

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ ІМЕНІ
ВАДИМА ГЕТЬМАНА**

**Навчально-науковий інститут
Інститут інформаційних технологій в економіці**

Кафедра математичного моделювання та статистики

Освітньо-професійна програма

Галузь знань

Спеціальність

Економічна кібернетика та Дата Сайнс

05 соціальні та поведінкові науки

051 економіка

Форма навчання: очна (денна)

КВАЛІФІКАЦІЙНА МАГІСТЕРСЬКА РОБОТА

на тему «Аналіз та моделювання цільової аудиторії в системах веб-аналітики»

(назва теми)

здобувача Васильєвої Катерини Сергіївни

(ПІБ, підпис)

Науковий керівник: кандидат економічних наук,
доцент Катуніна О.С.

(науковий ступінь, учене звання, ПІБ)

(підпис)

**Робота допущена до захисту перед екзаменаційною
комісією з атестації здобувачів вищої освіти (ЕК)**

Завідувач кафедри кандидат фізико-математичних наук,
професор Великоіваненко Г.І.

(підпис)

Київ 2023

ЗМІСТ

ВСТУП	3
РОЗДІЛ 1 ТЕОРЕТИЧНІ ЗАСАДИ ДОСЛІДЖЕННЯ ТА МОДЕЛЮВАННЯ ПОВЕДІНКИ СПОЖИВАЧІВ НА РИНКУ	6
1.1. Сутність та структура цільової аудиторії та поведінки споживачів у маркетингових дослідженнях.....	6
1.2. Особливості споживацької активності українців в умовах поточної економічної ситуації	11
1.3. Напрями дослідження та моделювання цільової аудиторії споживачів	14
РОЗДІЛ 2 МАТЕМАТИЧНІ МЕТОДИ ТА ІНСТРУМЕНТАРІЙ МОДЕЛЮВАННЯ СПОЖИВАЦЬКОЇ ПОВЕДІНКИ	18
2.1. Аналіз тенденцій поведінки цільової аудиторії засобами Google Analytics	4 18
2.2. Моделі сегментування цільової аудиторії та визначення латентних факторів поведінки споживачів	28
2.3. Моделювання поведінки споживачів на ринку методами машинного навчання	33
РОЗДІЛ 3 ПРАКТИЧНЕ ЗАСТОСУВАННЯ МАТЕМАТИЧНИХ МОДЕЛЕЙ ПОВЕДІНКИ ЦІЛЬОВОЇ АУДИТОРІЇ НА РИНКУ УКРАЇНИ.....	37
3.1. Аналіз результатів розвідувального аналізу та препроцесингу даних про активність цільової аудиторії на ринку	37
3.2. Аналіз результатів сегментування споживачів та визначення латентних факторів їхньої поведінки	47
3.3. Аналіз результатів моделювання поведінки цільової аудиторії на базі інструментів RFM та машинного навчання	60
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	74
ДОДАТКИ	77
Додаток А	77

ВСТУП

Беручи до уваги ряд кризових подій, які сталися у всесвітніх масштабах та масштабах нашої країни в тому числі, і безпосередньо вплинули на споживацькі настрої серед населення, необхідним є більш глибокий аналіз споживацької активності як українців, так і світових споживачів, що складає *актуальність* проведеного дослідження. Крім того, доречним є аналіз як поточних патернів цільової аудиторії підприємств, так і її поведінки в «стабільному» стані, адже, за позитивних прогнозів, це саме та поведінка, що властива споживачам при стабільній економічній ситуації.

Дослідження в сфері споживацької активності були проведені багатьма науковцями та суб'єктами бізнесу, серед яких Скригун Н.П., Романов А.А., Кучма І.З., Розумей С.Б., Семененко К.Ю., Сюсель Ю.В., Байдак Т.М., Болотова В.О., Зоська Я.В., Сторчовий Б.А. та інші. Значним вклад в статистичний аналіз та дослідження поведінки цільової аудиторії бізнесу був зроблений компанією Deloitte, який відображає зміни як в структурі поведінки, так і її зміну в динаміці.

Метою кваліфікаційної роботи є моделювання факторів формування попиту та поведінки споживачів на ринку та розробка комплексу моделей для аналізу цільової аудиторії на базі методів економіко-математичного моделювання, інтелектуального аналізу даних та машинного навчання.

Для досягнення поставленої мети у роботі вирішуються *завдання*:

- розглянути сутність, способи визначення цільової аудиторії та принципи формування поведінки споживачів на ринку, здійснити огляд літературних джерел та сучасних результатів наукових досліджень з теми роботи;
- проаналізувати специфіку та тенденції споживацької активності на світовому ринку та в Україні;
- здійснити огляд напрямів використання сучасного інструментарію дослідження товарних ринків та математичних моделей в маркетингових дослідженнях;

- дослідити фактори формування попиту та поведінки цільової аудиторії на ринку;
- сегментувати цільову аудиторію за комплексом характеристик та побудувати математичні моделі прогнозування здійснення покупок, що базуються на визначенні патернів споживацької активності;
- застосувати методи розвідувального аналізу та візуалізації даних для аналізу цільової аудиторії, виконати моделювання взаємозалежностей в системі факторів поведінки споживачів та спрогнозувати їхню активність на ринку на базі методів машинного навчання.

Об'єктом дослідження є поведінка споживачів та фактори формування цільової аудиторії на ринку.

Предметом дослідження є математичні методи та моделі аналізу тенденцій формування цільової аудиторії та поведінки споживачів у маркетингових дослідженнях.

Серед основних *методів дослідження* можна виділити: методи емпіричного і теоретичного аналізу були застосовані в процесі ознайомлення з теорією споживацької поведінки та поточним станом споживацької активності в Україні та світі; метод наукової абстракції був застосований в процесі виокремлення основних факторів, етапів підготовки та роботи з цільовою аудиторією бізнесу; веб-аналітика та статистичне спостереження були застосовані при аналізі та візуалізації статистичних даних про споживачів, отримані в результаті маркетингових кампаній досліджуваної фірми; за допомогою графічних методів були побудовані візуалізації проведеного дослідження; методи математичного моделювання допомагали про побудові моделей поведінки споживачів, її структури, розподілу цільової аудиторії на групи та прогнозування їх споживчої поведінки. Крім того, проведення дослідження гарантувалося методами комп'ютерного програмування та використанні пакетів машинного навчання в мові програмування Python.

Теоретичною й методичною основою дослідження є праці провідних вітчизняних та зарубіжних вчених з теорії й методології здійснення економіко-

математичного моделювання в маркетинговій діяльності, теорії та прикладних аспектів використання методів машинного навчання, наукові статті, статистична інформація.

Структура роботи. Робота складається зі вступу, трьох розділів, висновків, списку використаних джерел та додатків.

РОЗДІЛ 1

ТЕОРЕТИЧНІ ЗАСАДИ ДОСЛІДЖЕННЯ ТА МОДЕЛЮВАННЯ ПОВЕДІНКИ СПОЖИВАЧІВ НА РИНКУ

1.1. Сутність та структура цільової аудиторії та поведінки споживачів у маркетингових дослідженнях

Цільова аудиторія, за словами Скригун Н.П., є сукупністю потенційних або реально існуючих (фактичних) споживачів, які приймають рішення щодо купівлі або впливають на прийняття рішення [1]. Маркетингова кампанія, направлена на помилково визначену цільову аудиторію, не тільки не принесе позитивних результатів у помилково вибраної частини споживачів, але і може результувати у втрату можливості налагодження комунікації зі споживачами, які насправді є цільовою аудиторією даного товару.

Згідні з Романовим А.А., цільова аудиторія може перебувати в одному з семи станів купівельної готовності [2]:

- **Проінформованість**

Першочерговою задачею маркетолога є з'ясування того, чи знайомий цільовій аудиторії конкретний товар, його специфіка, якість, переваги перед іншими товарами, або про фірму, яка його виробляє. Представник цільової аудиторії може бути добре проінформованим про даний товар, що є гарним підґрунтям до здійснення покупки, може біти частково проінформованим, а може взагалі не знати про виробництво даного товару конкретною компанією.

- **Знання**

Цільова аудиторія може знати про існування фірми або продукції, що нею виготовляється, але не знати специфіку та переваги даних продуктів. Компанії рекомендується з'ясувати, яка частина потенційних покупців тільки чула про цей товар, яка частина володіє деякою інформацією про товар, і яка частина є знайомою

з товаром в деталях. З метою інформування споживача компанії також рекомендується проведення маркетингових кампаній, рекламних звернень та інші застосування інших засобів масової комунікації.

- Прихильність

На даному етапі є ймовірність ініціативи з боку споживача. Ініціатива покупці залежить значним чином від рівня його інформованості, а також від проведеної роботи компанії щодо формування позитивного ставлення до товарів власного виробництва. При позитивному усталенні споживача до товару, підприємству необхідно забезпечити найбільш прості способи придбання товару споживачем, однак при негативному усталенні фокус зміщується на виявлення причин даного характеру відносин.

- Перевага

Споживач може мати прихильність до даного товару, тобто існування позитивних асоціацій, пов'язаних з продуктом, однак це ще не показником того, що в порівнянні з аналогами споживач вибере саме цей товар. У цьому випадку компанії рекомендується здійснити заходи по формуванню переваги у покупців, підкреслюючи якість, особливості товару, а також можливі послуги, що супроводжують товар (наприклад, сервісне обслуговування) тощо.

- Переконаність

Можливою є ситуація, за якої цільова аудиторія віддаватиме перевагу даному товару перед іншими, але матиме сумніви щодо необхідності його придбання. Компанії в даній ситуації рекомендується провести роботу з цільовою аудиторією по наголошенню по на тому, які проблеми будуть вирішені після придбання товару, та переконати споживача в існуванні потреби до його купівлі.

Споживацька активність може бути визначена як деяка поведінка або окремі її патерни, яких притримуються споживачі або домогосподарства в процесі вибору, підготовки, здійснення покупки, і подальшого використання продукту або послуги [2]. На споживацьку активність може впливати велика множина факторів, серед яких психологічні, соціологічні, культурних елементи, а також безпосередній

вплив на споживача суб'єктами бізнесу з метою його заохочення до здійснення купівлі товару/послуги.

