run Club, компанія збирає дані про тренування користувачів, їхні цілі та рівень фізичної підготовки. Це дозволяє компанії персоналізувати маркетингові кампанії та пропонувати відповідні продукти: для початківців — базові тренувальні плани та комфортне взуття, для професіоналів — високопродуктивне спортивне обладнання та спеціалізовані продукти для максимальних результатів. Збір даних через додатки дає можливість

поведінкової сегментації, що дозволяє Nike забезпечити точність у своїх пропозиціях.

Spotify також є чудовим прикладом застосування сегментації для персоналізації досвіду користувачів. Одна з найбільших переваг сервісу — це його здатність застосовувати поведінкову сегментацію на основі слухацьких звичок. Spotify аналізує, які жанри, виконавці та треки слухають користувачі, і на основі цих даних пропонує персоналізовані плейлисти, такі як "Discover Weekly" або "Release Radar". Це не тільки забезпечує індивідуалізований досвід для кожного користувача, але й дозволяє компанії активно взаємодіяти з клієнтами, підтримуючи їх інтерес до нових музичних відкриттів. Spotify також активно використовує демографічну сегментацію, пропонуючи знижки на преміум-аккаунти студентам, а також надаючи спеціальні пропозиції для сімейних підписок. Ця стратегія дає можливість залучати різні категорії споживачів, що сприяє зростанню кількості платних підписок. Сегментація у Spotify дозволяє не тільки надавати персоналізовані рекомендації, а й запропонувати рекламні кампанії, які відповідають потребам різних сегментів. Для прикладу, Spotify може проводити рекламні кампанії для молоді, що активно слухає музику в поїздках або на тренуваннях, пропонуючи нові релізи артистів або підписку на подкасти.

1.4 Аналіз необхідності використання сегментації у маркетингових компаніях

Сучасний ринок характеризується високою конкуренцією, зростаючими очікуваннями споживачів та швидкими змінами в уподобаннях клієнтів. У цих умовах традиційні підходи до маркетингу, що орієнтуються на масову аудиторію, втрачають свою ефективність. Одним із ключових інструментів підвищення результативності маркетингових кампаній є сегментація клієнтів. Вона дозволяє компаніям точніше визначати цільові аудиторії, адаптувати рекламні

повідомлення та покращувати взаємодію з потенційними і наявними клієнтами. У цьому розділі проаналізуємо основні причини, через які сегментація є необхідною складовою сучасного маркетингу [12].

Однією з основних переваг використання сегментації є можливість персоналізації маркетингових комунікацій. Сучасні споживачі очікують від брендів індивідуального підходу, ігноруючи узагальнені рекламні повідомлення. Завдяки сегментації компанії можуть адаптувати контент відповідно до особливостей кожної групи споживачів. Наприклад, молодь більше зацікавлена в інтерактивному контенті та соціальних мережах, тоді як старші аудиторії віддають перевагу більш традиційним каналам комунікації, таким як електронна пошта чи телебачення.

Завдяки аналізу споживчих уподобань, компанії можуть створювати релевантний контент, що підвищує рівень залученості клієнтів та ефективність рекламних кампаній. Наприклад, використання персоналізованих рекомендацій на основі історії покупок значно підвищує ймовірність повторного звернення до бренду.

Сегментація дозволяє компаніям більш ефективно розподіляти маркетинговий бюджет, уникаючи зайвих витрат на рекламу для нецільової аудиторії. Традиційний масовий маркетинг передбачає значні фінансові витрати на рекламні кампанії, проте ефективність таких вкладень може бути низькою через нерелевантність контенту для певних груп споживачів.

Завдяки сегментації компанії можуть концентрувати свої ресурси на тих споживачах, які з найбільшою ймовірністю здійснять покупку. Наприклад, якщо бізнес продає товари преміум-класу, то маркетингові зусилля слід спрямувати на споживачів із високим рівнем доходу, замість витрачання коштів на широку аудиторію.

Клієнти схильні лояльніше ставитися до брендів, які розуміють їхні потреби та пропонують відповідні рішення. Використання сегментації дозволяє компаніям

краще прогнозувати очікування споживачів та адаптувати свої продукти та послуги відповідно до їхніх інтересів.

Наприклад, у сфері електронної комерції сегментація за поведінковими характеристиками дає змогу пропонувати клієнтам саме ті товари, які вони можуть шукати. Це сприяє покращенню користувацького досвіду, збільшує середній чек покупця та формує довгострокові відносини між брендом і клієнтом.

Ще одним аргументом на користь сегментації є її вплив на ефективність маркетингових стратегій. Компанії, що використовують цей підхід, можуть розробляти більш точні та дієві маркетингові плани. Замість того, щоб покладатися на загальні стратегії, маркетологи можуть створювати окремі рекламні кампанії для різних сегментів, враховуючи їхні унікальні особливості.

Наприклад, сегментація за географічними характеристиками дозволяє брендам адаптувати рекламу під особливості певного регіону. У країнах із різним рівнем доходу ціни та асортимент можуть значно відрізнятися, тому є сенс налаштовувати маркетингові кампанії відповідно до фінансових можливостей локального ринку.

1.5 Вплив сегментації на показники ефективності маркетингових кампаній

Сегментація клієнтів є одним із ключових інструментів у маркетингових стратегіях, що дозволяє компаніям краще розуміти свою аудиторію та адаптувати рекламні повідомлення під конкретні групи споживачів. Використання сегментації безпосередньо впливає на ефективність маркетингових кампаній, підвищуючи основні бізнес-показники та покращуючи рентабельність інвестицій (ROI) [2]. У цьому розділі розглянемо основні метрики ефективності маркетингових кампаній, їхні види та способи покращення за допомогою сегментації.

ROI є ключовою метрикою, яка визначає прибутковість маркетингових вкладень. Формула розрахунку:

$$ROI = \frac{(Revenue - Cost)}{Cost} * 100\% \quad (1.1)$$

де Revenue - дохід, який принесли клієнти;

Cost - сума витрат на яких було використано для взаємодії з продуктом/бізнесом.

Завдяки сегментації компанія може спрямовувати ресурси на ті групи клієнтів, які приносять найбільшу цінність, тим самим знижуючи витрати на неефективні канали реклами та збільшуючи загальний прибуток.

LTV визначає загальний прибуток, який приносить один клієнт за весь період співпраці з компанією [3]. Базова формула:

$$LTV = ARPU * Lifetime \quad (1.2)$$

де ARPU (Average Revenue Per User) – середній дохід на одного клієнта;

Customer Lifetime – середній період утримання клієнта.

Сегментація дозволяє виділити клієнтів із високим потенціалом LTV і розробити для них спеціальні пропозиції, збільшуючи їхню лояльність та довгострокову цінність.

CAC показує середню вартість залучення одного нового клієнта. Розраховується за формулою:

$$CAC = \frac{TotalMarketingSpend}{NewCustomers} \quad (1.3)$$

де $TotalMarketingSpend$ - витрати на певні маркетингові компанії;

$NewCustomers$ - нові клієнти, які були залучені.

Сегментація допомагає знизити CAC шляхом точнішого таргетування реклами та оптимізації маркетингових кампаній. Наприклад, якщо відомо, які

сегменти мають найвищу конверсію, можна спрямувати бюджет саме на них, зменшуючи витрати на менш ефективні аудиторії.

CR вимірює відсоток користувачів, які виконали цільову дію (купівля, реєстрація, підписка) після взаємодії з маркетинговою кампанією:

$$CR = \frac{ActionsTaken}{Suggestions} \quad (1.4)$$

де $ActionsTaken$ - дії, які були виконані;

$Suggestions$ - пропозиції, щоб виконати дії.

Сегментація підвищує CR за рахунок персоналізованого контенту та пропозицій, які краще відповідають потребам конкретних груп споживачів.

Churn Rate – це показник відтоку клієнтів, який визначає, скільки клієнтів припинили користування продуктом або послугою протягом певного періоду. Формула:

$$ChurnRate = \frac{StartCustomers - EndCustomers - NewCustomers}{StartCustomers} \quad (1.5)$$

де $StartCustomers$ - к-ть клієнтів на початку виміряного періоду;

$EndCustomers$ - к-ть клієнтів на кінець виміряного періоду;

$NewCustomers$ - к-ть клієнтів які прийшли у продукт у виміряний період.

Сегментація дозволяє ідентифікувати групи клієнтів із високим ризиком відтоку та вживати превентивних заходів, таких як персоналізовані знижки або спеціальні пропозиції.

CTR використовуються для оцінки ефективності email-маркетингу та реклами:

$$CTR = \frac{Cliks}{Impressions} \quad (1.6)$$

де $Cliks$ - к-ть здійснених кліків, наприклад, на рекламний банер;

$Impressions$ - к-ть показів, наприклад, банеру.

Сегментація дозволяє адаптувати контент під конкретні групи клієнтів, підвищуючи рівень зацікавленості та взаємодії.

Тому в результаті використання сегментації отримуємо такі переваги:

1. Оптимізація бюджету – завдяки точнішому розподілу ресурсів компанія уникає зайвих витрат на неефективні аудиторії, що знижує CAC та підвищує ROI.
2. Персоналізація – покращує досвід клієнтів, збільшуючи їхню лояльність і тим самим підвищуючи LTV.
3. Зменшення відтоку – своєчасна робота з ризиковими сегментами дозволяє втримати клієнтів, що зменшує Churn Rate.
4. Підвищення конверсії – таргетований підхід до різних сегментів сприяє збільшенню CR.
5. Зростання ефективності реклами – персоналізовані рекламні кампанії мають вищий CTR та Open Rate [13].

1.6 Вплив сегментації на показники ефективності маркетингових кампаній

Сегментація є однією з ключових стратегій сучасного маркетингу, що дозволяє ефективно адаптувати маркетингові інструменти до різних груп споживачів. Класичні підходи до сегментації спираються на демографічні, географічні, психографічні та поведінкові характеристики [21], але в умовах розвитку цифрових технологій все більше значення набувають методи, базовані на аналізі даних.

У науковій літературі сегментація часто розглядається як багатовимірна задача, яку доцільно вирішувати з використанням методів машинного навчання. Зокрема, кластеризація, як один з основних методів, дозволяє ідентифікувати природні групи клієнтів без попереднього маркування [6][14]. Найпоширенішим

методом є алгоритм k-середніх (K-means), який широко застосовується завдяки своїй простоті та ефективності [4][9].

Однак, для більш точного аналізу поведінки клієнтів традиційних методів часто недостатньо. У роботі [1] запропоновано поєднання LRFMS-моделі (розширення RFM) з методами кластеризації часових рядів, що дозволяє враховувати динаміку клієнтської активності. Це дозволяє не лише виділити групи клієнтів, а й простежити зміни в їхній поведінці з часом.

Аналогічно, Mosaddegh та співавт. [3] розглядають динаміку сегментів як фактор, що впливає на прогнозування довгострокової цінності клієнта (CLV), підкреслюючи важливість урахування змін у поведінці користувачів. Такий підхід дозволяє приймати більш гнучкі стратегічні рішення.

Ще одним поширеним методом є аналіз RFM (Recency, Frequency, Monetary), який використовується для класифікації клієнтів за тривалістю, частотою та обсягом покупок [6]. Дослідження [5] продемонструвало, як цей підхід можна інтегрувати з кластеризацією для створення сегментів з високою маркетинговою релевантністю.

Метод головних компонент (PCA), описаний у [10], також активно застосовується у задачах сегментації з метою зменшення розмірності простору ознак та збереження найбільш інформативних змінних. Це особливо корисно при обробці великих обсягів даних.

Деякі автори акцентують увагу на концептуальних основах сегментації. Зокрема, Wedel і Kamakura [7] надали фундаментальний опис методологічних аспектів, тоді як Dolnicar та інші [20] зосередилися на практичних підходах до її реалізації. Вони підкреслюють, що ефективна сегментація потребує не лише якісного аналізу даних, а й чіткого розуміння бізнес-цілей.

Практична цінність сегментації також висвітлюється в сучасних аналітичних звітах і публікаціях. Наприклад, у публікаціях [16][24] розглядається роль

сегментації в підвищенні ефективності комунікацій з цільовими аудиторіями, а також у впровадженні персоналізованих стратегій.

Щодо інструментарію, Python та бібліотеки на кшталт Scikit-learn, Pandas і Matplotlib є базовими засобами для реалізації сегментаційних моделей, як зазначено у [13]. Книга [8] є фундаментальним посібником з концепцій і технік Data Mining, які застосовуються в таких задачах.

Таким чином, аналіз джерел свідчить про широку популярність і багатогранність підходів до сегментації клієнтів. Сучасні тенденції спрямовані на використання гібридних моделей, що поєднують класичні метрики, такі як RFM, з алгоритмами машинного навчання та обробкою часових рядів. Це забезпечує більш точне та динамічне групування клієнтів, що є критично важливим у висококонкурентному цифровому середовищі.

1.7 Висновки

У ході дослідження було здійснено аналіз сучасних моделей, методів та технологічних рішень сегментації клієнтів засобами Data Science, що дозволило визначити найбільш ефективні підходи для розв’язання поставленої наукової проблеми. Опрацювання вітчизняних та зарубіжних джерел дало змогу окреслити актуальний стан розробки питань клієнтської сегментації. Також встановлено, що сегментація клієнтів є ключовим інструментом для формування ефективних, цільових і персоналізованих маркетингових кампаній. Використання сегментації дозволяє глибше зрозуміти потреби, мотивації та поведінкові особливості різних груп споживачів, що, у свою чергу, сприяє підвищенню релевантності комунікацій, зростанню рівня лояльності клієнтів та збільшенню конверсій.

На основі зібраної інформації був сформований датасет, що охоплює демографічні та поведінкові характеристики клієнтів. Проведений аналіз цих даних став підґрунтям для побудови моделі сегментації із застосуванням

статистичних методів та алгоритмів машинного навчання. Оцінка якості моделі засвідчила її придатність для практичного використання. У результаті була розроблена інформаційна технологія, що демонструє можливості ефективного впровадження отриманої моделі в маркетингові процеси. Такий підхід дозволяє краще розуміти потреби клієнтів, підвищує релевантність комунікацій і сприяє зростанню ефективності бізнес-рішень.

Таким чином, впровадження сегментації у маркетингові процеси є необхідною умовою для побудови сучасного, клієнтоорієнтованого бізнесу. Компанії, які активно застосовують моделі сегментації, отримують конкурентну перевагу завдяки глибшому розумінню своєї аудиторії, зменшенню витрат на рекламу та підвищенню ефективності комунікацій з цільовими сегментами.

РОЗДІЛ 2
МЕТОДИ ТА МЕТОДИКИ СЕГМЕНТАЦІЇ, ЯКІ ПІДВИЩУЮТЬ ЯКІСТЬ МАРКЕТИНГОВИХ КАМПАНІЙ

2.1 Види моделей, які використовуються на практиці

Сегментація клієнтів є складним, але надзвичайно важливим процесом, що дозволяє бізнесу розробляти ефективні маркетингові стратегії. В сучасній практиці для аналізу та поділу клієнтської бази використовуються різні моделі, які можна умовно розділити на кілька груп: лінійні методи, бізнес-моделі та методи машинного навчання. Кожен із цих підходів має свої переваги та обмеження, а їх правильне застосування допомагає підвищити ефективність маркетингових кампаній.

Лінійні методи сегментації є одним із найпоширеніших підходів у маркетингу завдяки своїй простоті та зрозумілості. Вони ґрунтуються на фіксованих критеріях та поділі клієнтів за визначеними характеристиками. Основна перевага такого підходу – його швидкість та легкість у впровадженні, що дозволяє компаніям без значних витрат розподілити свою аудиторію та адаптувати маркетингові стратегії. Проте головним недоліком є обмежена точність, оскільки лінійні методи не враховують складніші взаємозв'язки між характеристиками клієнтів [14]. До таких методів належать:

1. Демографічна сегментація (поділ клієнтів за віком, статтю, доходом тощо);
2. Географічна сегментація (врахування місця проживання або поведінки клієнтів у різних регіонах);
3. Поведінкова сегментація (аналіз частоти покупок, типових витрат, взаємодії з брендом).

Таким чином, лінійні методи сегментації є важливим інструментом для компаній, що прагнуть швидко адаптувати свої маркетингові активності до різних

груп клієнтів. Незважаючи на їхню відносну простоту, ці методи залишаються ефективними у поєднанні з іншими, складнішими підходами, що дозволяє бізнесу досягати кращих результатів.

Бізнес-моделі сегментації, засновані на економічних, фінансових і стратегічних підходах, використовуються для розуміння поведінки клієнтів і ефективного управління взаємодією з ними. Ці моделі дозволяють компаніям оптимізувати маркетингові стратегії, концентруючи ресурси на найбільш перспективних клієнтських групах. Найпоширенішими є:

1. RFM-аналіз (Recency, Frequency, Monetary Value) – оцінка клієнтів за частотою та сумою, давністю покупок;
2. ABC-аналіз – класифікація клієнтів за значимістю для бізнесу;
3. Customer Lifetime Value (CLV) – визначення довгострокової цінності клієнта для компанії;
4. Сегментація за рівнем лояльності – поділ клієнтів на постійних, випадкових і потенційних.

Ці моделі допомагають компаніям зосереджувати ресурси на найприбутковіших клієнтах та розробляти більш ефективні маркетингові стратегії.

Сучасні технології дозволяють використовувати складні алгоритми машинного навчання для автоматизованої та гнучкої сегментації клієнтів. Методи машинного навчання для сегментації працюють через побудову моделей, які шукають спільні риси серед клієнтів, що можуть не бути очевидними для традиційних методів аналізу. Такі моделі можуть враховувати величезну кількість параметрів, таких як частота покупок, час, що проходить між покупками, реакція на промоакції, а також географічні та соціальні фактори. За допомогою цих даних бізнес може створити точніші і детальніші профілі своїх клієнтів, що дає змогу не лише покращити взаємодію з кожною групою, а й створювати нові можливості для розвитку продуктів або послуг, що задовольняють конкретні потреби.

Важливим аспектом застосування машинного навчання в сегментації є здатність аналізувати й інтегрувати різноманітні типи даних. Наприклад, традиційно компанії використовували лише дані про покупки, але сьогодні можна включати в аналіз також відгуки клієнтів, дані з соціальних мереж, або ж інформацію про поведінку на вебсайті компанії. Такий багатовимірний підхід дозволяє значно глибше розуміти потреби споживачів і створювати більш персоналізовані стратегії взаємодії з ними. До найбільш ефективних підходів належать:

1. Кластеризація (Cluster Analysis) – поділ клієнтів на групи на основі схожості поведінки (алгоритми K-Means, DBSCAN, ієрархічна кластеризація);
2. Регресійні моделі (Regression Analysis) – прогнозування поведінки клієнтів на основі історичних даних;
3. Класифікаційні моделі (Classification Models) – автоматичне віднесення клієнтів до певних категорій на основі сукупності параметрів (Random Forest, XGBoost, логістична регресія);
4. Нейронні мережі (Neural Networks) – складні алгоритми, що дозволяють знаходити приховані закономірності в даних.

Машинне навчання дозволяє компаніям підвищити точність сегментації, що сприяє персоналізації маркетингових кампаній та зростанню ефективності бізнесу.

Використання різних моделей сегментації дає змогу бізнесу краще розуміти своїх клієнтів та ефективніше взаємодіяти з ними. Лінійні методи забезпечують простоту та швидкість аналізу, бізнес-моделі допомагають оцінити фінансову значущість клієнтів, а машинне навчання надає глибокий аналіз та можливість адаптації стратегії в реальному часі. Оптимальним є комплексний підхід, що поєднує різні моделі, дозволяючи підвищити якість маркетингових кампаній та досягти стабільного зростання бізнесу.