Споживацька активність є складним процесом, що включає в себе виявлення потреб споживача, збір даних про товари, ознайомлення з варіантами, прийняття рішення про покупку та подальшу оцінку товару/послуги. Впродовж цих етапів знаходяться під впливом, а, відповідно, і приймають рішення, залежно від особистих поглядів та цінностей, соціальних звичаїв, маркетингових кампаній, специфічних особливостей продукту, і навіть умов навколишнього середовища.

Розуміння специфіки цільової аудиторії має важливе значення для підприємств, і проводиться з метою створення останніми ефективних маркетингових планів та виготовлення і постачання саме тих товарів та послуг, що точно задовольняли б бажання та потреби клієнтів. З метою виокремлення тенденцій та закономірностей, прогнозування попиту і прийняття оптимального рішення щодо маркетингу відповідного товару, ціни, просування та розповсюдження підприємства проводять глибокий аналіз своїх (або в перспективі своїх) клієнтів, їх рівень задоволеності від продуктів, а також виокремлення груп споживачів за різними факторами особливостей поведінки при прийнятті рішення.

Аналіз активності та поведінки споживачів націлений виявлення [3]:

- ставлення споживачів до конкретних товарів або категорій товарів, а також до можливості їх заміни альтернативними (відносно брендів, продуктів тощо);
- мотивація споживачів та критерії вибору між різними товарами;
- характер поведінки споживачів в процесі пошуку продукту та здійснення покупки;
- ставлення оточення покупця до даного товару як фактор впливу на процес прийняття рішення суб'єктом споживання.

Класифікуючи цільову аудиторію підприємств за загальними патернами поведінки під час прийняття рішення, компанія може виокреми наступні групи:

1. Складна купівельна поведінка - така поведінка може спостерігатись, коли споживачі купують дорогий продукт і здійснюють подібні покупки

достатньо рідко. Покупець є активно залученим в процес аналізу ринку товарів, порівняння характеристик альтернатив та зваженого обмірковування ситуації з метою здійснення вдалої інвестиції.

2. Дисонансна поведінка, що знижує споживацьку активність - споживач глибоко занурений в процес покупки, але йому важко визначитись між двома альтернативними варіантами. «Дисонанс» може виникнути, коли споживач хвилюється, що пошкодує про свій вибір. Дана ситуація може вплинути на споживацьку активність довгою тривалістю прийняття рішення, а також можливою відмовою від здійснення покупки взагалі.
3. Купівельна поведінка за звичкою - звичні покупки характеризуються тим, що споживач поверхнево залучений до процесу вибору товару, однак добре визначений щодо бажаного об'єкту купівлі. Описана споживацька поведінка посереднім чином впливає на загальну споживацьку активність, однак іноді може негативним чином відобразитися на останній, адже покупець менш схильний до розширення інтересів та здійснення імпульсивного вибору, заснованого на прихильності до ново-відкритого товару.
4. Поведінка пошуку різноманітності характеризується здійсненням покупок споживачем не базуючись на необхідності заміни товару, бо він за певними характеристиками може не задовольняти споживача, а скоріше з вмотивованості споживача до змін та бажання розвитку. Дана поведінка є протилежною описаній раніше поведінці за звичкою. Поведінка пошуку різноманітності позитивно впливає на споживацьку активність загалом.

Сегментація цільової аудиторії велике значення для бізнесу, адже допомагає в процесі створення нових продуктів під конкретні умови задоволення потреб, коригування існуючих товарів, а також розвитку маркетингових кампаній задля розширення кількості нових споживачів та заохочення до більшої активності вже знайомих клієнтів.

Розглянемо один з існуючих варіантів сегментації цільової аудиторії за наступними показниками [3]:

1. Клієнти, що шукають задоволення потреби

Коли клієнти досліджують продукт або послугу, їхня поведінка може виявити цінну інформацію про те, які переваги, функції, цінності, варіанти використання чи проблеми є найбільш мотивуючими і які впливають на прийняття рішення щодо здійснення покупки. Коли клієнт наділяє одну чи декілька переваг більшою цінністю перед іншими, ці властивості товару є визначальними факторами мотивації, що спонукають клієнта прийняти рішення про покупку. Споживацька активність даного сегменту є вище середнього, адже споживачі мають проблемні і, відповідно, мотивуючі їх фактори, що можуть виступати як чинники підвищення активності.

2. Випадкові або сезонні

Активність даного сегменту цільової аудиторії підвищується при зростанні сезонних нагод до купівлі, як, наприклад, свята або сезон літнього відпочинку. Крім того, можна виділити і індивідуальні патерни, такі як, наприклад, день народження суб'єкта споживання, або його знайомих, певні звичаї, ритуали поведінки.

3. Віддання переваги певному бренду

Представниками даного патерну поведінки є покупці, що мають лояльний характер відносин з певним брендом або конкретним магазином. Тож, сегментація даних клієнтів можлива за рівнем лояльності, а їх споживацька активність може коливатися в залежності від випуску нової продукції або проведення маркетингових кампаній магазином.

4. Статус клієнта

Залежно від специфіки підприємства клієнти можуть мати різні статуси, однак можливі і спільні категорії:

- Нові покупці;
- Перспективні;
- Постійні покупці;

- Змінні покупці (колишні клієнти, які перейшли до конкурента) тощо.

1.2. Особливості споживацької активності українців в умовах поточної економічної ситуації

Компанією Delloite було проведено декілька досліджень споживацької активності українців у 2020 та 2022 роках. Розглянемо більш предметно перше з них.

У 2020 році представники компанії провели масове опитування українців щодо змін в обсягах, структурі їхніх витрат, поведінкових патернів та загальної споживацької активності [4]. До прикладу, розглянемо одне з питань в опитуванні (рис. 1.1).

? Як протягом наступних одного-двох років зміниться частота Ваших покупок у традиційних та онлайн-магазинах (офлайн vs онлайн)?



Рисунок 1.1– Розподіл населення українців за фактором очікуваної зміни споживацької поведінки у 2020 році [4]

Підбиваючи статистику відповідей тих покупців, які проголосували зміну або за не-зміну розподілу витрат в традиційних та онлайн магазинах, можемо побачити, що більше третини респондентів не збиралося змінювати споживацькі поведінкові патерни, і більше 50% збиралися змінювати патерни поведінки в сторону віддання переваги онлайн-магазинам. З точки зору аналітики, дана статистика є великим стимулом до розвитку інтернет-маркетингу та торгових

інтернет-майданчиків. З точки зору споживацької активності це може виступати як позитивним, так і негативним фактором, адже необхідні зусилля, які має прикласти споживач для здійснення покупки в онлайн магазині, мінімізуються, як і ймовірність здійснити ним імпульсивну покупку, яка найчастіше відбувається у фізичних точках компаній.

Як Ви оцінюєте зміну своєї купівельної активності під час карантину порівняно з докарантинним періодом у розрізі сегментів (відповіді «Без змін» та «Важко відповісти» не презентуються)?

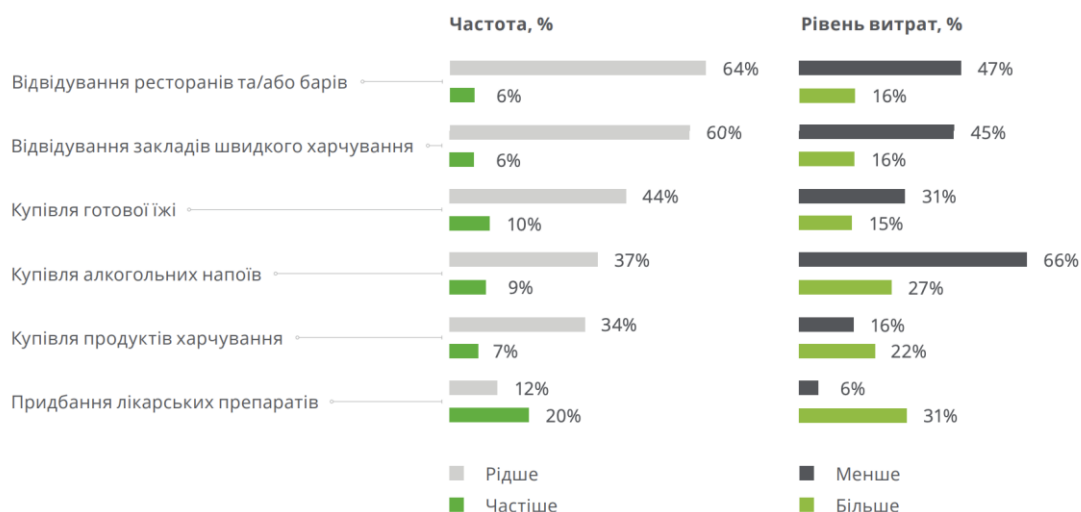


Рисунок 1.2 – Зміна купівельної активності споживачів в пост-ковідний період [4]

Аналізуючи статистику зміни споживацької активності суб'єктивно представлену респондентами про самих себе за показником частоти (frequency – однієї з RFM метрик) (рис. 1.2), можемо побачити, що він стрімко зростає у рамках відсоткової величини від 64% (відвідування ресторанів) до 12% (придбання лікарських препаратів). Остання зі статей, а саме придбання лікарських препаратів, є єдиною, споживацька активність якої є додатною – 12% респондентів знижують свою активність, в той час як 20% - підвищують.

Аналізуючи зміну споживацької активності за показником рівня витрат, можемо побачити скоріше негативну динаміку, тобто скорочення витрат. Особливо вирізняються такі статті скорочення як відвідування ресторанів та купівля алкогольних напоїв. Найбільше зростають витрати на купівлю лікарських препаратів.

Беручи до уваги більш актуальні тенденції сьогодення, спричинені розв’язанням воєнних дій на території України, компанією Delloite було проведене додаткове дослідження споживацької активності українців, змін в абсолютних розмірах витрат та їх структурі [5].