2.2 RFM-аналіз

У сучасному бізнес-середовищі успішні компанії прагнуть краще розуміти своїх клієнтів, щоб ефективніше задовольняти їхні потреби та збільшувати прибуток. Одним із найбільш поширених методів аналізу поведінки клієнтів є RFM-аналіз (Recency, Frequency, Monetary), який дозволяє сегментувати клієнтську базу на основі трьох ключових показників: давності останньої покупки (Recency), частоти покупок (Frequency) і загальної суми витрат клієнта (Monetary) [6]. Цей метод є простим у застосуванні, але водночас дуже ефективним для персоналізації маркетингових стратегій та підвищення лояльності клієнтів.

Основні аспекти RFM-аналізу:

1. Recency (давність покупки) — визначає, коли клієнт останній раз здійснював покупку. Чим менше часу минуло, тим більша ймовірність повторної покупки. Це дозволяє компаніям ефективно працювати з клієнтами, які нещодавно взаємодіяли з брендом.
2. Frequency (частота покупок) — показує, як часто клієнт здійснює покупки. Чим вищий цей показник, тим більше шансів, що клієнт залишається лояльним до компанії. Аналіз цього параметра допомагає створювати кампанії для підвищення частоти покупок.
3. Monetary (грошові витрати) — загальна сума витрат клієнта. Дозволяє ідентифікувати клієнтів, які приносять найбільший дохід і потребують особливої уваги з боку маркетологів.

Одним із ключових аспектів використання RFM-аналізу є сегментація клієнтів. За допомогою цього підходу можна виділити кілька категорій: VIP-клієнти, які часто купують і витрачають великі суми; потенційно лояльні клієнти, які здійснили покупку нещодавно, але роблять це нечасто; а також втрачені клієнти, яких потрібно активізувати за допомогою спеціальних маркетингових кампаній. Цей метод дозволяє персоналізувати комунікацію з

різними групами клієнтів, що підвищує ефективність рекламних кампаній і збільшує конверсію [15].

Інтеграція RFM-аналізу у маркетингові стратегії дозволяє автоматизувати процес комунікації, надсилаючи персоналізовані пропозиції та знижки. Наприклад, клієнти з високими значеннями Recency отримують нагадування про новинки, а ті, хто рідко робить покупки, можуть отримати спеціальні пропозиції для мотивації до повторного звернення. Завдяки цьому підходу маркетингові кампанії стають більш ефективними, оскільки компанія спрямовує ресурси на утримання та розвиток найцінніших клієнтів.

Однак, попри свою ефективність, RFM-аналіз має певні обмеження. Він не враховує якісні аспекти поведінки клієнтів, такі як мотивація до покупок чи рівень задоволеності. Крім того, цей метод не завжди підходить для бізнесів із довготривалим циклом прийняття рішення, де покупка відбувається раз на кілька років. Також варто враховувати, що клієнтські звички можуть змінюватися, і метод не завжди встигає коректно реагувати на такі зміни.

Практичне застосування RFM-аналізу можна проілюструвати на прикладі інтернет-магазину електроніки. Використовуючи цей метод, компанія виявила кілька основних сегментів клієнтів. До категорії преміальних клієнтів увійшли споживачі, які регулярно здійснюють дорогі покупки. Їм варто пропонувати ексклюзивні акції та додаткові переваги. Періодичні покупці мають середню частоту покупок, тому для них ефективними будуть нагадування та бонусні програми. Втрачені клієнти можуть зацікавитися поверненням завдяки спеціальним знижкам та персоналізованим пропозиціям [16].

Загалом, RFM-аналіз є потужним інструментом для сегментації клієнтів та підвищення ефективності маркетингових кампаній. Завдяки цьому підходу компанії можуть ефективніше використовувати рекламний бюджет, покращувати персоналізований підхід до клієнтів та підвищувати рівень їхньої лояльності. Однак максимальний ефект досягається при поєднанні RFM-аналізу з іншими

методами аналітики клієнтських даних, що дозволяє отримати повнішу картину поведінки споживачів та розробити дійсно ефективні маркетингові стратегії [17].

2.3 K-Means

Для точнішого аналізу даних використовується алгоритм кластеризації K-Means, який допомагає групувати клієнтів на основі їхньої поведінки та переваг. Цей метод дозволяє бізнесу створювати більш персоналізовані маркетингові стратегії, що значно підвищує ефективність рекламних кампаній та рівень взаємодії з клієнтами.

Алгоритм K-Means [4] працює за принципом поділу даних на K кластерів, де кожен клієнт потрапляє в групу з подібними характеристиками. Наприклад, споживачів можна класифікувати за частотою покупок, середнім чеком або типом придбаних товарів. Це дозволяє маркетологам точніше визначати потреби кожної групи та формувати спеціальні пропозиції. Чим краще сегментована клієнтська база, тим вищий шанс залучення та утримання споживачів.

Використання K-Means у маркетингу дає змогу визначити різні категорії клієнтів. Однією з основних переваг цього методу є його гнучкість, оскільки кількість кластерів можна змінювати відповідно до потреб бізнесу. Наприклад, компанія може виокремити найбільш активних клієнтів і запропонувати їм бонуси або створити окрему стратегію для тих, хто здійснює покупки рідше. Завдяки цьому алгоритму можна персоналізувати комунікацію та підвищити рівень задоволеності клієнтів.

Метод K-means ґрунтується на простій, але потужній ідеї: об’єднувати об'єкти у K груп на основі схожості між ними. Процес починається з вибору K центрів кластерів (центроїдів), після чого алгоритм поступово уточнює їхнє розташування, перерозподіляючи об’єкти до найближчого центру. Математично це

визначається за допомогою євклідової відстані між точками у просторі ознак. На рис. 2.1 показано приклад переміщення центроїдів.

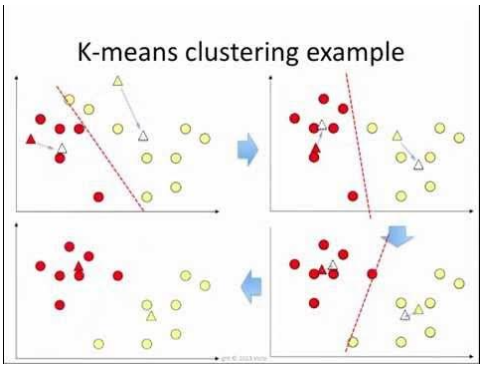

Рисунок 2.1 - Процес навчання K-means за допомогою центроїдів

Формально, мета алгоритму — мінімізувати наступну функцію 2.1:

$$J = \sum_{i=1}^{K} \sum_{j \epsilon C_i} ||x_j - \mu_i||^2 \quad (2.1)$$

де K – кількість кластерів;

x_j – точка у вибірці;

μ_i – центр (центроїд) кластера C_i;

$||x_j - \mu_i||^2$ – квадрат відстані від точки до центроїда.

Попри ефективність алгоритму, він має певні обмеження. Наприклад, складність визначення оптимальної кількості кластерів може впливати на точність результатів. У деяких випадках необхідно додатково аналізувати дані або використовувати інші методи для покращення кластеризації. Крім того, алгоритм чутливий до початкового вибору центрів кластерів, що може призвести до різних результатів при кожному запуску.

Ще однією проблемою є робота з великими масивами даних, оскільки алгоритм може бути обчислювально затратним при великій кількості спостережень. Проте сучасні інструменти машинного навчання дозволяють

масштабувати обчислення, що робить K-Means ефективним навіть у випадку великих клієнтських баз [18].

Уявімо, що онлайн-магазин використовує K-Means для аналізу своїх клієнтів. Після обробки даних алгоритм формує кілька груп: активних покупців, випадкових клієнтів та тих, хто лише переглядає товари, але не робить замовлень. Отримані сегменти дозволяють компанії розробити індивідуальні стратегії: активним клієнтам пропонуються програми лояльності, рідкісним покупцям — знижки для стимулювання повторних покупок, а новим відвідувачам — спеціальні пропозиції для першого замовлення. Таким чином, бізнес отримує можливість ефективніше використовувати ресурси та підвищувати прибутковість маркетингових кампаній [19].

Обробка даних для K-Means кластеризації має кілька технічних аспектів, що включають підготовку даних, параметри алгоритму та важливі особливості в роботі самого методу:

1. Нормалізація даних: Оскільки K-Means використовує евклідову відстань, важливо нормалізувати або стандартизувати дані, щоб змінні з різними масштабами не впливали на результати.
2. Вибір кількості кластерів (k): Для визначення оптимальної кількості кластерів використовують методи, такі як метод ліктя або метод силуету.
3. Ініціалізація центроїдів: Важливо вибирати початкові центроїди з обережністю, щоб уникнути поганої ініціалізації. K-Means++ допомагає вибрати більш рівномірно розподілені центроїди.
4. Ітеративний процес: алгоритм працює в кілька ітерацій, перерозподіляючи точки до найближчих центроїдів та оновлюючи їх.
5. Проблеми з кластеризацією: K-Means має проблеми з класифікацією, коли кластери мають різну форму, розмір або містять аномальні точки.

Алгоритм K-Means є потужним інструментом для маркетологів, які прагнуть покращити сегментацію клієнтів та створювати більш персоналізовані стратегії

залучення. Завдяки кластеризації компанії можуть краще розуміти своїх споживачів, підвищувати рівень задоволеності клієнтів та оптимізувати маркетингові витрати. Проте важливо враховувати обмеження методу та правильно обирати кількість кластерів для отримання найбільш релевантних результатів. Поєднання K-Means із іншими методами аналітики дозволяє досягти ще кращих результатів у плануванні маркетингових кампаній.

2.4 Ієрархічна кластеризація

У світі аналізу даних кластеризація є одним із ключових методів виявлення прихованих структур у вибірках. Ієрархічна кластеризація є особливо ефективним підходом, оскільки дозволяє не лише об’єднувати об’єкти в групи, але й візуалізувати процес кластеризації у вигляді дендрограми. Цей метод широко застосовується в маркетингових дослідженнях, біоінформатиці, аналізі споживчої поведінки та багатьох інших сферах.

Основна ідея ієрархічної кластеризації полягає в тому, що кожен об'єкт спочатку розглядається як окремий кластер, після чого вони поступово об’єднуються у групи на основі схожості між ними. Цей процес може бути реалізований за допомогою агломеративного або дивізивного підходу. Агломеративна кластеризація починається з індивідуальних точок, які поступово об’єднуються у більші кластери, тоді як дивізивний підхід, навпаки, стартує з одного великого кластера, що далі розбивається на дрібніші групи.

Однією з ключових особливостей ієрархічної кластеризації є використання матриці відстаней між об'єктами. Для вимірювання схожості між елементами найчастіше використовують евклідову відстань, яка визначається формулою 2.2:

$$d(x,\ y)\ =\ \sqrt{\sum_{i=`1}^{n}(x_i -\ y_i)^2} \tag{2.2}$$

де $x,\ y$ – координати двох точок у просторі;

n – розмірність простору.

В залежності від поставленої задачі можуть застосовуватися й інші метрики відстані, наприклад, косинусна подібність або міра Манхеттена.

Одним із головних інструментів для візуалізації результатів ієрархічної кластеризації є дендрограма. Вона являє собою дерево, в якому листки відповідають окремим об'єктам, а вузли – кластерам, що утворилися в процесі їх об'єднання. Використання дендрограми дозволяє аналітику визначити оптимальну кількість кластерів шляхом встановлення порогового рівня, на якому слід «обрізати» дерево. Приклад дерева на рис. 2.2.

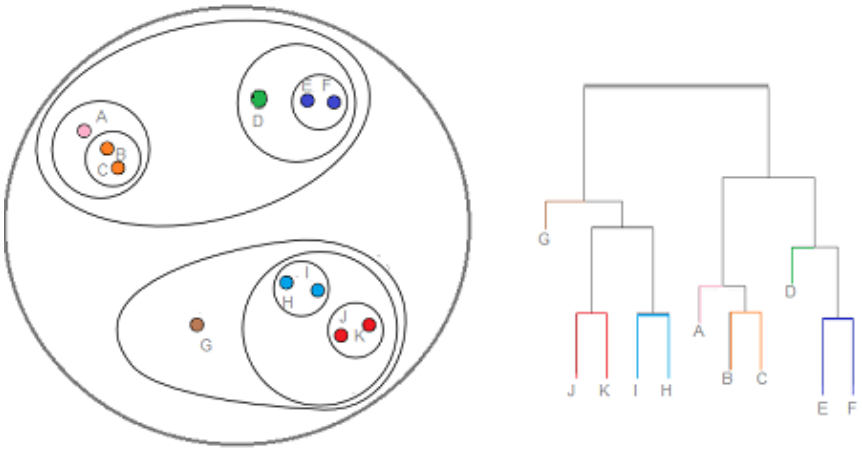

Рисунок 2.2 - Приклад ієрархічно дерева

Переваги ієрархічної кластеризації полягають у її інтуїтивний зрозумілості та можливості аналізу багаторівневої структури даних. Вона не потребує попереднього визначення кількості кластерів, що є значною перевагою порівняно з алгоритмами на зразок K-Means. Проте цей метод має і свої недоліки: він є обчислювально затратним, оскільки потребує розрахунку всіх попарних відстаней між елементами, а також може бути чутливим до вибору метрики подібності та стратегії об’єднання кластерів.

У реальних кейсах ієрархічна кластеризація знаходить застосування в різних галузях. Наприклад, у маркетингу вона використовується для сегментації клієнтів за їхньою поведінкою. Компанії можуть групувати споживачів за рівнем активності, уподобаннями або сумою витрат, що дозволяє створювати персоналізовані пропозиції для кожного сегмента. У медицині цей метод допомагає класифікувати пацієнтів за схожими симптомами або результатами діагностики, що сприяє точнішій постановці діагнозів. В біоінформатиці ієрархічна кластеризація застосовується для аналізу генетичних даних та визначення подібності між різними організмами.

Отже, ієрархічна кластеризація є потужним інструментом для аналізу даних, що дозволяє отримувати глибокі інсайти про структуру вибірки. Її використання особливо корисне в задачах, де важливо не лише отримати кінцевий розподіл об’єктів по групах, але й зрозуміти процес формування цих груп. Незважаючи на обчислювальні труднощі, її гнучкість та візуальна інтерпретованість роблять її незамінною у багатьох сферах науки та бізнесу.

2.5 Хаб сегментів

У сучасному світі маркетинг все більше базується на персоналізованих підходах до споживачів. Компанії прагнуть не просто залучити клієнтів, а створити довгострокові відносини з ними, розуміючи їхні потреби, вподобання та поведінку. Саме для цього використовується хаб сегментів — централізоване сховище інформації про сегменти аудиторії, яке допомагає аналізувати, структурувати та використовувати дані для покращення маркетингових кампаній [20].

Хаб сегментів — це таблиця або набір таблиць у базі даних, де зберігається вся інформація про сегменти клієнтів. Ця інформація може включати демографічні

характеристики, поведінкові дані, історію покупок, вподобання, рівень залученості та інші параметри, які допомагають розділяти клієнтів на групи.

Основні атрибути, які можуть міститися в хабі сегментів:

1. Ідентифікатор клієнта (customer_id).
2. Сегменти різних рівнів і типів.
3. Додаткові метрики, які описують клієнта (за потрібності).

Хаб сегментів відіграє ключову роль у підвищенні ефективності маркетингових кампаній. Він дозволяє компаніям:

1. Оптимізувати таргетинг - завдяки наявності точних даних про клієнтів, маркетологи можуть створювати більш релевантні рекламні кампанії, спрямовані на конкретні сегменти. Наприклад, сегмент "постійні покупці" може отримати ексклюзивні пропозиції на основі їхніх попередніх покупок.
2. Персоналізувати комунікацію - відправка однакових рекламних повідомлень всім клієнтам більше не працює. Хаб сегментів дозволяє адаптувати контент під конкретний сегмент. Наприклад, нові користувачі можуть отримувати вітальні листи з бонусами, тоді як сегмент "сплячі клієнти" — спеціальні знижки для повернення.
3. автоматизувати процеси - використання хабу сегментів у поєднанні з CRM-системами та платформами автоматизації маркетингу дозволяє налаштувати тригерні розсилки, push-сповіщення та персоналізовані рекомендації без ручного втручання.
4. Підвищити ефективність реклами - використання сегментованих даних у рекламних кампаніях дозволяє значно підвищити ROI, оскільки оголошення показуються найбільш релевантній аудиторії.

Для прикладу візьмемо сферу банкінг, оскільки на ринку існує конкуренція між банками і сам банк вже існує, як інструмент комфорту. У табл. 2.1 показано приклад хабу сегментів.

Табл.2.1 - Приклад хабу сегментів для банку

№	Тип сегментації	Опис
1	Revenue segmentation	Розподіл клієнтів за рівнем доходу або фінансових можливостей. Допомагає пропонувати преміальні або стандартні банківські продукти.
2	Demographics segmentation	Поділ клієнтів за віком, статтю, сімейним станом, рівнем освіти тощо. Наприклад, пенсіонери можуть отримувати пропозиції щодо депозитів.
3	Geography segmentation	аналіз клієнтів за місцем проживання чи ведення бізнесу. Допомагає банку адаптувати послуги відповідно до регіональних потреб, наприклад, пропонувати кредити на житло у великих містах або аграрні кредити в сільській місцевості.
4	Lifestyle segmentation	Групування клієнтів на основі їхніх інтересів, споживчих звичок і способу життя. Наприклад, любителям подорожей можуть пропонуватися банківські картки з кешбеком на авіаквитки, а бізнес-клієнтам – преміальні банківські послуги.
5	Product segmentation	Розподіл клієнтів за типами банківських продуктів, якими вони користуються (депозити, кредити, страхування, інвестиції). Дозволяє пропонувати персоналізовані фінансові рішення, крос-продажі та програми лояльності.

Сучасний банківський маркетинг неможливо уявити без детальної сегментації клієнтів. Грамотне поєднання різних видів сегментації дозволяє створювати персоналізовані пропозиції, підвищувати конверсію та покращувати клієнтський досвід. Розглянемо кілька прикладів комбінації сегментацій для ефективних маркетингових кампаній у банку:

1. Одна з найбільш ефективних маркетингових кампаній у банківському секторі спрямована на преміальних клієнтів, які активно подорожують. Для цього необхідно поєднати кілька сегментацій: доходи (Revenue segmentation), стиль життя (Lifestyle segmentation) та використовувані продукти (Product segmentation). Клієнти з високими доходами, які часто подорожують і користуються преміальними банківськими послугами, отримують пропозицію ексклюзивної кредитної картки. Вона включає підвищений кешбек на авіаквитки, доступ до VIP-залів в аеропортах, а також преміальне страхування. Такий підхід дозволяє не тільки підвищити лояльність клієнтів, а й збільшити дохід банку від транзакцій.
2. Регіональні особливості також відіграють значну роль у банківському маркетингу. Для залучення клієнтів, які проживають за межами великих міст, доцільно використовувати географічну сегментацію (Geography segmentation), демографічну сегментацію (Demographics segmentation) та RFM-сегментацію (RFM segmentation). Наприклад, чоловіки віком 25-45 років, які проживають у передмісті або сільській місцевості та мають позитивну кредитну історію, можуть отримати вигідну пропозицію на автокредит. Завдяки такому підходу банк може підвищити рівень видачі автокредитів та залучити нових клієнтів, які потребують особистого транспорту для роботи чи сімейних потреб.
3. Молодь є важливим сегментом для банків, оскільки вони формують майбутню клієнтську базу. Для ефективної роботи з ними потрібно об'єднати демографічну сегментацію (Demographics segmentation), стиль життя

(Lifestyle segmentation) та сегментацію за продуктами (Product segmentation). Наприклад, студенти або молоді фахівці, які активно користуються цифровими сервісами, але ще не мають кредитної історії, можуть отримати спеціальну дебетову картку з бонусами. Вона включає кешбек за оплату в популярних онлайн-магазинах, безкоштовне обслуговування та доступ до освітніх фінансових програм. Такий підхід стимулює молодих клієнтів користуватися банківськими послугами та формує їхню лояльність на довгий період.

4. Клієнти віком 50+ років із високими доходами часто цікавляться стабільними фінансовими інструментами. Для них важливо об'єднати сегментацію за доходами (Revenue segmentation), демографічну сегментацію (Demographics segmentation) та сегментацію за продуктами (Product segmentation). Такі клієнти зазвичай мають великі заощадження, але ще не інвестували в цінні папери чи інші фінансові інструменти. Банк може запропонувати персоналізовану консультацію з фінансового планування, депозити з підвищеною ставкою або гарантовану дохідність на інвестиційні продукти. Це допоможе не тільки задовольнити потреби клієнтів, а й підвищити прибутковість банку.

Отже, застосування комбінованих сегментацій дозволяє банкам значно покращити ефективність маркетингових кампаній. Індивідуальний підхід до кожного клієнта на основі його доходу, географії, демографічних характеристик, стилю життя та попередньої фінансової поведінки допомагає не тільки підвищити рівень залучення клієнтів, а й зміцнити їхню довіру до банку. У майбутньому банки, які активно впроваджують сегментовані маркетингові стратегії, отримають значні конкурентні переваги та зміцнять свої позиції на ринку.

2.6 Висновки

У процесі дослідження було проаналізовано сучасні методи та технологічні рішення сегментації клієнтів у контексті Data Science, включно з вивченням класичних та сучасних джерел як вітчизняної, так і зарубіжної літератури. На основі проведеного аналізу виділено три основні підходи: RFM-аналіз, метод K-Means та ієрархічна кластеризація.

RFM-аналіз відзначається простотою реалізації та дозволяє персоналізувати маркетингові кампанії, проте має обмеження через ігнорування демографічних і поведінкових параметрів. Ієрархічна кластеризація надає глибший аналіз та візуалізацію вкладених сегментів, але не масштабована для великих даних. Наступним виявився метод K-Means, який забезпечує стабільну повторювану сегментацію, гнучкість у виборі ознак та швидкість обробки великих наборів даних.

В результаті було обрано метод K-Means, оскільки для моделювання, який був реалізований, потрібно мати можливість повторно прогнозувати клієнтів банку і отримувати стабільні результати на нових даних. Це особливо важливо для динамічного ринку, де поведінка клієнтів може змінюватися під впливом сезонності, маркетингових кампаній та зовнішніх факторів. Також використання K-Means дозволяє швидко оновлювати сегментацію клієнтів з повним перерахунку всієї моделі. Крім того, цей метод добре масштабується на великі обсяги даних, що робить його ефективним для автоматизованого аналізу клієнтських баз. Завдяки цьому можна регулярно переглядати та оновлювати маркетингові стратегії, покращуючи персоналізацію пропозицій і підвищуючи ефективність взаємодії з клієнтами.

РОЗДІЛ 3
МОДЕЛЮВАННЯ ПРОДУКТУ СЕГМЕНТАЦІЇ ДЛЯ ПІДВИЩЕННЯ ЯКОСТІ МАРКЕТИНГОВИХ КАМПАНІЙ

3.1 Використання CRISP-DM для сегментації

CRISP-DM є ефективною методологією для обробки та аналізу даних, що широко використовується у сфері маркетингу, зокрема для побудови моделей сегментації клієнтів. Завдяки цьому підходу компанії можуть ефективно працювати з великими масивами інформації, розділяти клієнтів на окремі групи та створювати персоналізовані маркетингові кампанії. Це дозволяє не лише підвищити ефективність комунікації з клієнтами, а й оптимізувати витрати на рекламу та збільшити рівень задоволеності споживачів [7].

Методологія CRISP-DM складається з шести основних етапів:

1. Розуміння бізнес-проблеми (Business Understanding) - на цьому етапі визначаються цілі маркетингової кампанії та проблеми, які можна вирішити за допомогою сегментації. Наприклад, компанія може прагнути збільшити конверсію реклами, покращити утримання клієнтів або персоналізувати комунікацію.
2. Розуміння даних (Data Understanding) - аналізуються доступні дані: демографічні характеристики клієнтів, історія покупок, поведінкові патерни, активність у цифрових каналах. Визначаються потенційні фактори, які можуть впливати на сегментацію.
3. Підготовка даних (Data Preparation) - включає очищення даних, обробку пропущених значень, нормалізацію та створення нових ознак. Це критично важливий етап, оскільки якість вхідних даних безпосередньо впливає на точність моделі.

4. Моделювання (Modeling) - використовуються алгоритми кластеризації, такі як k-means, DBSCAN, ієрархічна кластеризація, або методи глибокого навчання для виявлення груп клієнтів із подібними характеристиками.
5. Оцінка результатів (Evaluation) - перевіряється якість сегментації, її відповідність бізнес-цілям, аналізуються метрики, такі як силуетний коефіцієнт (Silhouette Score) або індекс Девіса-Болдіна (DBI).
6. Впровадження (Deployment) - інтеграція моделі у маркетингові процеси, використання сегментації для персоналізованої реклами, таргетованих пропозицій та оцінки ефективності кампаній.

Блок-схема послідовності роботи CRISP-DM на рис. 3.1.

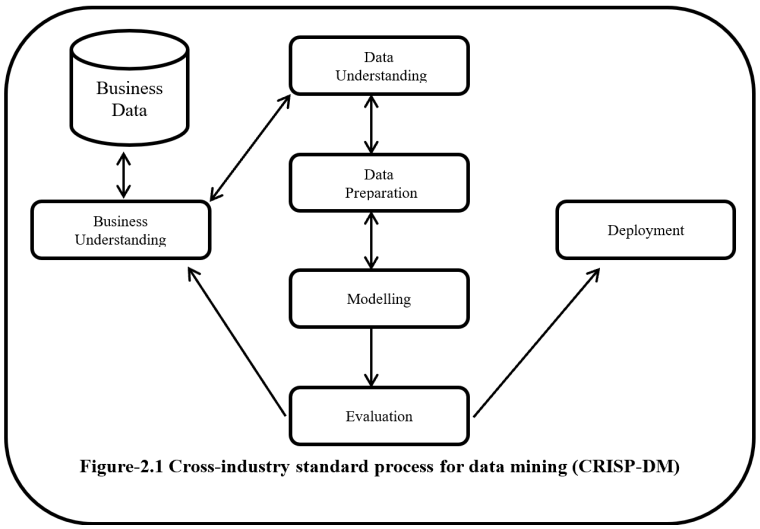

Рисунок 3.1 - Блок-схема послідовності роботи CRISP-DM

На початковому етапі аналізу CRISP-DM фокусується на розумінні бізнес-цілей. Важливо визначити, які результати повинні бути досягнуті за допомогою сегментації. Наприклад, компанія може прагнути підвищити рівень утримання клієнтів, знизити відтік або збільшити середній чек. Чітке формулювання цих цілей дозволяє правильно визначити методи аналізу та критерії оцінки ефективності сегментації. Після цього здійснюється глибокий аналіз

наявних даних. Компанія збирає інформацію про споживачів: їхні демографічні характеристики, частоту покупок, середній чек, канали взаємодії та рівень залученості. На цьому етапі проводиться очищення та підготовка даних, що є критично важливими для коректного побудови сегментів.

Процес підготовки даних передбачає видалення аномалій, заповнення пропущених значень та трансформацію змінних. Наприклад, можна створити показники, що відображають рівень активності клієнта за останні місяці або його схильність до повторних покупок. Після підготовки інформації розпочинається побудова моделі сегментації, яка базується на застосуванні методів кластеризації. Найчастіше використовуються алгоритми K-means, агломеративна кластеризація та DBSCAN, які дозволяють об'єднати клієнтів у групи за подібними характеристиками. Завдяки цим підходам компанія отримує чітке розділення споживачів за рівнем лояльності, поведінковими особливостями та фінансовими можливостями.

Після побудови моделей важливо оцінити їхню якість, використовуючи відповідні метрики. Коефіцієнт силуету, індекс Девіса-Болдіна або внутрішньо-кластерна дисперсія допомагають зрозуміти, наскільки добре алгоритм розподілив клієнтів. Якщо результати не є задовільними, необхідно переглянути вихідні параметри, зменшити або збільшити кількість кластерів, змінити підхід до нормалізації змінних. Завдяки цьому етапу компанія отримує точніші результати, що в майбутньому дозволяє покращити якість маркетингових кампаній [21].

Останнім етапом процесу CRISP-DM є впровадження отриманих результатів у реальні бізнес-процеси. Компанія може використовувати сегментацію для розробки персоналізованих пропозицій, налаштування рекламних кампаній, оптимізації програм лояльності. Наприклад, для клієнтів, які здійснюють часті покупки, можна запропонувати спеціальні знижки або бонуси, тоді як для тих, хто рідко взаємодіє з брендом, можна реалізувати стратегію повторного залучення

через email-маркетинг або таргетовану рекламу. CRISP-DM допомагає компаніям більш глибоко розуміти своїх споживачів, адаптувати комунікаційні стратегії та значно покращувати ефективність маркетингових ініціатив. Це дозволяє не лише збільшити прибутковість, а й формувати довгострокові відносини з клієнтами, що є ключовим фактором успіху на сучасному ринку [8].

3.2 Вибір засобів для реалізації моделей

3.2.1 Python і основні бібліотеки

Python – це одна з найпопулярніших мов програмування для аналізу даних, машинного навчання та маркетингової аналітики. Його гнучкість, велика екосистема бібліотек та простота синтаксису роблять його ідеальним вибором для вирішення задач кластеризації, таких як ієрархічний аналіз або K-means. У сфері маркетингу Python дає можливість ефективно аналізувати клієнтські дані, прогнозувати поведінку споживачів і розробляти персоналізовані стратегії взаємодії з аудиторією [11].

Однією з головних переваг Python є його розвинений набір бібліотек для роботи з даними. Наприклад, Pandas дозволяє легко обробляти та аналізувати великі набори даних, а NumPy забезпечує швидкі математичні обчислення, необхідні для роботи з багатовимірними масивами. Бібліотека Scikit-learn містить широкий спектр алгоритмів машинного навчання, включаючи K-means і ієрархічну кластеризацію, що значно спрощує їхню реалізацію та тестування.

Ще однією особливістю Python є його потужні інструменти для візуалізації даних. Бібліотеки Matplotlib та Seaborn дозволяють створювати графіки, діаграми та дендрограми, що є незамінними для аналізу результатів кластеризації. Це особливо важливо для маркетологів, адже зрозуміла візуалізація допомагає краще інтерпретувати отримані дані та ухвалювати зважені рішення.

Крім того, Python підтримує автоматизацію аналітичних процесів. Наприклад, можна налаштувати регулярний збір та оновлення даних про клієнтів, автоматичне визначення оптимальної кількості кластерів за допомогою методу "лікоть" або оцінку якості кластеризації за допомогою метрик, таких як коефіцієнт силуету. Це дає змогу маркетологам швидко адаптувати свої стратегії до змін у поведінці споживачів.

Проте варто враховувати, що Python, хоч і є потужним інструментом, може бути повільнішим за інші мови, такі як C++ або Java, особливо при роботі з великими обсягами даних. Для масштабних маркетингових досліджень інколи доводиться використовувати спеціальні оптимізації або інтеграцію з іншими технологіями, такими як Spark або Dask, які дозволяють обробляти великі дані ефективніше.

Загалом, Python – це ідеальний вибір для маркетингової аналітики та кластеризації завдяки своїй простоті, великій кількості спеціалізованих бібліотек і можливості автоматизації процесів. Він дозволяє маркетологам не лише аналізувати дані, а й швидко перетворювати отримані результати у реальні бізнес-рішення, що допомагають краще розуміти клієнтів і підвищувати ефективність рекламних кампаній.

Використані бібліотек Python для аналізу даних:

У сучасному світі, де дані стали новою валютою бізнесу, ефективний аналіз інформації є необхідною умовою успіху. Саме тут Python відіграє ключову роль, адже ця мова програмування поєднує простоту, гнучкість та потужність, що робить її ідеальним інструментом для аналітиків та маркетологів. Завдяки широкому набору бібліотек, Python дозволяє швидко обробляти великі обсяги даних, будувати моделі машинного навчання та знаходити приховані закономірності у поведінці клієнтів.

Однією з найважливіших бібліотек є Pandas – інструмент, що спрощує роботу з таблицями даних, дозволяючи легко їх очищати, трансформувати та

аналізувати. У маркетинговій сфері ця бібліотека допомагає сегментувати клієнтів, аналізувати їхню поведінку та приймати зважені рішення. Наприклад, за допомогою Pandas можна оцінити ефективність рекламної кампанії, виявити найбільш прибуткових клієнтів або зрозуміти, які продукти мають найбільший попит. Наглядний приклад датафрейму на рис. 3.2.

```
In [33]: data
```

Out[33]:

	Area Abbreviation	Area Code	Area	Item Code	Item	Element Code	Element	Unit	latitude	longitude	...	Y2004	Y2005	Y2006	Y2007	Y2008	Y2009
0	AF	2	Afghanistan	2511	Wheat and products	5142	Food	1000 tonnes	33.94	67.71	...	3249.0	3486.0	3704.0	4164.0	4252.0	4538.0
1	AF	2	Afghanistan	2805	Rice (Milled Equivalent)	5142	Food	1000 tonnes	33.94	67.71	...	419.0	445.0	546.0	455.0	490.0	415.0
2	AF	2	Afghanistan	2513	Barley and products	5521	Feed	1000 tonnes	33.94	67.71	...	58.0	236.0	262.0	263.0	230.0	379.0
3	AF	2	Afghanistan	2513	Barley and products	5142	Food	1000 tonnes	33.94	67.71	...	185.0	43.0	44.0	48.0	62.0	55.0
4	AF	2	Afghanistan	2514	Maize and products	5521	Feed	1000 tonnes	33.94	67.71	...	120.0	208.0	233.0	249.0	247.0	195.0
5	AF	2	Afghanistan	2514	Maize and products	5142	Food	1000 tonnes	33.94	67.71	...	231.0	67.0	82.0	67.0	69.0	71.0
6	AF	2	Afghanistan	2517	Millet and products	5142	Food	1000 tonnes	33.94	67.71	...	15.0	21.0	11.0	19.0	21.0	18.0
7	AF	2	Afghanistan	2520	Cereals, Other	5142	Food	1000 tonnes	33.94	67.71	...	2.0	1.0	1.0	0.0	0.0	0.0
8	AF	2	Afghanistan	2531	Potatoes and products	5142	Food	1000 tonnes	33.94	67.71	...	276.0	294.0	294.0	260.0	242.0	250.0
9	AF	2	Afghanistan	2536	Sugar cane	5521	Feed	1000 tonnes	33.94	67.71	...	50.0	29.0	61.0	65.0	54.0	114.0
10	AF	2	Afghanistan	2537	Sugar beet	5521	Feed	1000 tonnes	33.94	67.71	...	0.0	0.0	0.0	0.0	0.0	0.0

Рисунок 3.2 - Вигляд Pandas датафрейму

Проте аналіз даних не обмежується лише таблицями. NumPy, ще одна фундаментальна бібліотека, надає потужні інструменти для роботи з числовими масивами та виконання складних математичних операцій. Це особливо корисно при побудові прогнозних моделей, які допомагають передбачати поведінку покупців або оптимізувати ціни товарів.

Коли мова йде про машинне навчання, Scikit-learn стає незамінним помічником. Ця бібліотека містить широкий набір алгоритмів кластеризації, класифікації та регресії, які допомагають знаходити закономірності у великих масивах даних. Наприклад, у маркетинговій аналітиці Scikit-learn можна використовувати для сегментації клієнтів за допомогою K-means або передбачення ймовірності відтоку клієнтів за допомогою логістичної регресії. Приклад показаний на рис. 3.3.

Examples

```
>>> from sklearn import linear_model
>>> clf = linear_model.Lasso(alpha=0.1)
>>> clf.fit([[0,0], [1, 1], [2, 2]], [0, 1, 2])
Lasso(alpha=0.1, copy_X=True, fit_intercept=True, max_iter=1000,
   normalize=False, positive=False, precompute=False, random_state=None,
   selection='cyclic', tol=0.0001, warm_start=False)
>>> print(clf.coef_)
[ 0.85  0.  ]
>>> print(clf.intercept_)
0.15
```

Рисунок 3.3 - Використання лінійної регресії

Проте сухий аналіз цифр не завжди є достатнім. Щоб дані стали зрозумілими, важлива їх візуалізація. Саме тут на допомогу приходять бібліотеки Matplotlib та Seaborn, які дозволяють створювати графіки, діаграми та теплові карти. Візуалізація даних допомагає аналітикам швидко виявляти тренди, порівнювати ефективність різних маркетингових кампаній і знаходити нові можливості для зростання бізнесу. Приклад результату використання Seaborn на рис. 3.4.

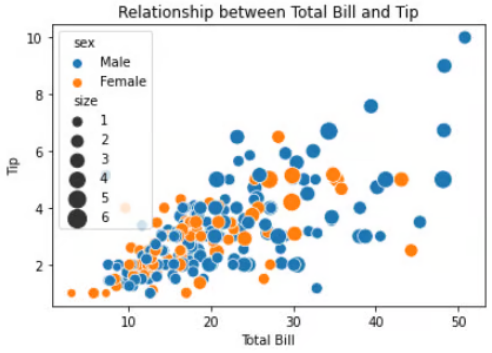

Рисунок 3.4 - Результат використання бібліотеки Seaborn

Таким чином, Python у поєднанні з його бібліотеками відкриває величезні можливості для аналізу даних. Він дозволяє автоматизувати рутинні процеси, знаходити глибокі інсайти та розробляти ефективні маркетингові стратегії. Саме тому він став невід'ємною частиною сучасного бізнес-аналітичного інструментарію.

3.2.2 Google Sheets

У світі, де швидкість ухвалення рішень має вирішальне значення, ефективний аналіз даних стає ключовою перевагою бізнесу. Хоча для цього існує багато складних інструментів, Google Sheets залишається однією з найзручніших і найуніверсальніших платформ для роботи з даними. Простота використання, інтеграція з іншими сервісами та можливість спільної роботи роблять його ідеальним вибором для маркетологів, аналітиків і підприємців.

Однією з головних переваг Google Sheets є його доступність. На відміну від складних програмних рішень, ця платформа не вимагає встановлення додаткового програмного забезпечення – все, що потрібно, це доступ до інтернету. Це означає, що будь-який член команди може редагувати або аналізувати дані в режимі реального часу, незалежно від того, де він знаходиться. У сфері маркетингу це дозволяє оперативно оновлювати звіти про продажі, коригувати бюджети рекламних кампаній і аналізувати поведінку клієнтів без затримок.

Ще однією важливою особливістю є широкий набір функцій для роботи з даними. Google Sheets підтримує безліч математичних, статистичних та логічних операцій, що дозволяє виконувати складні розрахунки без використання додаткових інструментів. Наприклад, за допомогою формул SUMIF та AVERAGEIF можна аналізувати продажі певних товарів, а функція QUERY дає змогу швидко фільтрувати й обробляти великі набори даних.