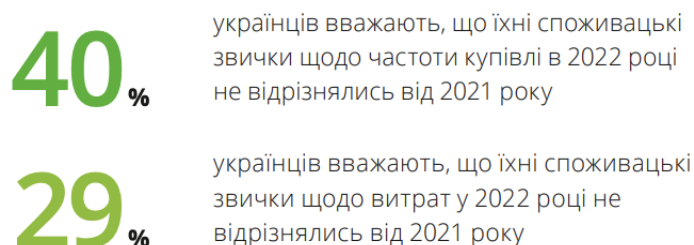


Рисунок 1.3 – Статистика змін споживацької поведінки українців у 2022 році [5]

За наведеною на рис. 1.3 статистикою можемо зробити висновки, що більшість українців змінила свої споживацькі звички щодо частоти купівлі в порівнянні з 2021 роком – 60%, та 71% українців вважають, що змінились їх споживацькі звички щодо витрат.

Однак за даною статистикою на даному етапі не можемо зробити висновки щодо підвищення/зниження активності споживачів, тільки про характер споживацької поведінки.

Аналізуючи показник частоти (рис. 1.4), можемо дійти висновку, що споживацька активність українців знизилась. Найбільшу різницю демонструє категорія відвідування закладів громадського харчування, найменшу – дитячі товари. В свою чергу, рівень витрат українців також зменшився в усіх досліджуваних категоріях, найбільші зміни зазнали косметичні товари та будівельні матеріали.

❓ Вкажіть, будь ласка, як ви оцінюєте зміну вашої купівельної активності в розрізі сегментів від 24 лютого 2022 року? (Варіант «Без змін» не презентується)

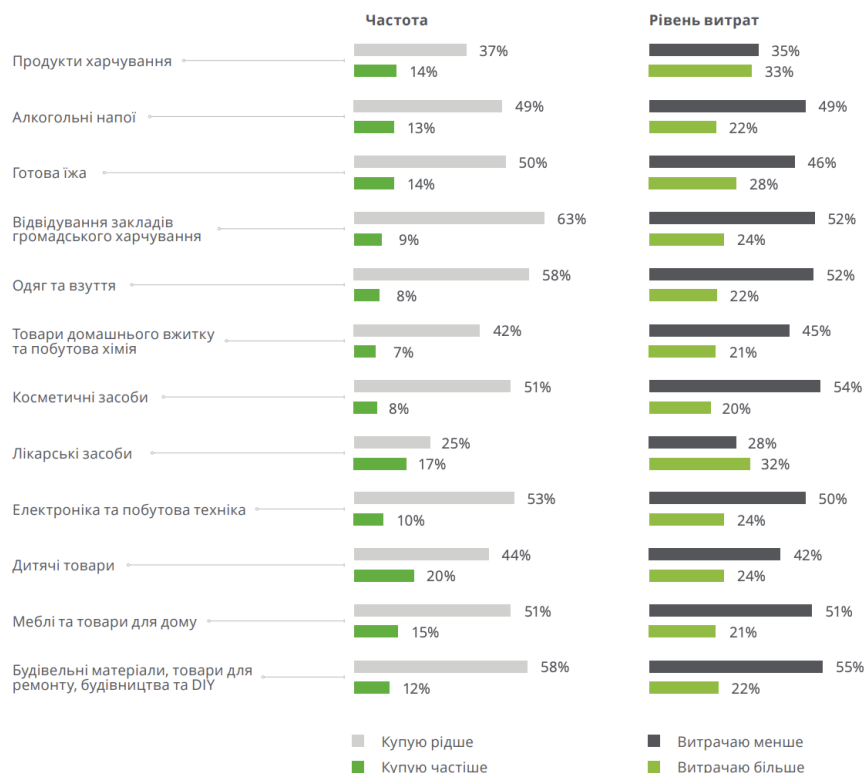


Рисунок 1.4 – Зміна купівельної активності споживачів у воєнний період [5]

1.3. Напрями дослідження та моделювання цільової аудиторії споживачів

В свою чергу, RFM аналіз є одним із маркетингових інструментів, що дозволяє охарактеризувати споживачів за наступними показниками [6]:

1. Час після здійснення останньої покупки (recency). Показник відображає у часовому вимірі кількість (переважно) днів, що пройшли після здійснення споживачем останньої покупки в даному магазині. Клієнти, які не так давно здійснили покупку, матимуть більш чітке враження від продукту та, скоріше за все, придбають або використають продукт знову.
2. Частота (frequency). Показник відображає відносну величину того як часто споживач робить покупки у компанії. Клієнти, які не є новими, і зробили

покупку хоча б один раз, імовірно зроблять покупку знову. Крім того, даний показник є прямо залежним від кількості та якості проведених рекламних кампаній підприємством.

3. Витрачені кошти (monetary). Показник відображає скільки грошей витратив клієнт за визначений період часу. Клієнти, які витрачають відносно більшу кількість коштів є більш цінними для бізнесу та мають високу ймовірність витратити кошти знову.

Період здійснення підрахунків може коливатися, і визначається підприємством залежно від цілей проведення аналізу.

Набір із описаних трьох значень відповідно до кожного споживача називається осередком RFM. Часто з даних показників знаходиться середнє значення і проводиться сортування клієнтів за отриманим RFM показником. Залежно від специфіки бізнесу деяким значенням може віддаватися більша перевага (вага) в порівнянні з іншими.

RFM аналіз також може бути корисним для неприбуткових організацій. Використання RFM-аналізу некомерційними та благодійними організаціями може бути корисним, наприклад, з позиції пошуку найкращих донорів, оскільки ті, хто робив пожертвування в минулому, з більшою ймовірністю зроблять пожертвування в майбутньому.

Крім того, організації, що не отримують прямий прибуток від споживачів товарів/послуг. Можуть також бути зацікавленими в проведенні такого аналізу. До прикладу, сайти, блоги, програми, які цінують читацьку аудиторію, кількість переглядів або взаємодію, можуть використовувати показники залученості замість грошової складової (monetary) в стандартному RFM аналізі.

Отже, модель RFM допомагає компанії визначити деяку універсальну ознаку, яку можна використовувати для проведення більш глибокого маркетингового аналізу та коригування маркетингової стратегії підприємства. За кожним із трьох основних компонент RFM показника клієнти отримують оцінку від 1 до 5 на основі спостережуваної купівельної активності споживача, де 5 вважається найкращим результатом [6]. Тому ідеальний клієнт мав би оцінку 5, 5,

5 за всіма трьома факторами. Клієнти з нижчими оцінками можуть бути віднесені до конкретних груп, відповідно до специфіки яких компанії необхідно буде здійснити дії для їх утримання та підвищення активності в порівнянні з вже наявною на поточний момент.

Так як мета дослідження включає в себе аналіз та моделювання як структурних особливостей поведінки споживачів, так і динаміку їх змін, в процесі роботи з даними доцільним є використання як статичних, так і динамічних методів та моделей машинного навчання (рис. 1.5).

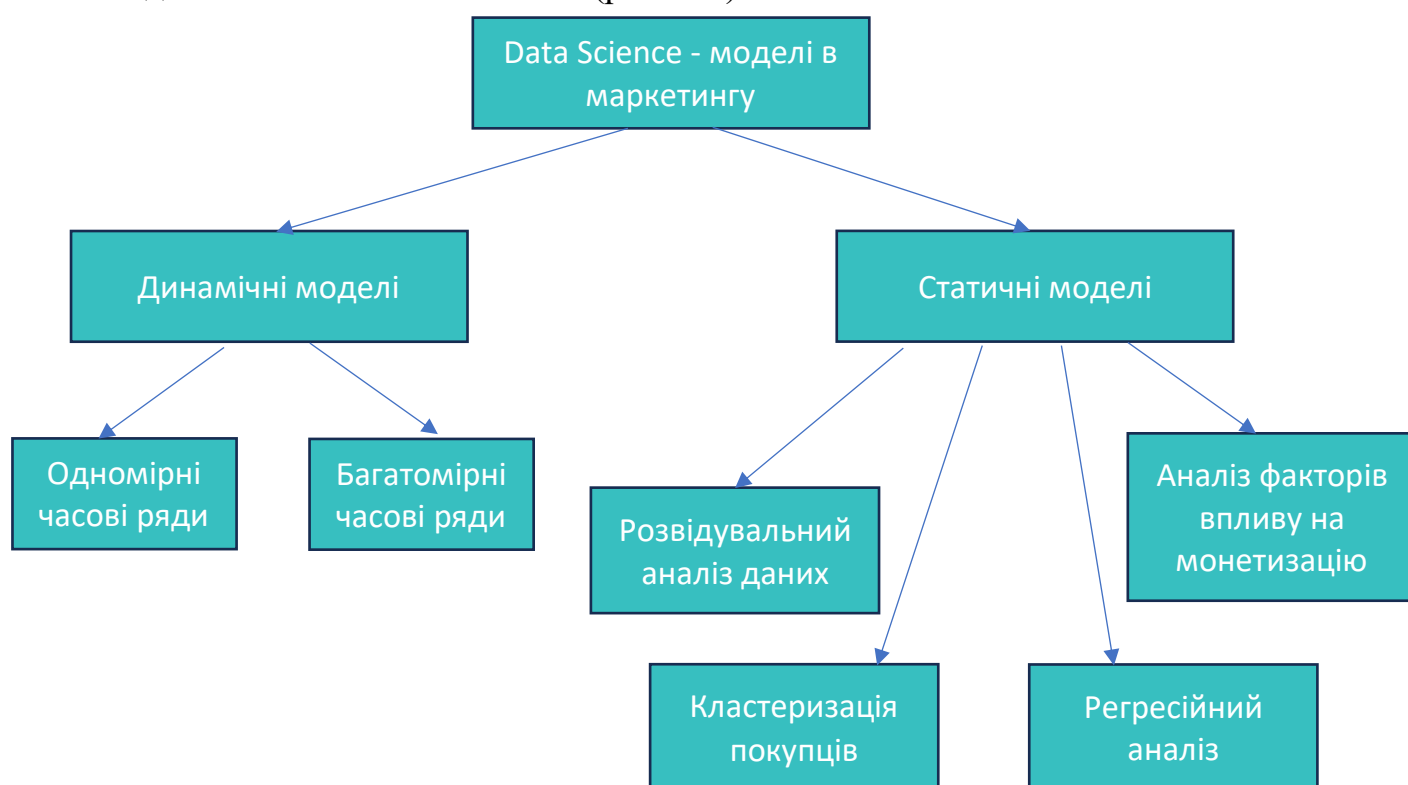


Рисунок 1.5 – Система моделей аналізу результатів маркетингових досліджень

Джерело: Розроблено автором на основі [7]

Метод визначення основних компонент є інструментом зменшення розмірності даних з великою кількістю змінних, які, крім того, можуть бути корельованими одне з одним. Даний метод є досить точним способом сформулювати нові ознаки серед існуючих факторів, або виділити головні з них, при цьому втративши мінімальну кількість інформації. Даний метод є точним інструментом визначення латентних факторів поведінки споживачів і може бути застосоване для структурного аналізу поведінки цільової аудиторії.

Метод К-середніх, що входить до блоку класифікації покупців, є інструментом кластерного аналізу, що дозволяє розділити цільову аудиторію бізнесу на окремі групи, подібні за деякими спільними характеристиками. Розділення споживачів за спільними патернами поведінки, які, в тому числі, можуть бути визначені методом головних компонент, є важливим для формування маркетингової стратегії підприємств та підвищення залученості клієнтів.

Методи регресійного аналізу, які, наряду з іншими методами, включають в себе рідж регресію, ласо регресію та еластичнет, досліджують динамічні дані споживацької активності та застосовуються для прогнозування поведінки цільової аудиторії. Дані методи відносяться до динамічних моделей, а саме блоку багатомірних часових рядів. Потужним інструментом в маркетинговій діяльності бізнесу може бути прогностичний аналіз основних RFM метрик, що в перспективі може результувати не тільки в удосконалення каналів комунікації з аудиторією, але і підготовку бізнесу до підвищення/спаду споживацької активності та здійснення заходів щодо її виведення на бажаний рівень.

РОЗДІЛ 2

МАТЕМАТИЧНІ МЕТОДИ ТА ІНСТРУМЕНТАРІЙ МОДЕЛЮВАННЯ СПОЖИВАЦЬКОЇ ПОВЕДІНКИ

2.1. Аналіз тенденцій поведінки цільової аудиторії засобами Google Analytics 4

Оскільки досліджуваний в Розділі 3 датасет відображає споживацької активності та характері здійснення покупок, включно з WEB-каналами, доцільним є розгляд поточного стану ринку засобами Web-аналітики. Розглянемо декілька графіків та діаграм, що відображатимуть актуальний стан споживацької активності за допомогою інструментів Google analytics та даних демо аккаунту [8], за допомогою якого можна отримати доступ до даних відвідуваності, монетизації та інших показників активності споживачів веб-сторінки магазину Google – Google Merchandise Store.

Розглянемо, до прикладу, статистику відвідуваності веб-сторінки магазин за останній місяць (рис. 2.1-2.2).

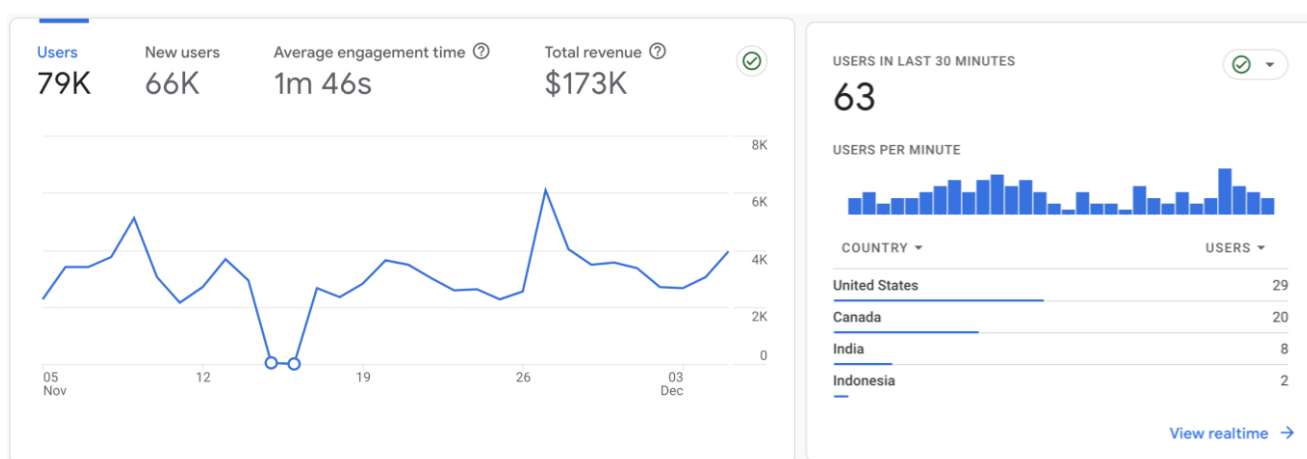


Рисунок 2.1 – Статистика відвідуваності Google Merchandise Store [8]

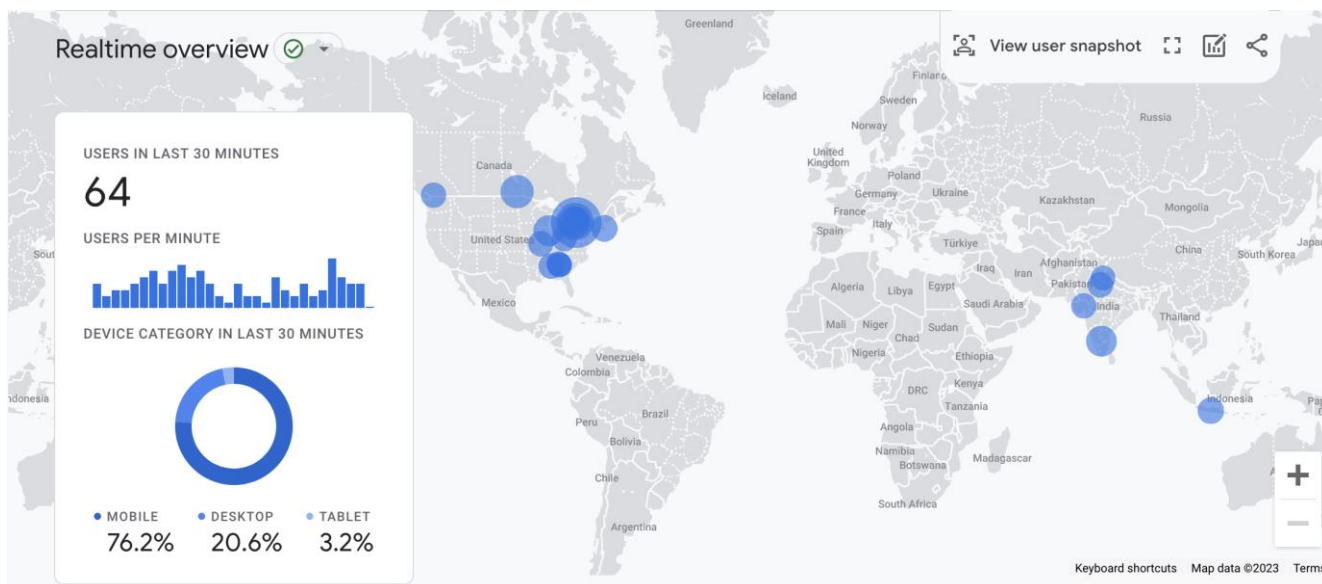


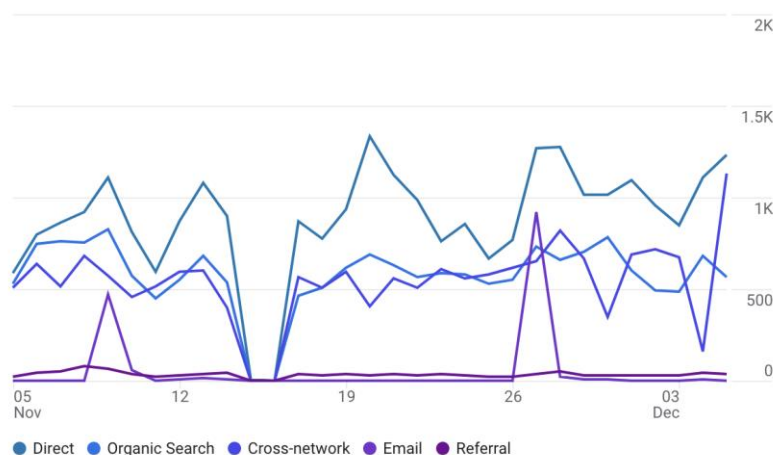
Рисунок 2.2 – Територіальний розподіл відвідувачів Google Merchandise Store [8]

Можемо спостерігати, що загальний показник відвідуваності (нові та вже знайомі користувачі) не є постійним, і коливається в межах 2 – 6 тис. На графіку можемо також спостерігати два аномальних спостереження зі значеннями 56 та 17 відвідувачів веб-сторінки 15.11.2023 та 16.11.2023 відповідно. Можемо зробити припущення про недоступність сервісу в деяких країнах протягом зазначених дат.

Крім того, аналізуючи територіальний графік розподілу користувачів сайту в режимі реального часу, можемо помітити, що найбільша кількість з них демонструє активність на території США, Канади та Індії.

На основі наведених раніше графіків можемо зробити припущення, що величина користувачів сайту не є стабільною величиною, і може коливатися в залежності від багатьох факторів, включно з днем збору інформації (робочий/вихідний день, чи проводиться свято в цей день, чи для всіх користувачів доступний сервіс на даний момент), країни, де зареєстрована компанія-власник магазину, специфіка товарів та відповідні рівні розвитку даних секторів економіки в інших країнах тощо.