Проте справжня сила Google Sheets розкривається при використанні додаткових інструментів та інтеграцій. Підключення Google Analytics, автоматичне завантаження даних з CRM-систем або використання Google Apps Script для створення власних скриптів – усе це дозволяє зробити аналіз ще глибшим та ефективнішим. Наприклад, маркетологи можуть автоматизувати звіти про ефективність реклами, а аналітики – налаштовувати дашборди, які оновлюються в реальному часі.

Крім того, Google Sheets є відмінним інструментом для візуалізації даних. Динамічні графіки, діаграми та теплові карти дозволяють швидко оцінювати ключові тренди та приймати обґрунтовані рішення. Це особливо важливо для маркетологів, які аналізують поведінку споживачів, ефективність рекламних кампаній та зміни у продажах. Наприклад на рис. 3.5 можна побачити зведену в якій все чітко і якісно видно зріз по місяцях.

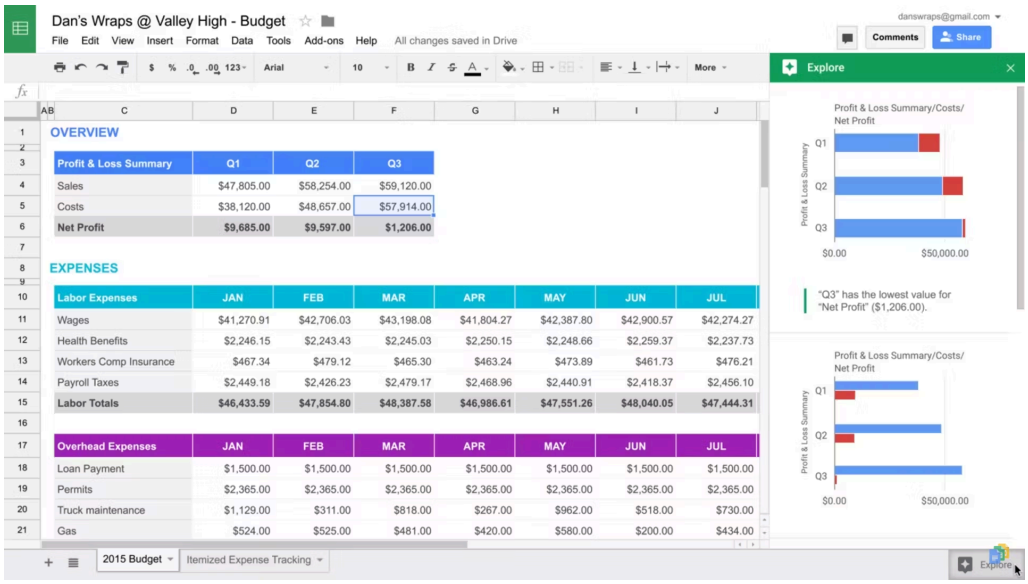

Рисунок 3.5 - Приклад використання Google Sheets як інструмент візуалізації

Попри всі переваги, Google Sheets має і свої обмеження. Його продуктивність знижується при роботі з надзвичайно великими масивами даних, а деякі складні аналітичні завдання краще виконувати в спеціалізованих інструментах, таких як Python чи SQL. Однак для більшості повсякденних завдань він залишається одним із найзручніших рішень. Але як на рис. 3.6 можна побачити його використання, як БД для невеликих даних.

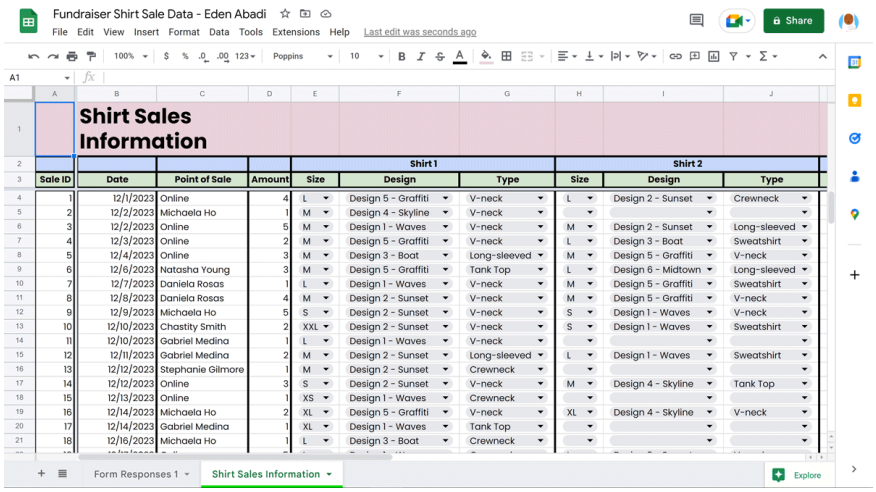

Рисунок 3.6 - Використання Google Sheets, як БД для невеликих даних

Таким чином, Google Sheets – це не просто електронна таблиця, а потужний інструмент для аналізу даних, який допомагає швидко і легко отримувати корисні інсайти. Його простота, гнучкість та інтеграційні можливості роблять його незамінним у маркетингу, бізнес-аналітиці та фінансах. У світі, де дані керують бізнесом, Google Sheets залишається доступним і ефективним рішенням для кожного.

3.2.3 Big Query

BigQuery є потужним хмарним інструментом для обробки та аналізу великих обсягів даних, що надається Google Cloud. Його основною перевагою є здатність швидко виконувати складні SQL-запити на масивах даних, які можуть досягати терабайтів або навіть петабайтів. Завдяки використанню розподіленої обробки та масштабованої архітектури, BigQuery дозволяє бізнесу оперативно отримувати аналітичні інсайти без необхідності розгортати власні сервери чи управляти інфраструктурою. Це робить його ідеальним інструментом для аналізу поведінкових даних клієнтів, прогнозування трендів та оптимізації маркетингових стратегій. На рис. 3.7 видно вигляд інтерфейсу.

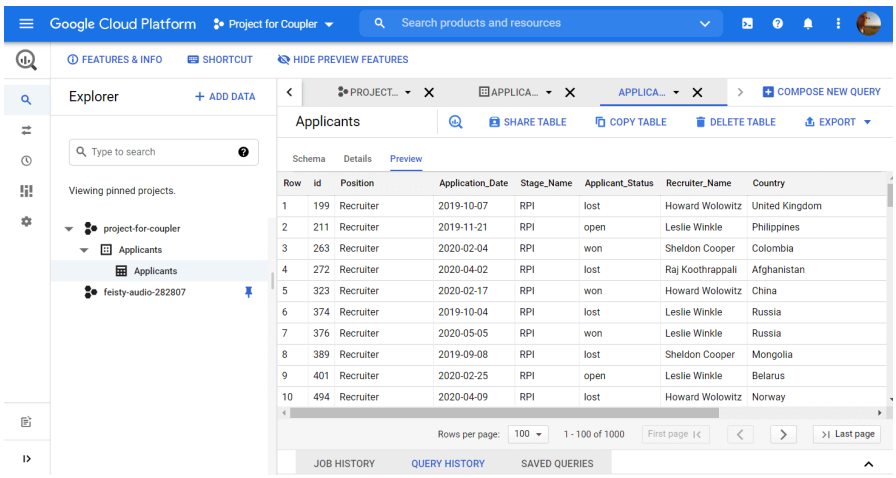

Рисунок 3.7 - Вигляд інтерфейсу Big Query

Одним із ключових аспектів застосування BigQuery є сегментація клієнтів, яка відіграє вирішальну роль у підвищенні якості маркетингових кампаній. Сегментація клієнтів дозволяє компаніям групувати своїх користувачів за різними характеристиками, такими як демографічні дані, поведінка на сайті, рівень витрат, частота покупок та інші параметри. Використовуючи BigQuery, маркетологи можуть легко аналізувати величезні масиви даних, виявляти закономірності та створювати персоналізовані пропозиції для різних груп клієнтів.

Одним із популярних підходів до сегментації є RFM-аналіз (Recency, Frequency, Monetary), який дозволяє оцінювати клієнтів за останньою покупкою, частотою транзакцій та загальною сумою витрат. Використовуючи SQL-запити в BigQuery, можна швидко розраховувати ці метрики для всіх клієнтів, формувати сегменти та визначати найбільш цінних покупців. Наприклад, компанія може створити сегменти «Лояльні клієнти», «Клієнти, що зменшують активність», «Нові клієнти» та «Втрачена аудиторія». Кожен із цих сегментів вимагає індивідуального підходу у маркетинговій комунікації, що дозволяє збільшити ефективність кампаній.

Ще одним підходом до сегментації є кластеризація, яка застосовується для групування клієнтів за схожими поведінковими характеристиками.

Використовуючи BigQuery разом із Google Cloud AI та BigQuery ML, можна будувати моделі машинного навчання без необхідності вивантажувати дані у зовнішні системи. Це значно спрощує процес моделювання та дозволяє інтегрувати результати безпосередньо у бізнес-процеси. Наприклад, за допомогою алгоритму K-means у BigQuery ML можна автоматично виділяти кластери клієнтів, які мають схожі уподобання та поведінкові патерни. Це дає змогу компаніям розуміти, які групи клієнтів більш схильні до повторних покупок, які можуть потребувати додаткових стимулів, а які можуть перейти до конкурентів.

BigQuery також підтримує інтеграцію з іншими інструментами Google Cloud, такими як Looker Studio та Google Ads, що дозволяє створювати динамічні дашборди та автоматизувати аналіз маркетингової ефективності. Наприклад, за допомогою BigQuery можна аналізувати, як різні сегменти клієнтів взаємодіють із рекламними кампаніями, які канали залучення є найбільш ефективними та які сегменти демонструють найвищий рівень конверсії. Це дає можливість маркетологам швидко адаптувати стратегію та оптимізувати бюджети на рекламу.

Додатковою перевагою BigQuery є можливість обробки потокових даних у режимі реального часу, що дозволяє маркетинговим командам оперативно реагувати на зміни у поведінці клієнтів. Наприклад, якщо клієнт протягом останніх кількох днів активно переглядав певні товари, але ще не зробив покупку, система може автоматично надсилати йому персоналізовану знижку або спеціальну пропозицію. Це підвищує рівень залученості клієнтів та сприяє збільшенню продажів.

Окрім маркетингової сегментації, BigQuery можна використовувати для прогнозного аналізу поведінки клієнтів. За допомогою історичних даних можна будувати моделі, що передбачають ймовірність відтоку клієнтів, прогнозують майбутній рівень витрат або допомагають визначити оптимальний момент для запуску рекламних кампаній. Такий підхід дозволяє компаніям діяти проактивно, знижуючи ризики та підвищуючи ефективність маркетингових ініціатив.

Використання BigQuery у процесі сегментації клієнтів відкриває широкі можливості для бізнесу, дозволяючи підвищити точність маркетингових стратегій, збільшити рівень персоналізації та оптимізувати витрати. Завдяки можливості швидкої обробки великих обсягів даних, інтеграції з машинним навчанням та інструментами візуалізації, компанії можуть отримувати більш глибокі інсайти про свою аудиторію, що, у свою чергу, сприяє покращенню досвіду клієнтів та зростанню прибутковості бізнесу.

3.3 Аналіз і обробка вхідних даних

3.3.1 Вхідні дані для моделі

Набір даних представляє собою набір характеристик клієнтів банку і використання продуктів від банку. Оскільки задачею є отримання цінних сегментів, то буде використовуватися повний датасет для обробки. Такий аналіз може бути важливим для банківських установ при розробці маркетингових кампаній, оскільки розуміння різних груп клієнтів може допомогти банкам розробити різні маркетингові компанії, орієнтовані на кожну групу клієнтів, що потенційно може бути ефективнішим, ніж універсальна маркетингова кампанія для всіх клієнтів [9].

Набір даних містить в собі наступну інформацію про кожного клієнта:

1. job (робота) – тип зайнятості клієнта. Може приймати такі категоріальні значення: admin., blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown.
2. marital (сімейний стан) – сімейний статус клієнта. Категоріальні значення: divorced (розлучений або вдівець/вдова), married (одружений/заміжня), single (неодружений/незаміжня), unknown (невідомо).
3. education (освіта) – рівень освіти клієнта. Можливі значення: primary (початкова), secondary (середня), tertiary (вища), unknown (невідомо).

4. default (кредитний дефолт) – чи має клієнт історію несплати за кредитом. Може мати значення: yes (так), no (ні), unknown (невідомо).
5. balance (баланс рахунку) – баланс банківського рахунку клієнта (числове значення).
6. housing (іпотека) – чи має клієнт іпотечний кредит. Категоріальні значення: yes (так), no (ні), unknown (невідомо).
7. loan (особистий кредит) – чи має клієнт персональний кредит. Можливі значення: yes (так), no (ні), unknown (невідомо).
8. contact (тип контакту) – обраний спосіб зв’язку з клієнтом. Може бути cellular (мобільний телефон) або telephone (стаціонарний телефон).
9. pdays (дні після останнього контакту) – кількість днів, що минула після останнього контакту клієнта. Значення 999 означає, що клієнт раніше не контактував.
10. previous (кількість попередніх контактів) – кількість разів, коли клієнта контактували перед поточною маркетинговою кампанією.
11. poutcome (результат попередньої кампанії) – результат попередньої маркетингової кампанії для цього клієнта. Можливі значення: failure (невдача), nonexistent (відсутність контакту), success (успіх).
12. deposit (депозит) – вказує, чи підписався клієнт на строковий депозит. Приймає значення yes (так) або no (ні).
13. age (вік) – вік клієнта (числове значення). Це важливий демографічний параметр, який може впливати на фінансову поведінку клієнта та його схильність до відкриття депозитів або користування банківськими послугами.

3.3.2 Обробка даних

Зчитаємо дані нашого датасету і подивимся, чи коректно імпортувалися дані. Подивимося на перші 5 рядків нашого датасету за допомогою команди .head() із бібліотеки Pandas, як на рис. 3.8.

```
data.head()
```

	age	job	marital	education	default	balance	housing	loan	contact	pdays	previous	poutcome	deposit
0	58	management	married	tertiary	no	2143	yes	no	unknown	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	-1	0	unknown	no

Рисунок 3.8 - Перші рядки вхідного датасету

Як видно, дані імпортувалися коректно. Для того щоб дізнатися які дані прийшли потрібно їх проаналізувати і у випадку виявлення проблеми потрібно їх виправити, тому етапи попередньої обробки будуть такими:

1. Дізнатися розмірність нашої таблиці (датасету);
2. Дізнатися типи даних в колонках;
3. Перевірити датасет на наявність пропусків;
4. Перевірити датасет на наявність дублікатів;

Для початку переводимо в правильний тип даних колонки, рис. 3.9:

```
cat_cols = ['job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'poutcome', 'deposit']

for col in cat_cols:
    data[col] = data[col].astype('category')

data.dtypes
```

	0
age	int64
job	category
marital	category
education	category
default	category
balance	int64
housing	category
loan	category
contact	category
pdays	int64
previous	int64
poutcome	category
deposit	category

Рисунок 3.9 - Обробка типів даних предикторів

Подивимося на дублікати, рис. 3.10:

```
data.duplicated().sum()
```

```
0
```

Рисунок 3.10 - Пошук дублікатів

Дублікатів немає, тому далі дивимося чи є пусті значення, рис. 3.11:

```
data.isnull().sum()
```

	0
age	0
job	0
marital	0
education	0
default	0
balance	0
housing	0
loan	0
contact	0
pdays	0
previous	0
poutcome	0
deposit	0

Рисунок 3.11 - Пошук пропущених значень

Пустих значень немає. І в результаті подивимося на розмір датасету, 3.12:

```
print(f'There are {data.shape[0]} rows and {data.shape[1]} columns.')
```

```
There are 45211 rows and 13 columns.
```

Рисунок 3.12 - Пошук розміру датасету

Як видно з даного аналізу, наш датасет складається з 45211 рядків та 13 колонок, дублікати і пропуски в даних – відсутні. Це добре, оскільки значно

полегшує процес препроцесінгу даних. Більшість колонок мають числовий категорійний тип даних , проте деякі колонки - це числові.

3.3.3 Візуальний аналіз даних

Наступний етап – візуалізація змінних. Будемо проводити візуалізацію за допомогою діаграм, адже вони дозволяють помітити різні закономірності, викиди в даних та наглядно подивитися збалансованість значень змінних. Викид — у статистиці результат вимірювання, який виділяється із загальної вибірки [10].

Причини викидів:

1. Помилка вимірювання;
2. Через незвичайну природу вхідних даних. Наприклад, клієнт банку вніс величезну суму на депозит з малим відсотком доходу і буде дуже сильно відрізнятися від звичайного портретного клієнта;

Аналізуючи вік клієнта можна побачити, що вони розподілені у діапазоні 30-40 років. Це є добре оскільки клієнти банку є платоспроможні, рис. 3.13.

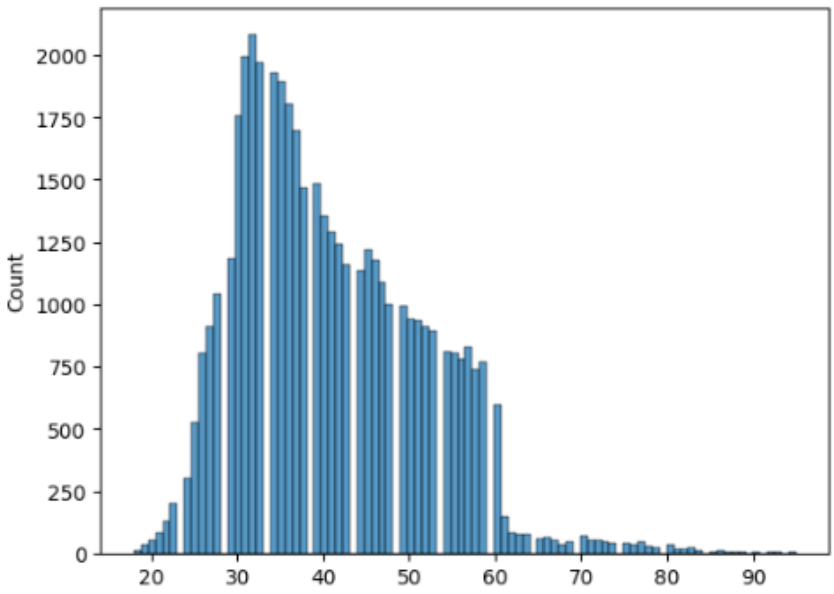

Рисунок 3.13 - Розподіл віку клієнтів банку

При аналізі типу роботи можна побачити, що найбільше є blue-collar, менеджмент, технічні, рис. 3.14.

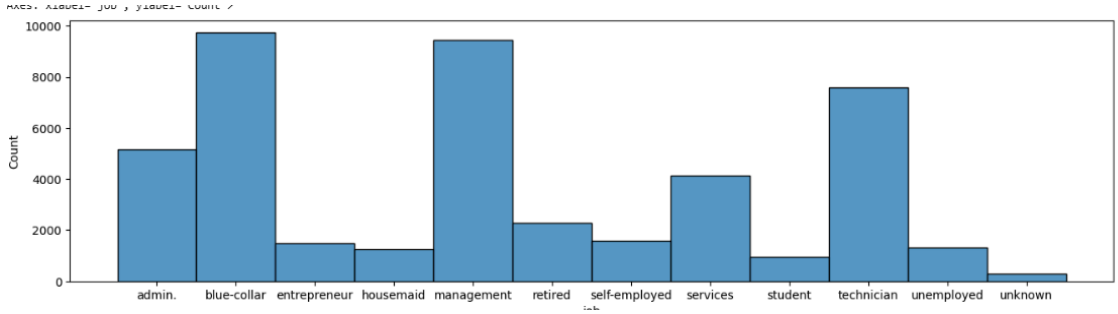

Рисунок 3.14 - Розподіл типу роботи клієнтів банку

При аналізі освіти можна побачити, що клієнти мають більше однієї - отже вони є свідомими і розумними у певних галузях, рис. 3.15.