New users by First user default channel group over time



New users by First user default channel group

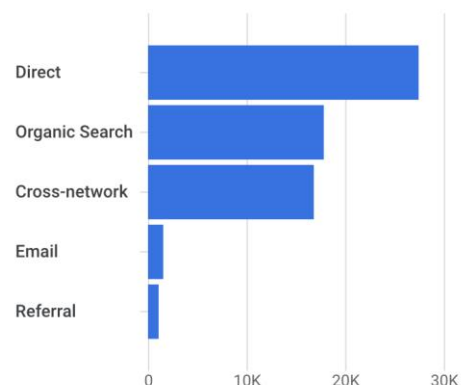


Рисунок 2.3 – Канали приходу нових користувачів Google Merchandise Store [8]

Розглядаючи більш специфічно канали приходу користувачів на сайт (рис. 2.3), можемо помітити, що найбільша кількість з них прийшла на сайт через канал «Direct», що означає прямий перехід по посиланню, тобто основною метою користувача протягом даного сеансу є саме виконання необхідних дій на сторінці компанії. Другим за абсолютними показниками є канал «Organic search», що позначає перехід на сторінку компанії через пошук в Google схожих товарів, що виставлені на досліджуваній сторінці. Загалом, можемо спостерігати, що кількість відвідувачів магазину, що надійшла з більшості доступних каналів є відносно стабільною на не часто змінюється за абсолютними показниками відвідуваності, окрім декількох піків на графіці каналу «Email», який позначає перехід за посиланням через рекламу, надіслану на пошту. На даному графіку, як і на попередніх, можемо бачити значний спад активності в період 15.11.2023 - 16.11.2023 (рис 2.3).

Звернемо більше уваги на показники монетизації та відповідні графіки (рис. 2.4).

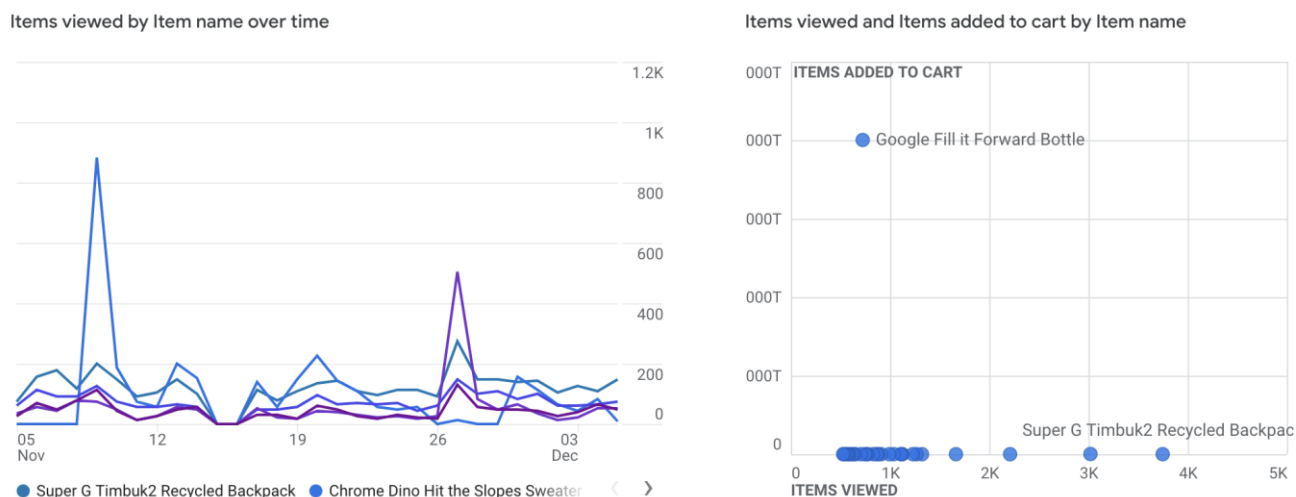


Рисунок 2.4 – Динаміка показника монетизації відповідних товарів Google Merchandise Store та їх розподіл за переглядами та додаванням в кошик [8]

На попередніх двох візуалізаціях можемо бачити динаміку переглянутих товарів покупцями та співвідношення між переглядами товарів та кількістю їх додавання в кошик. За першим графіком можемо зробити припущення, що найчастіше товари мають постійний попит з боку відвідувачів магазину, що підтверджується переважним утриманням кривих, що позначають відповідний товар, в рамках деяких величин. Однак, існують піки активності, що, за нашим припущенням, може бути пов'язаним з проведенням тієї чи іншої рекламної кампанії. Наприклад, такий товар як «Chrome Dino Hit the Slopes Sweater» мав показник додавання в кошик близький до нуля, але 09.11.2023 різко зріс до позначки в 900 тис., що перевищило всі інші товари. Крім того, бачимо ще один пік в додаваннях товарів в кошики 27.11.2023, але на цей раз вже декількох товарів одразу. Зробимо припущення, що поява даних локальних максимумів, які можемо спостерігати на графіку, була спричинена сезонними маркетинговими кампаніями «Black Friday». Тому, одною з перспектив для подальшого дослідження можемо визначити виявлення залежностей між даними, до представлені в досліджуваному данасеті про результативність проведених маркетингових кампаній з їх співпадінням або не співпадінням з глобальними кампаніями, такими як «Black Friday».

З другого графіку на рис. 2.4 можемо зробити наступне спостереження – відношення переглянутих товарів до товарів, доданих в кошик. Для більшості спостережень існує тенденція до низьких показників додавання в кошик в порівнянні з абсолютними величинами переглядів. Хоч товари і за віссю рівня переглядів мають достатньо великий розкид, однак всі наближаються до мінімальних позначок додаваності в кошик. Однак, серед інших товарів можемо спостерігати аномально високе співвідношення додаваності в кошик в порівнянні з переглядами, що може вказувати на те, що також була проведена маркетингова кампанія по залученню покупців. Переважання додавання в кошик над кількістю переглядів також може говорити про те, що на одне додавання в кошик приходилось декілька товарів одразу, тобто покупки здійснювались споживачами переважно в декілька одиниць даного товару за раз.

На основі графіків рис. 2.4 можемо зробити припущення, що кількість переглянутих товарів, та товарів, доданих в кошик, є переважно сталими величинами, однак за проведення рекламних кампаній та додаткового заохочення користувачів до купівлі можна отримати відповідні показники популярності товару в 4-5 разі вищі за звичайні.

Розглянемо також один з найбільш популярних маркетингових показників, а саме воронку продажів на рис. 2.5.

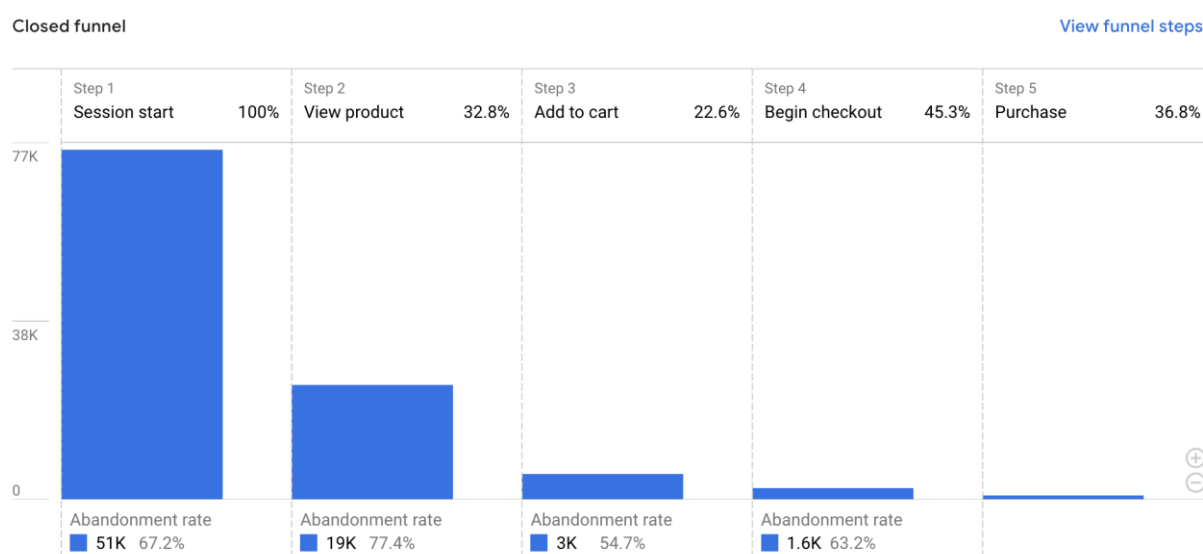


Рисунок 2.5 – Воронка продажів Google Merchandise Store [8]

Зауважимо, що кожна воронка продажів є специфічною для конкретного товару, однак всі вони мають спільні ключові етапи, що можуть бути трансформованими за рівнями деталізації/абстракції, однак логічно будуть присутніми в кожній воронці продажів (рис.2.6).



Рисунок 2.6 – Класична воронка продажів за Д. Бутенко [9]

Початковим етапом споживацької активності є початок сеансу, тобто відвідування веб-сторінки магазину. Споживач ознайомлюється із загальним каталогом товарів, оцінює візуальну складову сайту, і на даному етапі вже може стати об'єктом розповсюдження маркетингових кампаній. Даний показник береться як 100% для подальших обчислень. Після цього споживач може продовжити шлях до купівлі товару, перейшовши на сторінку конкретного товару. За гістограмою, показник переходів на сторінки товарів складає 33.4% від початкового числа відвідувачів. Споживацька активність може мати продовження в додаванні товару в кошик, що, за статистикою Google становить 22.8% від попереднього показника переглядів сторінок конкретного товару. Після додавання товару в кошик, при успішному ході подій, споживач може приступити до оформлення замовлення, що зазвичай стається в 44.7% випадків. Звернемо увагу, що це не еквівалентне оформленню замовлення, а позначає лише перехід на сторінку оформлення та введення необхідних даних. І кінцевий етап вдалої воронки

продажів для даної компанії є покупка, тобто повне внесення даних до замовлення та оплата, що стається за статистикою у 37.2% випадків.