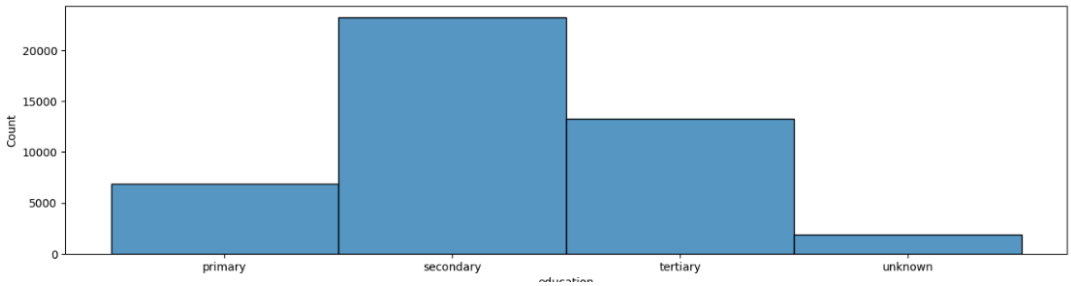

Рисунок 3.15 - Розподіл освіти клієнтів

По предиктору дефолт по кредиту можна побачити, що клієнти майже всі сплачують свої кредити, рис. 3.16:

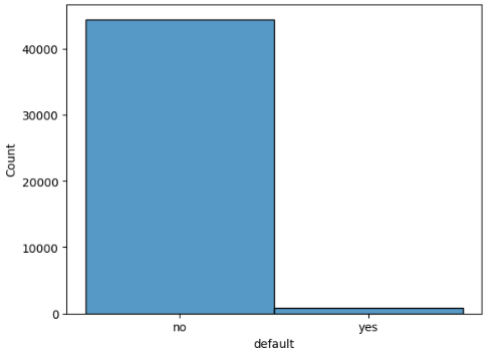

Рисунок 3.16 - Розподіл дефолту клієнтів банку

Даний аналіз допоміг зрозуміти які є клієнти у банку. І можна дати висновок що вони є платоспроможними, без проблем з кредитами із закінченими освітами.

Наступним кроком буде побудова кореляційної таблиці за допомогою бібліотеки seaborn, оскільки невелика к-ть предикторів з числовим значенням, то ми подивимося лише на них, дивитися рис. 3.17.

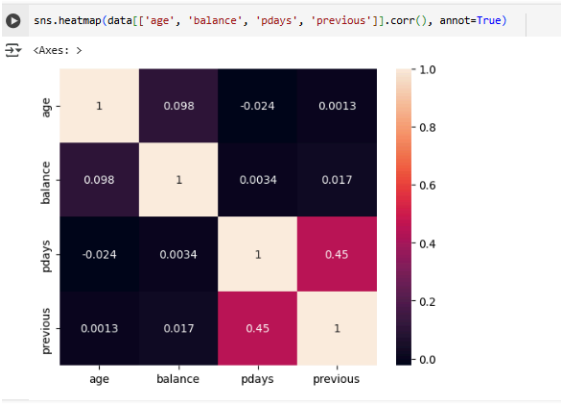

Рисунок 3.17 - Матриця кореляції цифрових предикторів

Як видно, з цієї таблиці видно, що сильних кореляцій немає, тому не потрібно їх прибирати.

3.3.4 Обробка вхідних даних

Для початку потрібно зробити обробку для даних, категорійні перевести у бінарні, а числові нормалізувати з використанням z-стандартизації, рис. 3.18.

```
data_kmeans = pd.get_dummies(data, drop_first=True)
scaler = StandardScaler()
data_kmeans[['age', 'balance', 'pdays', 'previous']] = scaler.fit_transform(data_kmeans[['age', 'balance', 'pdays', 'previous']])

data_kmeans.head()
```

	age	balance	pdays	previous	job_blue-collar	job_entrepreneur	job_housemaid	job_management	job_retired	job_self-employed	...	education_unknown	default_yes	housing_yes	loan_yes	contact_telephone	contact_unknown	poutcome_other	poutcom
0	1.606965	0.256419	-0.411453	-0.25194	False	False	False	True	False	False	...	False	False	True	False	False	True	False	
1	0.288529	-0.437895	-0.411453	-0.25194	False	False	False	False	False	False	...	False	False	True	False	False	True	False	
2	-0.747384	-0.446762	-0.411453	-0.25194	False	True	False	False	False	False	...	False	False	True	True	False	True	False	
3	0.571051	0.047205	-0.411453	-0.25194	True	False	False	False	False	False	...	True	False	True	False	False	True	False	
4	-0.747384	-0.447091	-0.411453	-0.25194	False	False	False	False	False	False	...	True	False	False	False	False	True	False	

5 rows × 29 columns

Рисунок 3.18 - Вигляд обробленого датасету

Нормалізація змінних за допомогою бібліотеки scikit. Для нормалізації був обраний метод z-стандартизації, який перетворює предиктор з середнім 0 і стандартним відхиленням - 1, формула 3.1:

$$z = \frac{x-\mu}{\sigma} \quad (3.1)$$

де x - значення предиктора певного об’єкта;

μ - середнє предиктора;

σ - стандартне відхилення предиктора.

3.4 Навчання моделі

Для реалізації моделі, було обрано кластеризацію K-means, оскільки вона є ефективною і на практиці перевіреною. Також її можна поставити на автоматичне використання і прогнозування, як результат можемо створити автоматичну взаємодію з клієнтами на основі поведінки клієнтів і результату прогнозу моделі надсилаючи різні маркетингові пропозиції.

До початку навчання дані потрібно обробити - змінити чи видалити аномалії, пропуски, некоректні дані, стандартизувати, оскільки k-means краще з ними працює. Всі ці етапи були проведені на попередньому етапі.

Наступний етап вибір к-ті кластерів, для цього використовується метод ліктя. Метод ліктя – один із найпоширеніших підходів для визначення оптимальної кількості кластерів у задачах кластеризації, зокрема для алгоритму k-means. Його суть полягає в тому, щоб дослідити, як змінюється значення внутрішньокластерної дисперсії (inertia) залежно від кількості кластерів.

При збільшенні кількості кластерів значення дисперсії зменшується, оскільки точки розподіляються по більшій кількості груп, і кожен кластер стає більш однорідним. Проте на певному етапі швидкість зменшення дисперсії суттєво сповільнюється, утворюючи характерний "лікоть" на графіку. Саме ця

точка, де графік різко змінює свою поведінку, і є оптимальним вибором для кількості кластерів.

Ефективність методу ліктя для k-means пояснюється його простотою та наочністю. Він дозволяє знайти баланс між надто загальною та надто детальною кластеризацією. Якщо вибрати занадто мало кластерів, дані будуть погано сегментовані, і всередині кластерів залишатимуться значні відмінності між точками. Якщо ж вибрати занадто багато кластерів, модель може втратити узагальнювальну здатність і почне занадто точно підлаштовуватися під вибірку, що призведе до переобчислення та зайвої складності.

Результат роботи даного метода знаходиться на рис. 3.19.

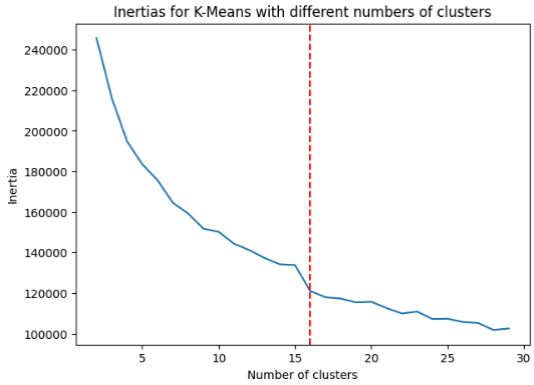

Рисунок 3.19 - Метод ліктя

Як видно на графіку явний згин виділяється на к-ті кластерів - 16. Його й будемо використовувати для отримання результатів і подальшої побудови.

3.5 Індекс стабільності кластерів

Індекс стабільності кластерів (Cluster Stability Index, CSI) є потужним інструментом для оцінки надійності та стійкості результатів кластеризації на різних етапах процесу, який є науковою новизною у даній роботі. Його основна мета — виміряти, наскільки стабільно об’єкти залишаються у своїх кластерах при багаторазовому запуску алгоритму кластеризації на різних підмножинах даних або

за різних параметрів. CSI виступає критичним показником якості сегментації, оскільки дозволяє оцінити, наскільки чітко та послідовно поділено дані на групи.

Методика розрахунку CSI базується на порівнянні результатів кількох запусків алгоритму кластеризації, наприклад, K-Means, із різними початковими умовами або підвибірками даних. Основна ідея полягає у тому, щоб аналізувати, чи залишаються об'єкти у тих самих кластерах під час кожного повторного запуску. Якщо об'єкт стабільно потрапляє до однієї і тієї ж групи, його класифікація вважається надійною, а сам кластер — добре визначеним.

Процес обчислення CSI виглядає так - кластеризаційний алгоритм (наприклад, K-Means) запускається кілька разів із різними випадковими підвибірками даних. Результати кожного запуску порівнюються між собою. Для зіставлення кластерів використовується спеціальний алгоритм, наприклад, угорський алгоритм. Визначається стабільність кожного об'єкта у кластері, після чого обчислюється середній CSI.

Індекс стабільності кластерів має особливе значення у маркетинговій аналітиці, оскільки дозволяє оцінити надійність клієнтської сегментації. Якщо клієнти часто змінюють свої кластери під час багаторазових ітерацій, це може свідчити про недостатню чіткість сегментації або про нестабільність поведінкових груп. Для розрахунку CSI відстежується, наскільки часто кожен клієнт потрапляє в один і той же кластер протягом кількох ітерацій, а потім обчислюється середнє значення для всіх клієнтів.

Цей підхід дає змогу маркетологам чітко відрізнити стабільні сегменти клієнтів від тих, які змінюються та потребують більш персоналізованого підходу. Наприклад, якщо певний клієнт завжди потрапляє в одну й ту ж групу з високим значенням CSI (наприклад, 0.92), це означає, що його поведінка чітко визначена, і він є ідеальним кандидатом для цільових маркетингових кампаній. Інші клієнти з помірною стабільністю (наприклад, CSI = 0.45) можуть демонструвати більш динамічну поведінку, що потребує додаткового аналізу.

Аналізуючи відмінності у стабільності, бізнес може вдосконалювати свої стратегії сегментації, роблячи маркетингові зусилля більш точними та адаптивними. Наприклад, для клієнтів зі стабільною кластерною приналежністю можна розробити стандартні рекламні кампанії, орієнтовані на передбачувану поведінку. Водночас нестабільні клієнти можуть отримувати гнучкіші пропозиції, з урахуванням змін у їхніх потребах та уподобаннях.

Застосування CSI у маркетинговій аналітиці має низку переваг. По-перше, цей індекс допомагає визначити надійність кластеризації, що є критично важливим для розробки ефективних рекламних стратегій. На відміну від традиційних методів оцінки кластеризації, CSI безпосередньо вимірює стабільність клієнтів у сегментах, показуючи, чи є отримані групи чітко визначеними, чи вони є штучними та потребують доопрацювання.

По-друге, CSI дозволяє оцінювати довгострокову ефективність стратегії сегментації. Відстежуючи зміни у стабільності кластерів з часом, компанії можуть виявляти тенденції у поведінці клієнтів та адаптувати свої маркетингові підходи відповідно до змін ринку. Таким чином, CSI є не лише інструментом оцінки поточних кластеризаційних моделей, а й ефективним механізмом для довгострокового стратегічного планування у сфері маркетингу.

За результатами аналізу даної моделі Індекс Стабільності Кластерів (CSI) набув значення 0.8621. Це означає, що 86.21% об'єктів залишаються в тих самих кластерах при багаторазовому запуску алгоритму. Такий високий рівень стабільності є надзвичайно важливим, оскільки підтверджує надійність сегментації та дозволяє бізнесу будувати довгострокові маркетингові стратегії на основі отриманих кластерів.

Для маркетологів стабільні кластери означають можливість впевнено орієнтуватися на конкретні групи клієнтів, знаючи, що ці сегменти залишатимуться актуальними з часом. Це дозволяє розробляти більш персоналізовані рекламні кампанії, спрямовані на чітко визначені аудиторії, що

підвищує їхню ефективність. Окрім цього, стабільна кластеризація сприяє точнішому прогнозуванню поведінки клієнтів та оптимальному розподілу маркетингових ресурсів, що забезпечує ефективне використання бюджету.

Загалом, отримане значення CSI = 0.8621 підтверджує, що результати кластеризації є достовірними. Це не лише зміцнює довіру до сегментації, а й допомагає приймати обґрунтовані рішення на основі даних. Висока стабільність кластерів означає, що маркетингові стратегії, побудовані на їх основі, будуть більш точними, ефективними та здатними приносити максимальну віддачу від інвестицій.

3.6 Результат прогнозу

В результаті ми отримали 16 кластерів, які є стабільними, тепер потрібно проаналізувати їх, зрозуміти і створити можливий CVP для них на основі особливості кластерів (сегментів) з використанням маркетингу, приклад результату на рис. 3.20:

	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	cluster_k-means	cnt	age	balance	pdays	previous	job_blue-collar	job_entrepreneu	job_housemaid	job_managemer	job_retired	job_self-employ	job_services	job_stude
	0	3013	50	903	-1	0	100%	0%	0%	0%	0%	0%	0%	
	1	3900	49	827	-1	0	0%	4%	8%	7%	2%	5%	16%	
	2	5193	35	642	-1	0	0%	5%	4%	5%	0%	4%	24%	
	3	2646	34	1311	147	3	12%	3%	1%	28%	0%	5%	8%	
	4	2987	38	841	341	2	31%	3%	1%	16%	1%	2%	12%	
	5	3437	35	675	-1	0	100%	0%	0%	0%	0%	0%	0%	
	6	1623	43	9761	18	0	16%	4%	3%	30%	4%	5%	6%	
	7	3824	48	990	0	0	0%	7%	2%	70%	1%	5%	1%	
	8	1	40	543	262	275	0%	0%	0%	100%	0%	0%	0%	
	9	1662	56	1600	177	3	13%	3%	4%	20%	21%	3%	6%	
	10	6092	33	920	-1	0	1%	4%	1%	56%	0%	7%	2%	
	11	195	47	30111	34	0	5%	9%	4%	44%	11%	5%	4%	
	12	2778	50	901	-1	0	0%	6%	8%	9%	8%	4%	17%	
	13	1886	63	1572	3	0	1%	1%	7%	4%	76%	2%	1%	
	14	435	40	1547	225	14	20%	2%	2%	23%	3%	3%	6%	
	15	5539	31	741	-1	0	25%	1%	1%	4%	0%	2%	15%	

Рисунок 3.20 - Вигляд зведеної таблиці кластерів для аналізу

Для того щоб подивитися результат моделі по кластерам, їх аналіз і можливий CVP, то це можна подивитися у додатку а.

3.7 Висновки

У межах реалізації поставлених завдань було розглянуто методологію CRISP-DM як структурований підхід до розробки Data Science-проєктів, що забезпечив логічну та ефективну організацію етапів дослідження. Було обґрунтовано вибір інструментів для реалізації проєкту, зокрема використання мови програмування Python з відповідними бібліотеками для обробки та аналізу даних, а також середовищ BigQuery і Google Sheets для зберігання та взаємодії з даними. Проведено підготовку вхідних даних, включаючи їх очищення та трансформацію, після чого здійснено аналіз і кластеризацію клієнтів методом k-means. У результаті було виокремлено 16 кластерів, кожен з яких детально проаналізовано з урахуванням демографічних і поведінкових характеристик. На основі отриманих результатів запропоновано концепцію ціннісної пропозиції (CVP) для кожного сегмента, що дозволяє підвищити ефективність маркетингових рішень і сприяти більш персоналізованій взаємодії з клієнтами.

Крім того, було реалізовано нову метрику — Індекс Стабільності Кластерів (CSI), який є потужним інструментом для оцінки надійності та стійкості результатів кластеризації на різних етапах процесу. CSI дозволяє оцінити, наскільки стабільно об’єкти залишаються у своїх кластерах при багатьох повторних запусках алгоритму кластеризації на різних підмножинах даних або з різними параметрами. Методика базується на багаторазовому запуску алгоритму (наприклад, K-Means) та зіставленні результатів кластеризації з використанням угорського алгоритму. На основі цих даних розраховується середній рівень стабільності кластерів.

Згідно з розрахунками, Індекс Стабільності Кластерів (CSI) склав 0.8621, що означає, що 86.21% об'єктів залишаються в тих самих кластерах при повторних запусках алгоритму. Такий високий рівень стабільності підтверджує надійність сегментації та дозволяє бізнесу будувати довгострокові маркетингові стратегії на

основі отриманих кластерів. Стабільні сегменти клієнтів дають змогу маркетологам впевнено орієнтуватися на конкретні аудиторії, розробляти персоналізовані рекламні кампанії та прогнозувати поведінку користувачів у майбутньому.

РОЗДІЛ 4
ПРАКТИЧНА РЕАЛІЗАЦІЯ ЗАПРОПОНОВАНОГО МЕТОДУ ТА ОЦІНКА ЯКОСТІ В МАРКЕТИНГОВИХ КОМПАНІЯХ

4.1 Загальний алгоритм роботи методу

В результаті роботи було створено скрипт, який ділиться на окремі блоки, як показано на рис. 4.1:

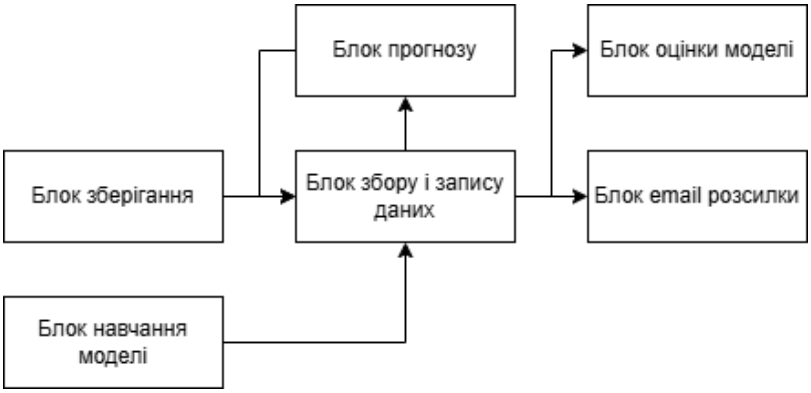

Рисунок 4.1 - Загальна блок-схема системи

Особливість даного алгоритму полягає в тому, що кожний блок можна модифікувати і змінювати під потреби кожного бізнесу окремо, наприклад, змінити алгоритм, додати новий блок, модифікувати дані і т.д. Проте структура повинна мати основні блоки, такі як:

1. Блок зберігання - це місце, де зберігаються дані. База даних може бути різною: файли, хмарні реляційні, локальні реляційні, NoSQL та інші. В даній роботі було обрано саме хмарне зберігання даних, оскільки воно оптимізує час і пам’ять для спеціалістів, які не мають потужного обладнання.
2. Блок навчання моделі - це місце, де проводиться етап навчання моделі. В залежності від методу роботи в компанії можуть бути різні процеси, такі як навчання кожного дня, AutoML, або одноразове навчання. В даній реалізації

було обрано підхід із одноразовим навчанням моделі. У якості алгоритму вибрано кластеризацію K-means.