Можемо зауважити, що конкретно для цього представника бізнесу – Google Merchandise Store – найменший відсоток «вдалої» споживацької активності, тобто слідування воронці продажів, спостерігається на етапі переходу від перегляду товарів до додавання товарів в кошик. Можемо зробити припущення, що необхідно підсилити інформативність та привабливість сторінок конкретних товарів для покупця з метою підвищення бажання останнього додати товар в кошик після його перегляду.

Найбільший показник слідування воронці продажів споживачами спостерігається на етапі переходу від додавання товару в кошик до початку оформлення замовлення. Дане спостереження може говорити про характерні патерни поведінки споживачів, а саме їх визначеність до купівлі товару після того, як був зроблений вибір додати його до кошика.

Розглянемо графік на рис. 2.7 в контексті впливу маркетингових кампаній на рівень споживацької активності.

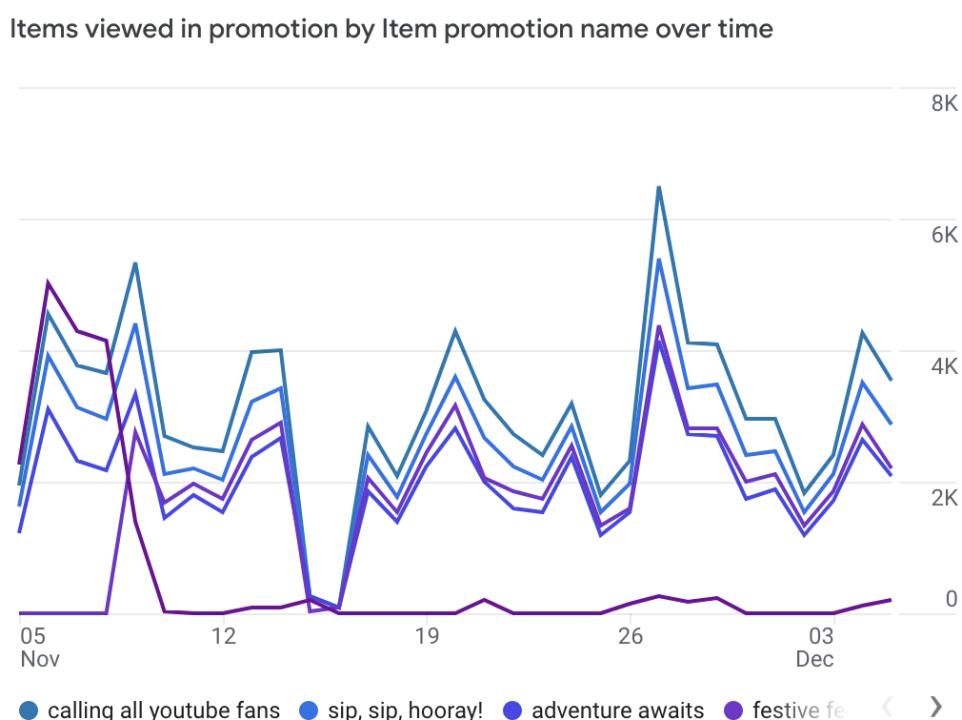


Рисунок 2.7 – Динаміка переглядів товарів Google Merchandise Store [8]

На візуалізації можемо бачити динаміку впливу конкретної промоакції на кількість переглянутих споживачами товарів (2 етап описаної вище воронки продажів), тобто у скільки переглядів товарів споживачами результувала конкретна промоакція. Цікавим є те, що показники переглядів за конкретними промоакціями знаходяться на постійних позиціях відносно одне одного. Динаміка змін абсолютної величини переглядів є змінюваною величиною, однак частка кожної з рекламних кампаній є постійною, тобто значення переглядів для них зростають або спадають синхронно. Виключенням є промоакція «aloha», що тримала провідну позицію серед інших на початок досліджуваного періоду, однак близько 10.11 результативність кампанії стрімко спала. Можемо зробити висновок про закриття даної кампанії, або зменшення об'ємів направлених в неї коштів.

Можемо зробити припущення, що вплив конкретної маркетингової кампанії є відносно сталою величиною, тобто при запуску даної кампанії вона досягає певного рівня результативності та зберігає дану позицію серед інших кампаній протягом переважного періоду свого існування.

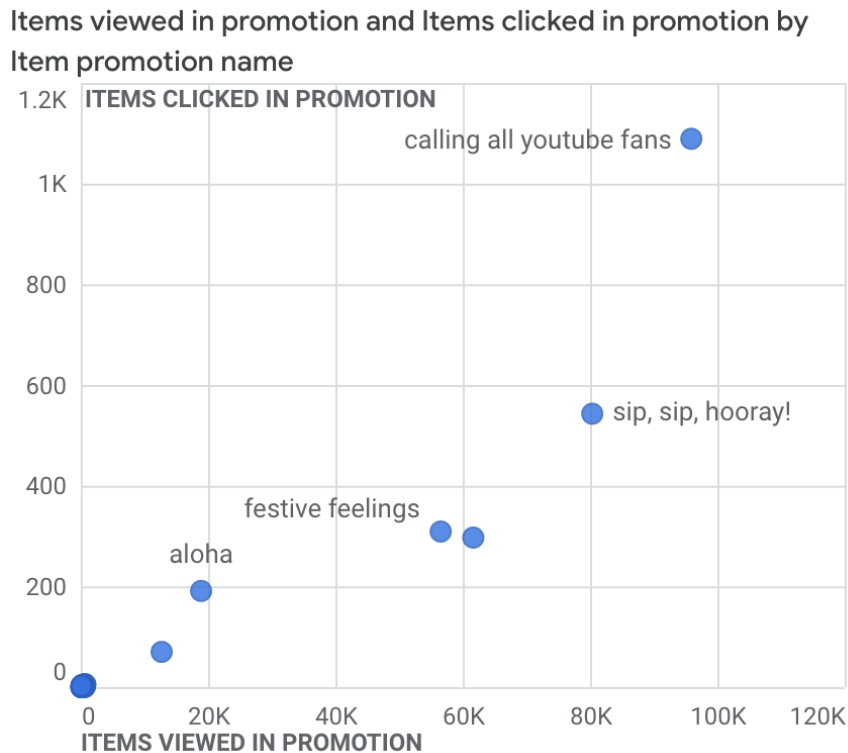


Рисунок 2.8 – Розподіл маркетингових кампаній Google Merchandise Store за показниками переглядів та додавань в кошик [8]

На графіку залежності переглянутих товарів та обраних (в кошик) товарів (рис. 2.8) можемо спостерігати залежність пряму залежність між показниками. Однак, на графіку також помітно, що деякі компанії мають низький рівень продовження споживачами воронки продажів, інші – навпаки високий. Наприклад, кампанія «adventure awaits», що на малюнку знаходиться поруч з «festive feelings», має значне переважання переглянутих товарів до товарів, доданих в кошик, в порівнянні з іншими промоакціями. Це може говорити про не такий вдалий результат, якого можна було очікувати від кампанії. В свою чергу, промоакція «calling all youtube fans» показала достатньо високі результати, і у відносному вимірі – співвідношення переглянутих товарів до доданих, і у абсолютному вимірі – обидва показники сягають максимальних значень порівняно з іншими промоакціями.

Враховуючи описаний вплив рекламних кампаній на показники воронки продажів, а саме етапи переглядів товарів та їх додавання до кошика, можемо зробити висновок, що доцільним буде розгляд подібних показників «повертаємості» споживачів для декількох рекламних кампаній в досліджуваному в Розділі 3 датасеті.

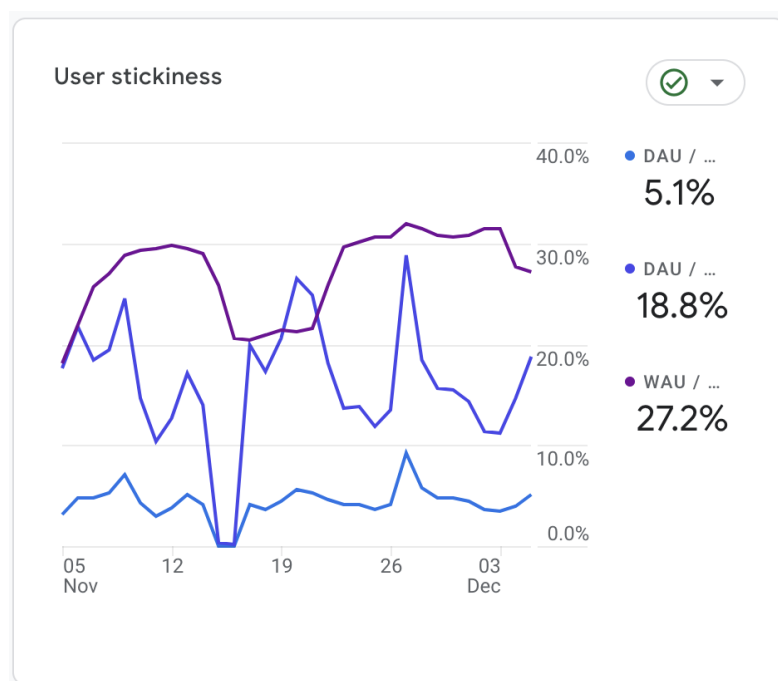


Рисунок 2.9 – Показники лояльності покупців Google Merchandise Store [8]

За допомогою Google Analytics також можемо здійснити деяку сегментацію споживачів за ознакою лояльності (рис. 2.9). Викликає цікавість вибір груп класифікації, адже вони були сформульовані не просто в формі абсолютних показників відвідуваності за день/тиждень/місяць, а у формі їх відношень одне до одного. Це дозволяє оперувати не тільки поточними абсолютними даними, але і аналізувати специфіку/характер поведінки споживачів в різному часовому контексті.

Тож, на поточний момент відношення кількості клієнтів, що відвідали магазин за тиждень, до кількості клієнтів, що відвідали магазин за місяць, сягає 30%, що горить про за останній тиждень активність споживачів не значним чином підвищилась. Крім того, це може казати про постійність даної величини, а отже про лояльність покупців магазину, адже величина в 30% не значним чином відрізняється від 25%, при якому відвідуваність за місяць була б розподілена рівномірно. Наступним розглянемо співвідношення відвідуваності за день до кількості відвідувань за тиждень, що за актуальною статистикою становить 7%. Так як при рівномірному розподілі частка тих клієнтів, що відвідують магазин за день по відношенню до тижня мала складати 14%, можемо зробити припущення, що за останні дні лояльність споживачів дещо знизилась. Однак, звернемо увагу, що ця інформація є оперативною, і говорить лише про споживацьку активність в поточному моменті, а не загалом в довгостроковій перспективі. Аналізуючи показник відношення відвідувань за день до відвідувань за місяць, можемо так само зробити висновок про оперативне зниження активності споживачів за останні дні, адже статистичний показник приймає значення 2%, тоді як при рівномірному розподілі мав би знаходитись біля позначки 3.3%.

Отже, на отриманих даних можемо зробити висновок, що лояльність споживачів дещо знизилась протягом останнього короткострокового періоду, однак більшість покупців можна класифікувати як лояльних.

2.2. Моделі сегментування цільової аудиторії та визначення латентних факторів поведінки споживачів

Метод головних компонент є одним з алгоритмів зменшення розмірності даних та полягає у «...виборі нової ортогональної системи координат у просторі спостережень» [10]. За першу головну компоненту береться напрямок, вздовж якого сукупність спостережень має найбільший показник дисперсії. Наступні головні компоненти обираються за схожим алгоритмом – відповідно до максимального значення дисперсії вздовж цього напрямку. Відповідно, при зростанні номера компоненти буде зменшуватись пов'язана з нею частка загальної дисперсії [10]. Оптимальна кількість головних компонент визначається суб'єктом моделювання.

Графічно метод головних компонент може бути представлений наступним чином. Уявімо, що маємо дві досліджувані змінні – довжину та ширину, які до того ж корелюють одне з одною. Проведемо вектор вздовж найбільшого розкиду дисперсії, а потім ще один перпендикулярно до першого (рис. 2.10).

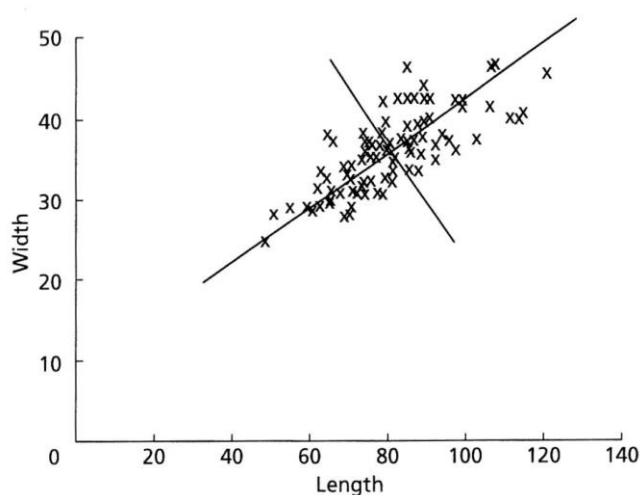


Рисунок 2.10 – Розподіл спостережень за значеннями показників «Довжина» - «Ширина» [11]

Після визначення цих векторів маємо можливість зробити перерахунок точок спостережень в новій системі координат та зобразити їх наступним чином:

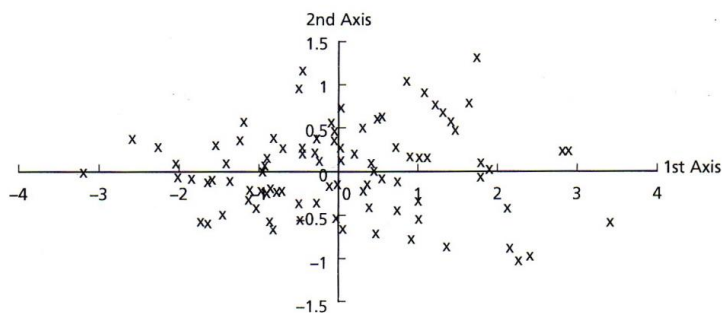


Рисунок 2.11 – Розподіл спостережень за значеннями головних компонент [11]

В новій системі координат (рис. 2.11) більша дисперсія зберігається вздовж першої вісі. Положення спостережень відносно одне одного не були змінені, змінилися лише їх координати. Крім того, новоутворені параметри не є корельованими між собою, на відміну від оригінальних. Нові параметри можуть нести інше логічне навантаження, ніж початкові, і відповідно доречним може бути присвоєння їм інших назв.

При роботі з датасетами з великою кількістю змінних дисперсії вздовж деяких змінних можуть бути відносно невеликими в порівнянні з іншими, а тому можуть бути проігноровані в користь напрямків більшої дисперсії.

Алгоритм К-середніх є одним з найпоширеніших в застосуванні до задач кластеризації. Алгоритм класифікує елементи в k груп або кластерів подібності. Як показник подібності використовується Евклідова відстань [12].

Алгоритм роботи алгоритму наступний:

- ініціалізація випадковим чином k точок, які визначаємо як кластерні центроїди;
- класифікація спостережень відповідно до його найближчого середнього значення (центроїду) та перерахунок координат центроїду відповідно як середнє значення координат спостережень, які на цьому етапі були віднесені до відповідного кластеру;
- повторення процесу протягом заданої кількості ітерацій, або до моменту, коли при перерахунку координат центроїдів, вони не змінять своє значення в порівнянні з попередньою ітерацією [12].

Початкова ініціалізація центрів кластерів також може проводитись декількома способами, серед яких присвоєння центроїдам значення координат випадкових спостережень з датасету, або випадкових величин в межах розкиду спостережень, однак не рівні їм.

Загальний перелік можливих способів ініціалізації середніх точок є наступним [13]:

- вибір найближчих до початку координат;
- вибір найближчих до центрів зосередження множини спостережень в просторі ознак;
- вибір в найвіддаленіших від основного зосередження спостережень;
- вибір найвіддаленіших від початку координат;
- точки рівномірно віддалені від початку координат на задану відстань;
- виріб, заставаний на логіці, відповідній до специфіки конкретного набору даних;
- вибір випадковим чином.

Випадковий вибір значень центроїдів у деяких випадках може бути невдалим, тому на практиці часто виконується запуск k-середніх кілька разів з різними початковими значеннями центроїдів [14].

Для оцінки точності вибору кількості кластерів використовуватимемо в подальшому дослідженні такі метрики як коефіцієнт силуету, індекс Девіса-Болдіна та індекс Калінські-Харабаша.

Коефіцієнт силуету — один з показників, які використовуються для обчислення ефективності методу кластеризації. Його значення коливається від -1 до 1 [13]:

- 1: означає, що кластери знаходяться на великій відстані один від одного та мають чітку умовну межу одне між одним.
- 0: означає, що кластери незначно пов'язані одне з одним, або відстань між кластерами незначна.
- -1: означає, що спостереження віднесені до відповідних кластерів невдало.

Даний показник базується на «згуртованості» спостережень в межах конкретного кластеру (внутрішньо-кластерна відстань) і різниці у спостереженнях між різними кластерами (між-кластерна відстань), щоб в подальшому здійснити обчислення на основі цих двох величин.

В загальному формулюванні, індекс обчислюється шляхом знаходження відстані від спостереження до центроїда кластеру, до якого воно належить, і подальшого порівняння з відстанню від спостереження до найближчого центроїда кластеру, до якого спостереження не належить. За вдалого розбиття на кластери друга відстань має переважати над першою. Подібний алгоритм застосовується до кожного спостереження і визначається загальний показник для реалізованої кластеризації.

Під час розрахунку «силуетного» індексу для кожного кластера C_j ($j = 1, 2, \dots, k$) і кожного спостереження кластера X належить S обчислюється величина s_i - с показник силуету (формула 2.1).

$$s(i) = \frac{B(i) - A(i)}{\max\{A(i), B(i)\}} \quad (2.1)$$

де A_i – середня відстань між спостереженням X належить C_j і усіма спостереженнями кластера C_j ;

B_i – мінімальна відстань між спостереженням X_i належить C_j і спостереженнями X_l належить C_l [15].

Показник силуету для кластера C_j обчислюється за формулою 2.2.

$$S_j = \frac{1}{m} \sum_{i=1}^m s(i) \quad , \quad (2.2)$$

де m – кількість спостережень у кластері C_j [15].

Загальний силуетний індекс точності розбиття на кластери обчислюється за формулою 2.3.

$$GS = \frac{1}{k} \sum_{j=1}^k S_j \quad , \quad (2.3)$$

де k – обрана кількість кластерів [15].

Найкращим розбиттям на кластери вважається таке, при якому індекс силуету наближається на 1.

Іншим показником оцінки якості кластеризації є індекс Девіса-Болдіна, який вираховується як середнє значення від суми максимальних відношень між спостереженнями до відповідних між центроїдів та відстанями між центроїдами (формула 2.4).

$$DBI = \frac{1}{K} \sum_{k=1}^K \max_{j \neq k} \left[\frac{\sigma_k + \sigma_j}{d(c_k, c_j)} \right] \quad (2.4)$$

де σ_k та σ_j – міжкластерна відстань кластерів k та j відповідно;

d – міжкластерна відстань між k та j кластерами [16].

Беручи до уваги спосіб визначення коефіцієнту, як функції розкиду спостережень всередині кластеру до відстані між кластерами, більш точною вважатиметься кластеризація із меншим значенням індексу.

Даний індекс також можна визначити наступним чином. Охарактеризуємо відносний розкид спостережень у двох кластерах як половину суми середніх відстань їх елементів до центрів, розділену на відстань між центроїдами. Додатково охарактеризуємо розкид спостережень в кластері його максимальною величиною відносно розкиду в інших кластерах. Отже, індекс можна визначити як середній розкид спостережень за всіма кластерами [17].