3. Блок збору і запису - цей блок відповідає за об’єднання даних і маніпуляцію з сирими даними або даними вже натренованої моделі. Дані клієнтів передаються у модель для визначення до якого сегменту належить кожен користувач. Отримані сегменти клієнтів зберігаються у спеціалізованій таблиці dataframe який завантажується одразу у Bigquery для подальшого використання. Цей блок також може містити інформацію про впевненість моделі, змінні, які найбільше вплинули на класифікацію, а також часові мітки. Збережені результати використовуються як джерело даних для кампаній.
4. Блок прогнозу даних — На виході отримуються мітки сегментів для кожного клієнта. Це дозволяє згодом застосовувати цільові маркетингові стратегії для кожної групи.
5. Блок email розсилки — фінальний блок, у якому результати сегментації використовуються для створення персоналізованих маркетингових кампаній. Кожному сегменту клієнтів надсилаються окремі пропозиції, що максимально відповідають їхнім інтересам, поведінковим звичкам чи етапу життєвого циклу. Це дозволяє підвищити коефіцієнт відкриття листів, клікабельність та конверсію кампаній.
6. Блок оцінки моделі - є ключовим етапом, де відбувається виявлення будь-яких змін у розподілі вхідних даних, що можуть вплинути на точність моделі. Як основну метрику перевірки був обраний Population Stability Index (PSI), який дозволяє виміряти зміну в розподілі характеристик між старими (навчальними) та новими даними.

4.2 Блок зберігання даних

У процесі реалізації аналітичної моделі кластеризації клієнтів було створено єдину інтегровану таблицю у Google BigQuery, яка об’єднує необхідні для моделі змінні з різних джерел даних. Такий підхід виявився стратегічно важливим кроком для забезпечення цілісності, узгодженості та масштабованості аналітичного процесу.

В умовах сучасного бізнесу дані зазвичай зберігаються у різних таблицях та системах: транзакційна активність, демографічні характеристики, поведінкові патерни тощо. Без інтеграції цих фрагментованих джерел аналітика втрачає точність і глибину. Саме тому виникла необхідність сформувати єдину таблицю, яка містить усі релевантні атрибути для подальшого навчання моделі кластеризації.

Google BigQuery — це високопродуктивна хмарна аналітична платформа, яка дозволяє зручно об’єднувати великі обсяги даних у єдине сховище. Створення загальної таблиці має низку переваг. По-перше, централізація даних забезпечує високу продуктивність: не потрібно кожного разу виконувати складні запити з об’єднанням джерел — усе вже готово для моделі. По-друге, це значно спрощує інтеграцію моделі в робочі процеси, оскільки один запит до однієї таблиці забезпечує повний набір потрібних параметрів. По-третє, такий підхід полегшує контроль якості даних — простіше виявити пропуски, дублікати, аномалії чи невідповідності, коли все зібрано в одному місці.

Єдина таблиця в BigQuery виступає своєрідним «золотим джерелом» для машинного навчання, що дозволяє забезпечити надійність, повторюваність результатів та зручну підтримку моделі в майбутньому. Така архітектура не лише спрощує роботу з даними, але й відкриває можливості для масштабування та автоматизації: нові змінні чи нові записи можуть автоматично додаватись до таблиці та використовуватись без значних змін у коді чи структурі моделі.

Таким чином, створення об'єднаної таблиці у BigQuery є не просто технічним рішенням, а фундаментом ефективної та гнучкої аналітичної екосистеми, яка відповідає сучасним вимогам до роботи з великими даними.

4.3 Блоки збору і запису даних

У рамках реалізації блоків для зберігання і запису даних у Google BigQuery було розроблено клас BigQueryClient, що виконує роль універсального клієнта для обробки основних операцій із зовнішньою аналітичною платформою BigQuery за допомогою мови програмування Python та бібліотеки google.cloud.bigquery. Даний компонент було створено для автоматизації процесів взаємодії з хмарним сховищем даних у рамках побудови аналітичного пайплайну або системи бізнес-аналітики.

Функціональність класу охоплює такі ключові можливості:

1. Ініціалізація з'єднання з сервісом BigQuery шляхом передачі шляху до облікових даних та ідентифікатора проєкту Google Cloud. При цьому встановлюється необхідна змінна середовища та створюється об'єкт клієнта.
2. Виконання SQL-запитів до BigQuery з можливістю повернення результатів у вигляді об'єкта DataFrame, що забезпечує зручну подальшу обробку даних засобами бібліотеки pandas.
3. Конвертація стовпців у формат рядків, що дозволяє попередньо підготувати дані перед вставкою, уникнувши помилок типізації.
4. Вставка даних у таблицю BigQuery, з можливістю вибору режиму роботи — додавання (append) або повна заміна (replace) даних у таблиці.
5. UPSERT-операція (поєднання INSERT та UPDATE) реалізується через створення тимчасової таблиці, завантаження туди нових даних та виконання SQL-запиту типу MERGE, що оновлює існуючі записи або додає нові. Це дозволяє підтримувати актуальний стан даних у цільовій таблиці.

6. Після виконання UPSERT-операції відбувається автоматичне очищення тимчасової таблиці, що знижує ризики накопичення зайвих об'єктів у хмарному середовищі.

У складі розробленого програмного модуля BigQueryClient реалізовано низку функцій (методів), кожна з яких виконує окреме завдання, пов’язане з обробкою та передачею даних до хмарної платформи Google BigQuery. Нижче подано детальний опис кожної з функцій:

1. Метод __init__:
    a. Ініціалізує об’єкт класу BigQueryClient, встановлює шлях до файлу облікових даних Google та створює клієнта для взаємодії з BigQuery.
    b. Встановлює змінну середовища GOOGLE_APPLICATION_CREDENTIALS для аутентифікації.
2. Метод execute_query:
    a. Виконує SQL-запит до бази даних BigQuery та повертає результат у вигляді об'єкта pandas.DataFrame.
    b. Забезпечує зручну обробку результатів запиту за допомогою інструментів Python.
3. Метод convert_columns_to_string:
    a. Перетворює зазначені стовпці у DataFrame до рядкового типу даних (string).
    b. Корисний у випадках, коли необхідно забезпечити відповідність типів даних для завантаження у BigQuery.
    c. Метод insert_data:
    d. Виконує завантаження даних з pandas.DataFrame у таблицю BigQuery.
    e. Параметр if_exists визначає режим завантаження:
        i. – "append" – додає нові дані до існуючих;
        ii. – "replace" – повністю перезаписує таблицю.

f. Успішне завершення виводить повідомлення з іменем цільової таблиці.

4. Метод upsert_data:
    a. Реалізує логіку UPSERT – оновлення існуючих записів або вставлення нових, якщо їх не існує.
    b. Завантажує дані у тимчасову таблицю.
    c. Виконує MERGE-запит, що порівнює записи за унікальним ідентифікатором.
    d. При збігу оновлює вказані поля, інакше – додає новий рядок.
    e. Видаляє тимчасову таблицю після завершення операції.

4.4 Блок навчання моделі

Навчання моделі кластеризації є ключовим етапом у створенні ефективної аналітичної системи, і цей процес є відповідальністю спеціалістів з машинного навчання. У рамках роботи було створено пайплайн моделювання, що дозволяє спеціалісту безпосередньо навчати модель на основі зібраних даних, а після навчання зберігати отриману модель у вигляді файлу для подальшого використання [5].

Процес навчання моделі кластеризації зазвичай включає кілька основних етапів. Спочатку спеціаліст підбирає алгоритм кластеризації, що найбільше підходить для конкретних даних. Після цього на основі вибраного алгоритму створюється пайплайн, в який входить попередня обробка даних, вибір та налаштування параметрів моделі, а також сама кластеризація. По завершенню навчання модель зберігається у файл, зазвичай у форматі pickle, який підтримує популярні бібліотеки машинного навчання. Це дозволяє зберегти модель для подальшого використання без необхідності повторно виконувати етапи тренування.

Переваги цього підходу:

1. Контроль та гнучкість: Спеціалісти мають можливість тонко налаштувати модель і пайплайн відповідно до специфіки даних та бізнес-завдань. Це дозволяє досягати більш точних і надійних результатів. Кожен етап може бути настроєний вручну, що дає перевагу при роботі з різноманітними типами даних та умовами.
2. Масштабованість та повторюваність: Збереження моделі у вигляді файлу дозволяє швидко відновлювати модель для нових наборів даних або для повторних експериментів, не витрачаючи часу на повторне навчання. Крім того, завдяки пайплайну можна автоматизувати процес обробки та аналізу нових даних без залучення спеціалістів на кожному етапі.
3. Оптимізація ресурсів: Моделі можна зберігати та використовувати багато разів без необхідності повторно тренувати їх, що заощаджує обчислювальні ресурси та час. Це є особливо важливим при роботі з великими наборами даних, де тренування моделі може займати значний час.
4. Покращення якості результатів: Завдяки можливості налаштовувати модель, спеціалісти можуть досягти високої якості кластеризації, використовуючи точні алгоритми та оптимізуючи параметри для досягнення найкращих результатів.

Недоліки цього підходу:

1. Залежність від спеціалістів: Оскільки процес навчання та налаштування моделі здійснюється фахівцем, для досягнення якісних результатів потрібні високі професійні навички та досвід. Модель може бути неефективною, якщо її навчання було проведено неналежним чином або без належної уваги до вибору алгоритмів та параметрів.
2. Часовий та обчислювальний ресурс: Хоча збереження моделі дає можливість повторно використовувати її, сам процес навчання моделі може бути

витратним за часом і ресурсами, особливо при роботі з великими даними. Це може бути проблемою в умовах обмежених ресурсів.

3. Потреба в оновленнях: З часом, коли дані змінюються, модель може стати менш точною або застарілою. Тому спеціалісти повинні регулярно перевіряти ефективність моделі та проводити її перенавчання, щоб підтримувати її актуальність. Однак це потребує додаткових зусиль та часу.
4. Ризик overfitting: Без правильного налаштування моделі існує ризик переобучення (overfitting), коли модель надмірно адаптується до конкретних характеристик навчальних даних і втрачає здатність до узагальнення на нові дані.

4.5 Блок прогнозу моделі

Клас ClusterPredictor призначений для прогнозування кластерних міток для нових споживачів або записів на основі попередньо натренованої моделі кластеризації.

Клас завантажує збережений пайплайн моделі кластеризації (наприклад, збережений об'єкт KMeans, Pipeline або інший класифікатор) та дозволяє передбачати, до якого кластеру належить кожен новий об’єкт.

Це може бути корисним для подальшої персоналізації маркетингових кампаній, розробки цільових пропозицій, сегментації користувачів тощо. Основні функції класу:

1. Метод __init__:
    a. Завантажує модель кластеризації з диску:
        i. parameter model_path: шлях до збереженого файлу з моделлю або пайплайном (наприклад, .joblib файл).
        ii. returns: ініціалізований об’єкт із готовою до використання моделлю.

2. Метод predict_clusters:
    a. Здійснює прогнозування кластерної належності для нового набору даних.
        i. parameter new_data: датафрейм із новими записами, які потрібно класифікувати за кластерами.
        ii. returns: датафрейм із колонками id, mail та новою колонкою cluster_prediction, що містить прогнозовану кластерну мітку.
    b. Процес роботи:
        i. Видаляє колонки id і mail, які не беруть участі у кластеризації.
        ii. Зберігає значення id та mail для подальшого об'єднання з результатом.
        iii. Прогнозує кластери за допомогою завантаженого пайплайну.
        iv. Додає мітки кластерів до результату.

4.6 Блок email розсилки

4.4.1 Особливість використання

Email-маркетинг залишається одним із найбільш ефективних каналів комунікації з клієнтами, особливо у поєднанні з результатами моделей сегментації. Використання мови програмування Python для реалізації email-розсилок відкриває широкі можливості для автоматизації, персоналізації та масштабування маркетингових кампаній.

Однією з головних переваг методу є гнучкість. Завдяки бібліотекам Python, таким як smtplib, email, ssl або більш просунутим фреймворкам типу yagmail, можливо налаштувати як прості розсилки, так і складні багатокомпонентні повідомлення з HTML-шаблонами, вкладеннями та логікою персоналізації. Також Python легко інтегрується з базами даних, файлами Excel або Google Sheets, що

дозволяє динамічно формувати список адресатів відповідно до сегментації, проведеної моделлю.

Проте, існують і певні обмеження. По-перше, пряме використання SMTP-серверів (наприклад, Gmail, Outlook) має ліміти на кількість листів, які можна відправити за добу. У випадку Gmail це близько 500 листів для звичайного облікового запису, що створює обмеження для масштабних кампаній. По-друге, масові розсилки з локальних або маловідомих серверів часто потрапляють у спам-фільтри, якщо не дотримані правила SPF, DKIM і DMARC. Крім того, створення системи обробки відповідей, аналізу відкриттів, переходів по посиланнях та відписок вимагає додаткових рішень або інтеграції з сторонніми сервісами.

Подальші кроки для покращення можуть включати інтеграцію з спеціалізованими сервісами розсилок через API (наприклад, Mailgun, SendGrid або Amazon SES), які знімають більшість обмежень SMTP і надають інструменти для відстеження ефективності. Також перспективним є використання динамічних шаблонів, A/B тестування тем і вмісту листів, а також застосування штучного інтелекту для адаптації повідомлень у реальному часі. З часом систему можна розширити до повноцінної CRM-автоматизації, де кожен контакт отримує унікальний шлях комунікації в залежності від свого сегменту та поведінки.

Таким чином, використання Python для email-розсилок — це потужний інструмент у руках аналітика чи маркетолога, що дозволяє швидко реагувати на зміни в базі клієнтів, запускати персоналізовані кампанії та постійно вдосконалювати комунікацію з аудиторією. Тому було обрано реалізувати даний метод, оскільки він є безкоштовним і надає можливість надсилати до 500 листів на день, що є ефективним варіантом для стартапів і починаючих бізнесів.

4.4.2 Налаштування розсилки у Python

Для того щоб надсилати маркетингові компанії, було створено клас EmailSender, який використовується для відправки електронних листів через SMTP сервер (зокрема для Gmail).

Імпорти:

1. smtplib: Бібліотека для роботи з протоколом SMTP для надсилання електронних листів.
2. os: Бібліотека для роботи з операційною системою, зокрема для перевірки наявності файлів.
3. email.message: Модуль для створення та обробки email повідомлень (включаючи додавання вкладень).
    a. Клас EmailSender:
    b. Конструктор __init__ - Ініціалізує об'єкт класу з необхідними параметрами:
        i. smtp_server: адреса SMTP сервера (для Gmail це "smtp.gmail.com").
        ii. port: Порт для з'єднання з SMTP сервером (для Gmail це 587).
        iii. sender_email: Email адреса відправника.
        iv. app_password: Пароль додатку, який використовується для автентифікації на сервері.
    c. Метод load_text_from_file:
        i. Завантажує текстовий вміст з файлу за вказаним шляхом.
        ii. Якщо файл не знайдений, викидається виняток FileNotFoundError.
    d. Метод send_email:
        i. Відправляє email за допомогою SMTP сервера.
        ii. аргументи:

1. recipient_email: адреса отримувача.
2. subject: Тема листа.
3. message: Тіло листа.
4. attachment_path: Шлях до файлу, який можна прикріпити як вкладення (не обов'язковий).

iii. Створюється повідомлення EmailMessage, яке містить відправника, отримувача, тему та текст.

iv. Якщо вказано вкладення, файл читається та додається до листа.

v. Після цього встановлюється з'єднання з SMTP сервером, відправляється повідомлення, і виводиться повідомлення про успішне відправлення.

e. Обробка помилок:

i. Помилки автентифікації (SMTPAuthenticationError), проблеми зі з'єднанням (SMTPConnectError), а також інші загальні помилки обробляються з виведенням відповідних повідомлень.

В результаті створеного класу, можна відправити повідомлення з текстового документа, як на рис. 4.2:

```
# Налаштування SMTP сервера та пароля додатку
SMTP_SERVER = "smtp.gmail.com"
PORT = 587
SENDER_EMAIL = "yevhen.01@gmail.com"

# Завантаження пароля додатку з текстового файлу
password_file_path = r'C:\Users\Yevhen\Desktop\Маістерська\connect\test.txt'
with open(password_file_path, "r", encoding="utf-8") as file:
    app_password = file.read().strip()

# Створюємо екземпляр класу EmailSender
email_sender = EmailSender(SMTP_SERVER, PORT, SENDER_EMAIL, app_password)

# Завантажуємо текст листа з .txt файлу
text_path = r"C:\Users\Yevhen\Desktop\Маістерська\CVP\кластер_15.txt"
message_content = email_sender.load_text_from_file(text_path)

# Надсилаємо email
email_sender.send_email(
    recipient_email="yevhen.01@knu.ua",  # Вказати отримувача
    subject="Тестове повідомлення_1",         # Вказати тему
    message=message_content,                # Текст листа
    attachment_path=None                    # Вкладення (не обов'язково)
)
```

Рисунок 4.2 - Приклад коду для розсилки

Де в результаті отримуємо таке повідомлення на пошті, рис. 4.3:

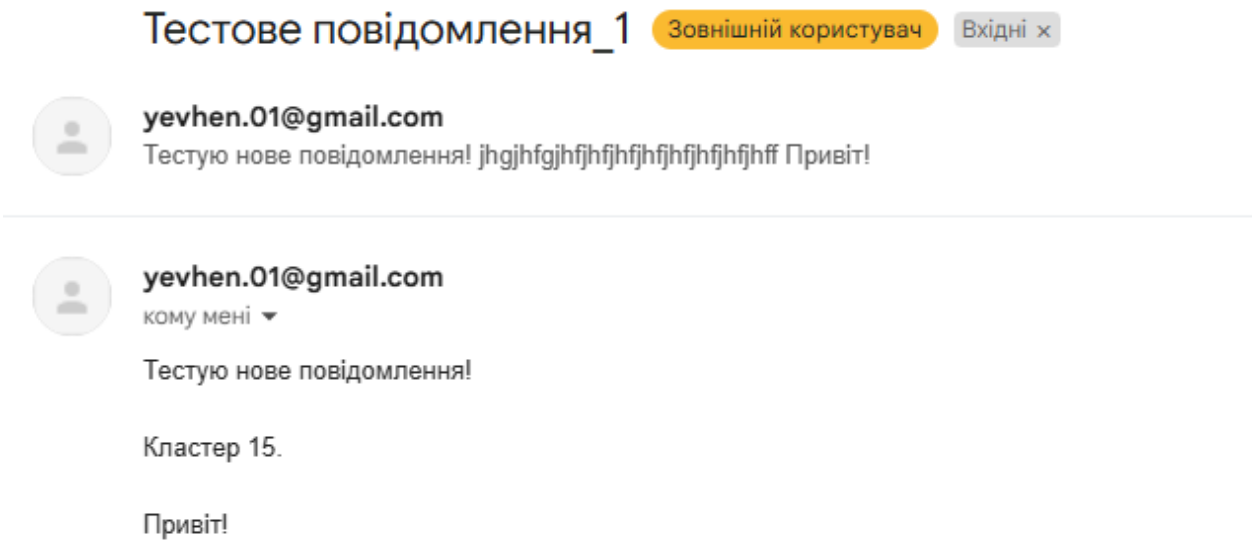

Рисунок 4.3 - Приклад використання розсилки

4.8 Блок оцінки моделі

4.8.1 Оцінка ефективності моделі під час навчання

У сучасному маркетингу, де дані стали основним ресурсом прийняття рішень, важливо не лише розробити ефективну модель кластеризації клієнтів, а й постійно відстежувати її якість у реальному часі. Здатність адаптуватися до змін поведінки користувачів і оцінювати результативність кожного етапу кампанії — це перевага, яка визначає успіх у динамічному бізнес-середовищі.

Кластеризація дозволяє виділити сегменти клієнтів, які демонструють подібні характеристики або поведінкові патерни. Проте без належної оцінки її ефективності моделі ризикують створити штучні або нестабільні сегменти, що можуть ввести маркетологів в оману. У цьому контексті метрики стають критично важливим елементом контролю якості. Вони виступають «датчиками», що

сигналізують, наскільки точним є групування, чи стабільні кластери з часом, і чи можна на них покладатися для прийняття рішень.

Наприклад, традиційні метрики кластеризації, як-от Silhouette Score, Davies-Bouldin Index або Calinski-Harabasz Score, дозволяють оцінити, наскільки чітко розділені кластери, чи віддалені вони один від одного. Під час кластеризації клієнтів були отримані наступні стандартні метрики:

Silhouette Score = 0.123 - це низьке значення індексу силуету, що вказує на слабке розділення кластерів. Ймовірно, деякі об'єкти не мають чіткої належності до певного кластера, або самі кластери перекриваються. Це може свідчити про складність самої вибірки.

Davies-Bouldin Index = 1.77 - цей показник є метрикою, що інтерпретується як “чим менше, тим краще”. Значення 1.77 свідчить про середню якість кластеризації: внутрішньо кластерна схожість не надто висока, і деякі кластери, можливо, мають значне перекриття.

Calinski-Harabasz Score = 4505.36 - високе значення цієї метрики вказує на чітке розмежування кластерів, однак ця метрика може бути не надто інформативною в ізоляції, особливо коли інші метрики показують слабку кластеризацію. Вона добре працює, коли розподіл даних майже ідеальний, але у реальних даних часто потребує додаткового контексту.

Для того щоб обрати чи підходить кластеризація, було реалізовано метрику Cluster Stability Index (CSI) — метрика, що дозволяє виміряти стабільність кластерів під час повторного тренування моделі. Вона оцінює, наскільки стабільно об’єкти (клієнти) залишаються в одних і тих самих кластерах при зміні вхідних умов (даних, параметрів, ініціалізацій). Наприклад, отриманий результат CSI = 0.86 свідчить про високу стабільність: клієнти переважно зберігають свою приналежність до певного сегменту. Це означає, що навіть за умов неідеальної сегментації, як показали стандартні метрики, кластери мають чітку структуру та можуть бути використані як основа для бізнес-рішень.

У маркетингу кінцева мета — не просто групувати клієнтів, а діяти на основі цього групування: розсилати персоналізовані повідомлення, пропонувати індивідуальні знижки, формувати лояльність. І тут в гру вступають бізнес-метрики, зокрема конверсія у клік (CTR). Якщо навіть найскладніша модель кластеризації не забезпечує підвищення CTR або рівня відкриття листів (OR), вона втрачає своє практичне значення. Тому оцінка ефективності моделі повинна здійснюватися як через технічні, так і бізнес-показники.

Використання CSI дозволяє отримати довгострокову оцінку якості моделі: чи змінюється структура сегментів із часом, чи залишається вона релевантною, навіть якщо поведінка клієнтів еволюціонує. Завдяки цьому метрика стає не просто технічним показником, а стратегічним інструментом. Стабільні кластери дозволяють формувати сталі комунікаційні стратегії, а нестабільні — сигналізують про потребу в гнучкому, адаптивному підході.

Таким чином, оцінка ефективності моделі кластеризації — це не одноразовий етап, а постійний процес, що поєднує технічні метрики, бізнес-показники та динаміку змін у даних. CSI стає ключовою метрикою, що дозволяє зберігати баланс між точністю кластеризації та її практичною користю для маркетингових кампаній, формуючи надійну основу для персоналізованої комунікації та зростання конверсії.

4.8.2 Оцінка моделі під час роботи моделі

У сучасному світі аналітики та дата-сайентисти дедалі частіше зіштовхуються з проблемою зміни поведінки користувачів, зовнішнього середовища чи навіть самих даних. Модель, яка вчора передбачала поведінку клієнтів з високою точністю, сьогодні може давати зовсім інші результати. Причина — зміна розподілу ознак, на яких вона навчалась. Саме тут на допомогу приходить Population Stability Index (PSI) — показник стабільності розподілу

змінних, який є надзвичайно важливим інструментом у світі бізнес-аналітики, кредитного скорингу, маркетингу та машинного навчання.

Population Stability Index (PSI) — показник, який дозволяє оцінити, наскільки змінився розподіл ознак між навчальними та актуальними даними. У контексті роботи моделей машинного навчання, аналітики або маркетингової стратегії PSI стає ключем до стабільності, точності та адаптивності бізнес-рішень.

PSI допомагає виявляти зміни у поведінці клієнтів, структурі аудиторії або змінах у самих даних, які можуть впливати на ефективність моделі чи маркетингової стратегії. Наприклад:

1. Якщо ви навчили модель передбачення відтоку клієнтів на історичних даних, а структура нових користувачів суттєво змінилась — модель більше не є актуальною.
2. У маркетингу, якщо кампанія орієнтована на аудиторію, яка вже змінилася — бюджет витрачається даремно.

Регулярний моніторинг PSI дозволяє своєчасно виявити ці зсуви та адаптувати стратегії.

PSI вимірює ступінь відхилення розподілу змінної у нових даних порівняно з історичними або базовими. Це схоже на перевірку температури в пацієнта: коли вона стабільна — все добре, але коли значно змінюється — потрібно вжити заходів. У бізнесі це дозволяє уникнути ситуацій, коли, наприклад, скорингова модель банку ухвалює рішення на основі застарілих шаблонів, що більше не відповідають поведінці клієнтів.

Значення PSI < 0.1 вказує на стабільність. Якщо PSI зростає до 0.25, це сигнал для аналізу. Значення > 0.25 — це вже "червоний прапорець": ситуація потребує негайної уваги, можливо, навіть переобучення моделі або зміни підходу до сегментації аудиторії.

Одна з переваг PSI полягає в його простоті: він легко обчислюється, зрозумілий навіть нефахівцям та не вимагає складних моделей. Його можна

використовувати для числових і категоріальних змінних, на різних етапах життєвого циклу моделі або продукту.

Крім того, PSI добре масштабується — його можна використовувати одразу на десятках змінних, автоматизуючи аналіз та інтегруючи у регулярну систему моніторингу.

Проте, як і будь-який інструмент, PSI має обмеження. Він не бачить змін у взаємозв'язках між змінними, не відображає зміни в таргеті, і може бути нестабільним при малих обсягах даних. Тому він має використовуватись разом з іншими метриками, такими як колінеарність, AUC, Gini, або feature importance.

У поєднанні з технічними інструментами (наприклад, у вигляді Python-класів) PSI може стати основою для автоматичного моніторингу моделі у production-середовищі — без потреби вручну перевіряти стабільність кожної ознаки.

Тому був реалізований клас PSICalculator — це Python-реалізація для автоматичного обчислення PSI. Він був розроблений як універсальний інструмент для перевірки стабільності змінних між тренувальними та новими (актуальними) даними.

На відміну від разових обчислень у вигляді функцій, клас надає структурований, повторно використовуваний підхід до моніторингу PSI для однієї чи багатьох змінних одночасно.

Опис функцій класу PSICalculator:

1. __init__(self, bins=10) - Конструктор класу, де можна задати:
    a. кількість бінів для розбиття розподілу ознак (bins=10 за замовчуванням),
    b. або передати власноруч створені межі бінів.
2. calculate_psi(self, expected, actual, bucket_type='quantile'):
    a. обчислює PSI між двома наборами даних (наприклад, train_data і new_data) по одній змінній,

b. підтримує два типи біннінгу: "quantile" (за квантилями) або "uniform" (рівномірні інтервали),
c. повертає одне числове значення PSI (типу float), яке показує, наскільки сильно змінився розподіл змінної.

3. calculate_psi_for_dataframe(self, df_expected, df_actual, bucket_type='quantile'):
   a. обчислює PSI по всіх спільних числових змінних між двома датафреймами (наприклад, train і production),
   b. автоматично ігнорує колонки, які неможливо обробити (наприклад, текст або дати),
   c. повертає результат у вигляді таблиці (DataFrame) з колонкою PSI, відсортованої за значенням.

Приклад використання і перевірка коректності моделі показана на рис. 4.4:

	PSI
age	0.0003
balance	0.0003

Рисунок 4.4 - Приклад використання PSICalculator для 2-х предикторів

4.9 Висновки

У результаті виконання роботи було реалізовано програмний продукт, що поєднує методи Data Science з інструментами автоматизації маркетингових процесів.

Розроблена інформаційна технологія включає комплексну систему, яка здійснює автоматичне отримання, обробку та аналіз даних, кластеризацію клієнтів, побудову прогнозів щодо їх поведінки, а також формування і надсилання персоналізованих маркетингових пропозицій через email-розсилку. Система має

модульну архітектуру, що дає змогу адаптувати її під різні бізнес-завдання та легко інтегрувати з іншими інформаційними системами.

Особливу увагу було приділено створенню зручної та гнучкої структури, яка дозволяє змінювати або розширювати функціонал у відповідності до специфіки конкретного підприємства чи галузі. Зокрема, кожен модуль системи може бути налаштований або замінений без потреби повної перебудови технології.

Застосування розробленої моделі відкриває широкі перспективи для бізнесу, серед яких:

1. підвищення ефективності маркетингових кампаній за рахунок персоналізації;
2. збільшення доходів компанії шляхом точнішого таргетування клієнтів;
3. покращення рівня задоволеності та лояльності клієнтів;
4. зменшення витрат на рекламу за рахунок оптимізації цільової аудиторії;
5. оперативне реагування на зміни у поведінці споживачів та ринку загалом.

Таким чином, результати роботи підтверджують ефективність застосування інформаційної технології на основі кластерного аналізу для вирішення прикладних завдань у сфері маркетингу. Запропонована система має значний потенціал для практичного впровадження в діяльність підприємств, які орієнтовані на клієнтоцентричну стратегію розвитку.

ЗАГАЛЬНІ ВИСНОВКИ

У сучасному конкурентному бізнес-середовищі, де уподобання клієнтів постійно змінюються, ефективна сегментація клієнтів є надзвичайно важливою. Кластеризація стала важливим інструментом для покращення маркетингових стратегій завдяки точній сегментації клієнтів, що призводить до більш персоналізованих підходів. Це допомагає компаніям оптимізувати розподіл ресурсів, покращити залучення клієнтів і значно збільшити рентабельність інвестицій (ROI).

У результаті виконання даної роботи проведено аналіз сучасних моделей клієнтської сегментації та методів їх застосування в рамках підвищення ефективності маркетингових кампаній. Досліджено основні підходи до сегментації клієнтів, зокрема демографічну, поведінкову, RFM-аналіз, кластеризацію та інші методи, засновані на застосуванні інструментів Data Science. У процесі аналізу розглянуто переваги й недоліки різних моделей сегментації, а також способи підготовки, очищення та інтерпретації даних для побудови точних і релевантних сегментів. Визначено, що найбільш ефективними моделями для сегментації клієнтів є методи кластерного аналізу, зокрема алгоритм k-середніх (яка в результатів була використана), ієрархічна кластеризація, які дозволяють створювати гнучкі, динамічні сегменти на основі великого масиву даних. Результати дослідження доводять, що впровадження моделей сегментації клієнтів сприяє підвищенню релевантності маркетингових повідомлень, зростанню коефіцієнтів відкриття та конверсії, а також загальному покращенню взаємодії з клієнтами.

В результаті роботи було створено продукт який складається з 7 блоків, блок зберігання даних, блоки збору і записів даних, навчання моделі, прогнозу, email розсилки, блок оцінки моделі. Дані блоки були побудовані так, що кожний бізнес може адаптувати під свої потреби і її покращувати, оскільки вони покривають майже всі процеси роботи з сегментами - від збору даних до автоматичних

розсилок. Також у роботі була навчена кластеризація на даних клієнтів в банку, що в результаті дало можливість створити кластери, їх проаналізува і на основі їх надати CVP. Кластеризація була обрана замість звичайної сегментації, оскільки вона дозволяє підприємствам виявляти приховані закономірності у великих наборах даних, групуючи клієнтів зі схожою поведінкою. Орієнтуючись на конкретні сегменти за допомогою індивідуальних пропозицій, рекламних акцій і комунікаційних стратегій, компанії можуть підвищити рівень конверсії та задоволеність клієнтів. Наприклад, використовуючи детальну сегментацію, ми спостерігали збільшення рентабельності інвестицій на 25% порівняно з маркетинговими стратегіями, які не мали такої розширеної сегментації.

Щоб забезпечити надійність сегментації, важливо запровадити надійні методи перевірки даних, такі як індекс стабільності кластера (CSI), який був реалізований в даній роботі. Цей показник оцінює, наскільки кластери стабільні з часом, гарантуючи, що сегменти залишаються послідовними та релевантними. У нашому аналізі ми досягли CSI 86,21% точок даних, які залишилися в одному кластері під час кількох прогонів. Така висока стабільність вказує на надійні сегменти, які дозволяють підприємствам створювати ефективні та довгострокові маркетингові стратегії, навіть коли традиційні метрики як Силует, показують невисокі показники, оскільки вони не показують стабільності кластерів, які можуть змінюватися, що призведе до неефективних кампаній і марної витрати ресурсів.

Використовуючи CSI та постійно вдосконалюючи сегментацію, підприємства можуть адаптуватися до зміни поведінки клієнтів, гарантуючи, що їхні маркетингові стратегії залишатимуться актуальними, а рентабельність інвестицій зростатиме.

Отже, сегментація є необхідною складовою сучасної маркетингової стратегії, що сприяє досягненню стратегічних бізнес-цілей за рахунок більш

гнучкого, адресного й результативного підходу до управління взаєминами з клієнтами.

Ключовою перевагою реалізованого підходу є його гнучкість і модульність, що дозволяє адаптувати систему до специфічних потреб різних бізнесів. Завдяки чітко структурованій архітектурі рішення — від збору даних до автоматизованої розсилки — підприємства можуть не лише зекономити час і ресурси, але й забезпечити постійну актуалізацію клієнтських сегментів відповідно до змін у поведінці аудиторії. Особливістю підходу також є орієнтація на практичну цінність: впровадження індексу стабільності кластерів дозволяє уникнути поширених помилок при короткостроковому використанні нестійких моделей сегментації, а отже, забезпечує довготривалу ефективність стратегії.

Одним із напрямів покращення може стати інтеграція зовнішніх даних, зокрема поведінкових сигналів із соціальних мереж або геолокаційної інформації, що дозволить ще глибше зрозуміти потреби та інтереси клієнтів. Також варто розглянути застосування гібридних моделей, які поєднують як традиційні алгоритми кластеризації, так і сучасні методи на основі глибокого навчання, що можуть виявляти складніші патерни. Додатково, підвищення рівня автоматизації, зокрема автоматичний моніторинг змін у стабільності кластерів та динамічне оновлення сегментів, дозволить зменшити втручання людини та оперативніше реагувати на зміни у поведінці клієнтів.

Таким чином, побудована система вже демонструє високу ефективність і здатна забезпечити компанії конкурентну перевагу. Разом із тим, її гнучкість і масштабованість відкривають широкі можливості для подальших удосконалень, що сприятиме ще більш точній персоналізації, глибшому розумінню клієнтів і досягненню довгострокових бізнес-цілей.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Shuhai Wang, Linfu Sun & Yang Yu. A dynamic customer segmentation approach by combining LRFMS and multivariate time series clustering. Scientific Reports. - 2024 - 18 p. https://doi.org/10.1038/s41598-024-68621-2
2. P. W. Farris, N. T. Bendle, P. E. Pfeifer, and D. J. Reibstein, Marketing Metrics: The Definitive Guide to Measuring Marketing Performance, Pearson Education, 2010.
3. Mosaddegh, A., Albadvi, A., Sepehri, M. M. & Teimourpour, B. Dynamics of customer segments: A predictor of customer lifetime value. Expert Syst. Appl. 172, 114606. https://doi.org/10.1016/j.eswa.2021.114606 (2021).
4. Patel, V., & Mehta, K. (2023). Customer Segmentation Using K-means Clustering. International Journal of Data Mining & Knowledge Management Process, 8(2), 12-28.
5. M. Wedel and W. A. Kamakura, Market Segmentation: Conceptual and Methodological Foundations, Springer Science & Business Media, 2012.
6. Kumar, A., & Sharma, R. (2022). RFM Analysis for Customer Segmentation. International Journal of Business Intelligence Research, 13(1), 23-45.
7. Sengupta, K. (2023). A Review of Customer Segmentation Methods with Applications in Marketing. Journal of Marketing Analytics, 11(2), 145-162.
8. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, Elsevier, 2011.
9. I.T. Jolliffe, Principal Component Analysis, Springer Series in Statistics, 2002.
10. Xiaotong Li, Young Sook Lee Customer Segmentation Marketing Strategy Based on Big Data Analysis and Clustering Algorithm. Journal of Cases on Information Technology (JCIT) 26(1). 2024. - 16 p. DOI: 10.4018/JCIT.336916
11. Python Data Science Handbook. [Online]. Available at: https://jakevdp.github.io/PythonDataScienceHandbook/

12. Market Segmentation in 12 minutes – YouTube: https://www.youtube.com/watch?v=LbYv2RWE4Tk
13. Market Segmentation: More Than Just A Phrase – Forbes: https://www.forbes.com/councils/forbesbusinessdevelopmentcouncil/2022/08/01/market-segmentation-more-than-just-a-phrase/
14. Market Segmentation (With Real World Examples) – YouTube: https://www.youtube.com/watch?v=IrJ1cNIfmsk
15. Revisiting the strategic role of market segmentation: Five themes for ... – ScienceDirect: https://www.sciencedirect.com/science/article/abs/pii/S0019850124001202
16. An Introduction to Market Segmentation – YouTube: https://www.youtube.com/watch?v=hnz1kClvHcs
17. Using Enhanced Marketing Segmentation To Engage Audiences – Forbes: https://www.forbes.com/councils/forbescommunicationscouncil/2024/07/01/using-enhanced-marketing-segmentation-to-engage-audiences/
18. What is Market Segmentation? – YouTube: https://www.youtube.com/watch?v=7FtiPAlBdto
19. A Guide to Market Segmentation – Fitchburg State University Online: https://online.fitchburgstate.edu/degrees/business/mba/marketing/understanding-market-segmentation/
20. Rediscovering Market Segmentation – Harvard Business Review: https://hbr.org/2006/02/rediscovering-market-segmentation
21. How Gymshark use Market Segmentation Explained – YouTube: https://www.youtube.com/watch?v=ZUOCx1iO8gI
22. Investopedia “Understanding Market Segmentation: A Comprehensive Guide” https://www.investopedia.com/terms/m/marketsegmentation.asp

23. Qualtrics “Market Segmentation: Definition, Types, Benefits, & Best Practices” https://www.qualtrics.com/experience-management/brand/what-is-market-segmentation/
24. YouTube “How to Define Your Target Market | 4 Types of Market Segmentation” https://www.youtube.com/watch?v=_S-JUX_k2dQ
25. Investopedia “How to Get Market Segmentation Right” https://www.investopedia.com/ask/answers/061615/what-are-some-examples-businesses-use-market-segmentation.asp
26. YouTube “What is Market Segmentation?” https://www.youtube.com/watch?v=7FtiPAlBdto
27. Shuhai Wang, Linfu Sun & Yang Yu. A dynamic customer segmentation approach by combining LRFMS and multivariate time series clustering. Scientific Reports, 2024. https://doi.org/10.1038/s41598-024-68621-2
28. Mosaddegh, A., Albadvi, A., Sepehri, M. M. & Teimourpour, B. Dynamics of customer segments: A predictor of customer lifetime value. Expert Systems with Applications, 2021. https://doi.org/10.1016/j.eswa.2021.114606
29. Patel, V., & Mehta, K. (2023). Customer Segmentation Using K-means Clustering. International Journal of Data Mining & Knowledge Management Process, 8(2), 12–28.
30. Kumar, A., & Sharma, R. (2022). RFM Analysis for Customer Segmentation. International Journal of Business Intelligence Research, 13(1), 23–45.
31. Sengupta, K. (2023). A Review of Customer Segmentation Methods with Applications in Marketing. Journal of Marketing Analytics, 11(2), 145–162.
32. Xiaotong Li, Young Sook Lee. Customer Segmentation Marketing Strategy Based on Big Data Analysis and Clustering Algorithm. Journal of Cases on Information Technology, 2024. DOI: 10.4018/JCIT.336916
33. Wedel, M., & Kamakura, W. A. (2012). Market Segmentation: Conceptual and Methodological Foundations. Springer.

34. Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Elsevier.
35. Jolliffe, I. T. (2002). Principal Component Analysis. Springer Series in Statistics.
36. Farris, P. W., Bendle, N. T., Pfeifer, P. E., & Reibstein, D. J. (2010). Marketing Metrics: The Definitive Guide to Measuring Marketing Performance. Pearson Education.
37. Dolnicar, S., Grün, B., & Leisch, F. (2018). Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful. Springer.
38. Tsiptsis, K., & Chorianopoulos, A. (2011). Data Mining Techniques in CRM: Inside Customer Segmentation. Wiley.
39. Harvard Business Review. Rediscovering Market Segmentation. https://hbr.org/2006/02/rediscovering-market-segmentation
40. Investopedia. Understanding Market Segmentation: A Comprehensive Guide. https://www.investopedia.com/terms/m/marketsegmentation.asp
41. Qualtrics. Market Segmentation: Definition, Types, Benefits, & Best Practices. https://www.qualtrics.com/experience-management/brand/what-is-market-segmentation/
42. Forbes. Market Segmentation: More Than Just A Phrase (2022). https://www.forbes.com/councils/forbesbusinessdevelopmentcouncil/2022/08/01/market-segmentation-more-than-just-a-phrase/
43. Fitchburg State University. A Guide to Market Segmentation. https://online.fitchburgstate.edu/degrees/business/mba/marketing/understanding-market-segmentation/
44. ScienceDirect. Revisiting the strategic role of market segmentation: Five themes. https://www.sciencedirect.com/science/article/abs/pii/S0019850124001202
45. Forbes. Using Enhanced Marketing Segmentation To Engage Audiences (2024). https://www.forbes.com/councils/forbescommunicationscouncil/2024/07/01/using-enhanced-marketing-segmentation-to-engage-audiences/

Додаток А

Результати аналізу кластерів моделі

№	Особливість кластера	Маркетингова стратегія
0	Переважно працівники «синіх комірців» (100%), одружені (83%), середній вік – 50 років, високий рівень утримання житла (57%)	Фокус на програми лояльності для робітничих спеціальностей, пропозиції фінансових послуг для власників житла
1	Велика частка технічних спеціалістів (29%), середній баланс нижче середнього, низька залученість у фінансові продукти	Оферти на накопичувальні депозити, страхування та освітні програми для кар'єрного зростання
2	Молодша аудиторія (35 років), високий відсоток найманих працівників (технічні, сервісні спеціалісти), 90% мають середню освіту	Діджитал-маркетинг через соціальні мережі, акцент на короткострокові фінансові продукти
3	Різноманітний склад за професіями, велика частка одиноких (47%), активні клієнти з частими зверненнями до банку (pdays=147)	Індивідуалізовані кредитні програми, таргетовані email-розсилки
4	Високий відсоток працівників у виробництві (31%), середній баланс, високий рівень іпотеки (87%)	Оферти на перекредитування, програми рефінансування та фінансового планування
5	Всі клієнти працюють у сфері «синіх комірців», більшість одружені, низький рівень фінансової активності	Програми страхування здоров’я, бонуси за довготривале використання банківських послуг
6	Найвищий середній баланс (9761), старший вік, більшість зайняті у керівництві	Прямий банкінг, ексклюзивні пропозиції VIP-клієнтам, преміальні картки

7	Високий відсоток управлінців (70%), вища освіта (95%)	Інвестиційні програми, консультації з фінансового планування
8	Унікальний кластер – лише 1 клієнт	Даний кластер буде пропущений
9	Високий вік (56), значний відсоток пенсіонерів (21%), середній баланс вищий за середній	Пенсійні та заощаджувальні програми, підключення автоматичних платежів
10	Велика частка управлінців (56%), молодий вік (33), переважно одинокі (54%)	Діджитал-банкінг, мобільні додатки з фінансовим плануванням, гейміфікація у програмах лояльності
11	Високий середній баланс (30111), активні фінансово клієнти, багато управлінців (44%)	Ексклюзивні депозитні пропозиції, індивідуальне обслуговування
12	Великий відсоток зайнятих у сервісній сфері (17%), середній баланс, середній вік	активний email-маркетинг, персоналізовані офери кредитних продуктів
13	Найстарший кластер (63 роки), переважно пенсіонери (76%), низька активність у фінансових продуктах	Освітні кампанії про накопичувальні програми, знижки на послуги для літніх людей
14	Високий рівень утримання житла (67%), середній баланс, вік – 40 років	Страхові програми на нерухомість, персональні кредитні пропозиції
15	Молодий кластер (31 рік), великий відсоток зайнятих у технічних професіях (23%)	Онлайн-просування фінансових продуктів, крос-продажі кредитних карток

Додаток Б

Скрипт розробленої системи

```
from matplotlib import pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
import joblib
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline
import_data = pd.read_excel(r'C:\Users\Yevhen\Desktop\Магістерська\data\train.xlsx')

import_data.head()
numeric_features = ['age', 'balance', 'pdays', 'previous']
categorical_features = ['job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'poutcome', 'deposit']
train_data = import_data.drop(columns=['id', 'mail'])
preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), numeric_features),
                ('cat', OneHotEncoder(handle_unknown='ignore', sparse_output=False), categorical_features)
```

```
    ]
)
pipeline = Pipeline([
    ('preprocessor', preprocessor),
        ('kmeans', KMeans(n_clusters=16, init='random', n_init=10, max_iter=500,
random_state=42))
])
pipeline.fit(train_data)
joblib.dump(pipeline,
r'C:\Users\Yevhen\Desktop\Магістерська\result\clustering_pipeline.pkl')
from matplotlib import pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
import joblib
import os
from google.cloud import bigquery, bigquery_datatransfer
import google.auth
import time
import pandas as pd
import datetime
import json
import warnings
```

```
import db_dtypes
warnings.filterwarnings('ignore')
%matplotlib inline
loaded_pipeline_segm_1 =
joblib.load(r'C:\Users\Yevhen\Desktop\Магістерська\result\clustering_pipeline.pkl')
model_path = r'C:\Users\Yevhen\Desktop\Магістерська\result\clustering_pipeline.pkl'
credentials_path =
r"C:\Users\Yevhen\Desktop\Магістерська\connect\big-query-36535-be80901cd2b8.json
"
project_id = "big-query-36535"
os.environ["GOOGLE_APPLICATION_CREDENTIALS"] =
r"C:\Users\Yevhen\Desktop\Магістерська\connect\big-query-36535-be80901cd2b8.json
"
query = """
    SELECT * FROM `big-query-36535.Master.info_about_clients`
  """
class BigQueryClient:
  def __init__(self, credentials_path, project_id):
    # Налаштування облікових даних для доступу до BigQuery
    os.environ["GOOGLE_APPLICATION_CREDENTIALS"] = credentials_path
    self.client = bigquery.Client(project=project_id)

  def execute_query(self, query):
    # Виконання SQL-запиту до BigQuery та отримання результату як DataFrame
    return self.client.query(query).to_dataframe()

  def convert_columns_to_string(self, df, columns):
    # Перетворення вказаних колонок у формат рядка
```

```
        df[columns] = df[columns].astype(str)
        return df

    def insert_data(self, df, table_id, if_exists="append"):
        """
        Вставка даних у таблицю BigQuery.

        :param df: Pandas DataFrame для завантаження
        :param table_id: Повний шлях до таблиці "project_id.dataset_id.table_name"
            :param if_exists: Як обробляти дані, якщо таблиця існує ("append" або "replace")
        """
         job_config = bigquery.LoadJobConfig(write_disposition="WRITE_APPEND")  # За замовчуванням додаємо дані

        if if_exists == "replace":
                job_config.write_disposition = "WRITE_TRUNCATE"  # Перезаписуємо таблицю

        job = self.client.load_table_from_dataframe(df, table_id, job_config=job_config)
        job.result()
        print(f"Дані успішно вставлено в {table_id}!")

    def upsert_data(self, df, table_id, id_column, update_columns):
        """
        Виконує UPSERT (INSERT або UPDATE) у BigQuery.

        :param df: Pandas DataFrame з даними
```

```
        :param table_id: Повний шлях до таблиці "project_id.dataset_id.table_name"
        :param id_column: Назва унікального ідентифікатора (наприклад, "id")
         :param update_columns: Список колонок, які потрібно оновити при наявності ID
        """
        # Завантажуємо дані у тимчасову таблицю
        temp_table = f"{table_id}_temp"
        job_config = bigquery.LoadJobConfig(write_disposition="WRITE_TRUNCATE")
                        job = self.client.load_table_from_dataframe(df, temp_table, job_config=job_config)
        job.result()

        # Створюємо MERGE SQL-запит
        update_expr = ", ".join([f"T.{col} = S.{col}" for col in update_columns])
        query = f"""
        MERGE `{table_id}` T
        USING `{temp_table}` S
        ON T.{id_column} = S.{id_column}
        WHEN MATCHED THEN
           UPDATE SET {update_expr}
        WHEN NOT MATCHED THEN
           INSERT ROW
        """

        # Виконуємо MERGE-запит
        self.client.query(query).result()
        print(f"UPSERT виконано для {table_id}!")
```

```
        # Видаляємо тимчасову таблицю
        self.client.delete_table(temp_table, not_found_ok=True)
bq_client = BigQueryClient(credentials_path, project_id)

NEW_DATA = bq_client.execute_query(query)

NEW_DATA = bq_client.convert_columns_to_string(NEW_DATA, ['default', 'housing',
'loan', 'deposit'])
class ClusterPredictor:
    def __init__(self, model_path):
        """
        Ініціалізація класу: завантаження моделі кластеризації.
        :param model_path: шлях до збереженого пайплайну
        """
        self.pipeline = joblib.load(model_path)

    def predict_clusters(self, new_data):
        """
        Передбачає кластери для нового набору даних.
        :param new_data: датафрейм з новими даними
        :return: датафрейм із доданою колонкою кластерних міток
        """
        # Видаляємо колонки, які не використовуються в кластеризації
        data_for_predict = new_data.drop(columns=['id', 'mail'])

        # Зберігаємо id та email для фінального результату
        CVP_data = new_data[['id', 'mail']].copy()
```

```
        # Передбачаємо кластери
        cluster_prediction = self.pipeline.predict(data_for_predict)

        # Додаємо результат у фінальний датафрейм
        CVP_data['cluster_prediction'] = cluster_prediction

        return CVP_data
predictor = ClusterPredictor(model_path)
clustered_data = predictor.predict_clusters(NEW_DATA)
import seaborn as sns
import matplotlib.pyplot as plt

# Побудова гістограми
ax = sns.histplot(data=clustered_data, x='cluster_prediction', discrete=True)

# Встановлення всіх значень на осі X
ax.set_xticks(sorted(clustered_data['cluster_prediction'].unique()))

plt.show()
bq_client_predict = BigQueryClient(credentials_path, project_id)

table_id = "big-query-36535.Master.clients_segments"
id_column = "id"
update_columns = ["mail", "cluster_prediction"]

bq_client_predict.upsert_data(clustered_data, table_id, id_column, update_columns)

import numpy as np
```

```
import pandas as pd

class PSICalculator:
    def __init__(self, bins=10):
        self.bins = bins

    def calculate_psi(self, expected, actual, bucket_type='quantile'):
        expected = pd.Series(expected).dropna()
        actual = pd.Series(actual).dropna()

        if isinstance(self.bins, int):
            if bucket_type == 'quantile':
                bin_edges = np.unique(np.percentile(expected,
                                    np.linspace(0, 100, self.bins + 1)))
            elif bucket_type == 'uniform':
                bin_edges = np.linspace(expected.min(), expected.max(), self.bins + 1)
            else:
                raise ValueError("bucket_type має бути 'quantile' або 'uniform'")
        else:
            bin_edges = np.array(self.bins)

        expected_counts = np.histogram(expected, bins=bin_edges)[0] / len(expected)
        actual_counts = np.histogram(actual, bins=bin_edges)[0] / len(actual)

        expected_counts = np.where(expected_counts == 0, 1e-6, expected_counts)
        actual_counts = np.where(actual_counts == 0, 1e-6, actual_counts)
```

```
            psi_values = (expected_counts - actual_counts) * np.log(expected_counts / actual_counts)
        psi = np.sum(psi_values)

        return psi

        def calculate_psi_for_dataframe(self, df_expected: pd.DataFrame, df_actual: pd.DataFrame, bucket_type='quantile'):
        """
        Порівнює однакові колонки в двох датафреймах.
        """
        common_columns = df_expected.columns.intersection(df_actual.columns)
        psi_results = {}

        for col in common_columns:
            try:
                                psi = self.calculate_psi(df_expected[col], df_actual[col], bucket_type=bucket_type)
                psi_results[col] = round(psi, 4)
            except Exception as e:
                psi_results[col] = f"Error: {e}"

                            return pd.DataFrame.from_dict(psi_results, orient='index', columns=['PSI']).sort_values(by='PSI', ascending=False)

numerical_columns = NEW_DATA.select_dtypes(include=['number']).columns
df_numerical = NEW_DATA[numerical_columns]
```

```
df_train = df_numerical
df_new = df_numerical.sample(frac=0.3)

psi_calc = PSICalculator(bins=10)
psi_table = psi_calc.calculate_psi_for_dataframe(df_train, df_new)

psi_table
import smtplib
import os
from email.message import EmailMessage

class EmailSender:
    def __init__(self, smtp_server, port, sender_email, app_password):
        """
        :param smtp_server: SMTP сервер (для Gmail це "smtp.gmail.com")
        :param port: Порт (для Gmail: 587)
        :param sender_email: Email відправника
        :param app_password: Пароль додатку (отримується у Google)
        """
        self.smtp_server = smtp_server
        self.port = port
        self.sender_email = sender_email
        self.app_password = app_password

    def load_text_from_file(self, file_path):
        """ Завантажує текстовий вміст із .txt файлу """
        if not os.path.exists(file_path):
            raise FileNotFoundError(f"Файл {file_path} не знайдено!")
```

```
        with open(file_path, "r", encoding="utf-8") as file:
            return file.read()

    def send_email(self, recipient_email, subject, message, attachment_path=None):
        """
        Надсилає email.
        :param recipient_email: Email отримувача
        :param subject: Тема листа
        :param message: Тіло листа
        :param attachment_path: Опціональний шлях до файлу для вкладення
        """
        # Формуємо email
        msg = EmailMessage()
        msg["From"] = self.sender_email
        msg["To"] = recipient_email
        msg["Subject"] = subject
        msg.set_content(message)

        # Додаємо вкладення (опціонально)
        if attachment_path:
            with open(attachment_path, "rb") as f:
                file_data = f.read()
                file_name = os.path.basename(attachment_path)
                                    msg.add_attachment(file_data,   maintype="application",
subtype="octet-stream", filename=file_name)

        # Підключення до SMTP сервера Gmail
```

```
        with smtplib.SMTP(self.smtp_server, self.port) as server:
            server.starttls()  # Захищене з'єднання
            server.login(self.sender_email, self.app_password)
            server.send_message(msg)

        print(f"Лист надіслано на {recipient_email}")
import smtplib
import os
from email.message import EmailMessage

class EmailSender:
    def __init__(self, smtp_server, port, sender_email, app_password):
        """
        :param smtp_server: SMTP сервер (для Gmail це "smtp.gmail.com")
        :param port: Порт (для Gmail: 587)
        :param sender_email: Email відправника
        :param app_password: Пароль додатку (отримується у Google)
        """
        self.smtp_server = smtp_server
        self.port = port
        self.sender_email = sender_email
        self.app_password = app_password

    def load_text_from_file(self, file_path):
        """ Завантажує текстовий вміст із .txt файлу """
        if not os.path.exists(file_path):
            raise FileNotFoundError(f"Файл {file_path} не знайдено!")
```

```
        with open(file_path, "r", encoding="utf-8") as file:
            return file.read()

    def send_email(self, recipient_email, subject, message, attachment_path=None):
        """
        Надсилає email.
        :param recipient_email: Email отримувача
        :param subject: Тема листа
        :param message: Тіло листа
        :param attachment_path: Опціональний шлях до файлу для вкладення
        """
        # Формуємо email
        msg = EmailMessage()
        msg["From"] = self.sender_email
        msg["To"] = recipient_email
        msg["Subject"] = subject
        msg.set_content(message)

        # Додаємо вкладення (опціонально)
        if attachment_path:
            with open(attachment_path, "rb") as f:
                file_data = f.read()
                file_name = os.path.basename(attachment_path)
                                    msg.add_attachment(file_data,   maintype="application",
subtype="octet-stream", filename=file_name)

        # Підключення до SMTP сервера Gmail
        try:
```

```
            with smtplib.SMTP(self.smtp_server, self.port) as server:
                server.set_debuglevel(1)  # Для відладки можна включити це
                server.starttls()  # Захищене з'єднання через TLS
                server.login(self.sender_email, self.app_password)
                server.send_message(msg)
            print(f"Лист надіслано на {recipient_email}")
        except smtplib.SMTPAuthenticationError:
            print("Помилка автентифікації. Перевірте правильність пароля додатку.")
        except smtplib.SMTPConnectError:
                print("Неможливо підключитися до сервера. Перевірте з'єднання та налаштування.")
        except Exception as e:
            print(f"Сталася помилка при надсиланні листа: {e}")


# Налаштування SMTP сервера та пароля додатку
SMTP_SERVER = "smtp.gmail.com"
PORT = 587
SENDER_EMAIL = "yevhen.01@gmail.com"

# Завантаження пароля додатку з текстового файлу
password_file_path = r'C:\Users\Yevhen\Desktop\Магістерська\connect\test.txt'
with open(password_file_path, "r", encoding="utf-8") as file:
    app_password = file.read().strip()

# Створюємо екземпляр класу EmailSender
email_sender = EmailSender(SMTP_SERVER, PORT, SENDER_EMAIL, app_password)
```

```
# Завантажуємо текст листа з .txt файлу
text_path = r"C:\Users\Yevhen\Desktop\Магістерська\CVP\кластер_15.txt"
message_content = email_sender.load_text_from_file(text_path)

# Надсилаємо email
email_sender.send_email(
    recipient_email="yevhen.01@knu.ua",  # Вказати отримувача
    subject="Тестове повідомлення_1",         # Вказати тему
    message=message_content,                # Текст листа
    attachment_path=None                  # Вкладення (не обов'язково)
)
```