Іншою метрикою для оцінки точності кластеризації є індекс Калінські-Харабаша, також відомий як критерій співвідношення дисперсії. Даний показник характеризує те, як сильно дане спостереження є схожим на спостереження з відповідного йому кластеру в порівнянні зі спостереженнями з інших кластерів [18]. Схожість елементів всередині кластеру оцінюється на основі відстані від координат спостережень у кластері до його центроїда, а розділення спостережень спирається на відстань між центроїдами кластерів до глобального центроїда (формула 2.5).

$$CH = \left[\frac{\sum_{k=1}^K n_k \|c_k - c\|^2}{K - 1} \right] / \left[\frac{\sum_{k=1}^K \sum_{i=1}^{n_k} \|d_i - c_k\|^2}{N - K} \right], \quad (2.5)$$

де n_k – кількість точок k -того кластеру;

c_k координати центроїду k -того кластера відповідно;

c – глобальний центроїд;

N – загальна кількість множини спостережень [18].

Чим вищим є величина індексу, тим більш впевненими можна бути, що кластери щільні та добре розділені. Серед переваг цього методу можна виділити інтуїтивну інтерпретацію, адже вище значення індексу відповідає більш щільному розподілу спостережень на кластери.

2.3. Моделювання поведінки споживачів на ринку методами машинного навчання

Досліджуваний нами датасет [7] є відображенням споживацької активності покупців торгової мережі, зібрані за період в три роки. Маючи велику кількість змінних в датасеті, що відображають специфіку споживацької поведінки та особистісні характеристики цільової аудиторії підприємства, для задачі регресії були вибрані такі моделі як рідж регресія, ласо регресі та еластікнет.

Вибір цих моделей був спричинений тим, що деякі змінні в досліджуваному датасеті є корельованими між собою. Нижче наведені змінні датасету та відповідні коефіцієнти кореляції між ними:

- Income та MntMeatProducts - 0.73550
- Income та MntWines - 0.70949
- NumCatalogPurchases та MntMeatProducts - 0.75001

Крім того, зазначені вище моделі мають часте використання в регресійних задачах зниження розмірності, що також є доречним у випадку досліджуваного датасету.

Рідж (гребнева регресія) є методом оцінки коефіцієнтів моделі лінійної регресії, у випадку, коли незалежні змінні корельовані між собою (формула 2.6).

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}, \quad (2.6)$$

де λ - коефіцієнт штрафу;

β_i – оцінки параметрів моделі [19].

Рідж-регресія додає квадрат коефіцієнта як штрафний термін до функції втрат [20]. В такому випадку, ми, замість того, щоб класично мінімізувати суму квадратів залишків, додатково обтяжуємо параметри моделі β 's штрафним термом α . Цей терм є попередньо визначеною константою, помноженої на квадрат суми вектора коефіцієнтів β .

Крім того, рідж регресія є інструментом виконання L2 регуляризації – вона додає коефіцієнт суми квадратів оцінок параметрів – α (або λ) у цільову функцію, що мінімізується – функцію залишків. α може приймати різні значення [21]:

- при $\alpha = 0$: модель перетворюється на класичну лінійну регресію і матиме відповідні коефіцієнти без участі регуляризації;
- при $\alpha = \infty$: коефіцієнти моделі будуть «оштрафовані» на велике значення і наблизатимуться до 0;
- при $0 < \alpha < \infty$: величина α визначатиме допустиме навантаження на параметри моделі.

Крім того, необхідно зауважити, що на вхід моделі необхідно подавати вже нормалізовані дані, що і буде виконано нижче в Розділі 3.

Ласо регресія, також відома як L1 регуляризація, є подібною до рідж регресії, за виключенням деяких особливостей (формула 2.7).

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}, \quad (2.7)$$

де λ - коефіцієнт штрафу;

β_i – оцінки параметрів моделі [22].

На відміну від рідж регресії, штраф в регресії ласо накладається не на суму квадратів параметрів моделі, а на суму їх абсолютних величин [22]. Як результат, маємо можливість прирівнювання деяких коефіцієнтів моделі в точності до нуля.

Гіперпараметр α виконує аналогічну функцію, як і в рідж регресії – є регулятивною константою між мінімізацією суми квадратів залишків та абсолютними величинами коефіцієнтів. Розподіл можливих значень α та її вплив на параметри моделі визначається аналогічно до рідж регресії:

- $\alpha = 0$: модель є класичною лінійною регресією;
- $\alpha = \infty$: коефіцієнти прирівнюються до нуля;
- $0 < \alpha < \infty$: параметри моделі визначаються залежно від величини α .

Іншою моделлю, яка є компромісом між рідж регресією та ласо регресією, виступає модель еластікнет, штрафний вираз якою відображений у формулі 2.8.

$$\sum_{j=1}^p (\alpha |\beta_j| + (1 - \alpha) \beta_j^2) \quad , \quad (2.8)$$

де α - коефіцієнт відношення моделей рідж регресії та ласо регресії;

β_i – оцінки параметрів моделі [22].

Еластікнет є лінійною регресійною моделлю, яка включає в себе два штрафних фактори – з рідж та ласо регресій, що відповідно накладаються на функцію втрат. Обидва ці штрафні фактори збалансовуються гіперпараметром α , що може приймати значення від 0 до 1:

- $\alpha = 0$: нейтралізуємо штраф моделі ласо та маємо рідж регресію;
- $0 < \alpha < 1$: збалансовуємо вагу рідж-ласо регресій для нашої моделі;
- $\alpha = 1$: нейтралізуємо штраф моделі рідж та маємо ласо регресію.

Використання моделі еластікнет має наступні переваги [23]:

- вибір ознак: регресія еластікнет може виконувати вибір ознак, зменшуючи параметри непідходящих змінних до нуля. Це створює модель з меншою кількістю змінних, яка є легшою для розуміння та інтерпретації, а також менш схильною до перенавчання;

- робастість: еластікнет є більш «стабільною» моделлю, в порівнянні з рідж та ласо регресіями окремо, адже вона поєднує в собі сильні сторони обох методів. Еластікнет може обробляти корельовані змінні та змінні різних масштабів;
- висока застосованість: модель показує високі результати в порівнянні з іншими видами лінійних регресій, особливо у випадку великої кількості змінних.

РОЗДІЛ 3

ПРАКТИЧНЕ ЗАСТОСУВАННЯ МАТЕМАТИЧНИХ МОДЕЛЕЙ ПОВЕДІНКИ ЦІЛЬОВОЇ АУДИТОРІЇ НА РИНКУ УКРАЇНИ

3.1. Аналіз результатів розвідувального аналізу та препроцесингу даних про активність цільової аудиторії на ринку

Взятий для дослідження датасет [7] містить детальні дані про покупців магазину, що в перспективі може бути використаним для кращого розуміння бізнесом покупців та вдосконалення маркетингових кампаній залежно від специфіки поведінки, потреб та вподобань основних сегментів споживачів (рис. 3.1). Дані описують споживацьку активність покупців на американському ринку, репрезентуючи період економічної стабільності. Говорячи більш предметно про датасет, можемо бачити наступні характеристики покупців:

- ID: унікальний ідентифікатор
- Year_Birth: рік народження
- Education: рівень освіти
- Marital_Status: сімейний стан
- Income: річний рівень доходу
- Kidhome: кількість немовлят
- Teenhome: кількість дітей-підлітків
- Recency: кількість днів після останнього придбання товару
- Complain: має значення 1 якщо покупець надсилав скаргу за останні 2 роки, 0 – в протилежному випадку

Умовний блок «Продукти» містить інформацію про структуру витрат споживача за останні 2 роки (в \$):

- MntWines: витрачено на вино
- MntFruits: витрачено на фрукти
- MntMeatProducts: витрачено на м'ясо
- MntFishProducts: витрачено на рибу
- MntSweetProducts: витрачено на солодощі
- MntGoldProds: витрачено на дорогоцінні метали

Умовний блок «Реклама» характеризує реакції покупця на рекламні кампанії:

- NumDealsPurchases: кількість покупок, здійснених зі знижкою
- AcceptedCmp1: 1 якщо покупець здійснив покупку під час першої рекламної кампанії, 0 – в протилежному випадку
- AcceptedCmp2: 1 якщо покупець здійснив покупку під час другої рекламної кампанії, 0 – в протилежному випадку
- AcceptedCmp3: 1 якщо покупець здійснив покупку під час третьої рекламної кампанії, 0 – в протилежному випадку
- AcceptedCmp4: 1 якщо покупець здійснив покупку під час четвертої рекламної кампанії, 0 – в протилежному випадку
- AcceptedCmp5: 1 якщо покупець здійснив покупку під час п'ятої рекламної кампанії, 0 – в протилежному випадку
- Response: 1 якщо покупець здійснив покупку під час останньої рекламної кампанії, 0 – в протилежному випадку

Умовний блок «Реклама» містить кількісну інформацію про канал здійснення покупки споживачем:

- NumWebPurchases: кількість покупок, здійснених через сайт компанії
- NumCatalogPurchases: кількість покупок, здійснених через сайт каталог
- NumStorePurchases: кількість покупок, здійснених у точці торгової мережі

- NumWebVisitsMonth: кількість відвідувань веб-сайту компанії за останній місяць.

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts
0	5524	1957	Graduation	Single	58138.00000	0	0	58	635	88	546	172	88
1	2174	1954	Graduation	Single	46344.00000	1	1	38	11	1	6	2	1
2	4141	1965	Graduation	Together	71613.00000	0	0	26	426	49	127	111	21
3	6182	1984	Graduation	Together	26646.00000	1	0	26	11	4	20	10	3
4	5324	1981	PhD	Married	58293.00000	1	0	94	173	43	118	46	27
5	7446	1967	Master	Together	62513.00000	0	1	16	520	42	98	0	42
6	965	1971	Graduation	Divorced	55635.00000	0	1	34	235	65	164	50	49
7	6177	1985	PhD	Married	33454.00000	1	0	32	76	10	56	3	1
8	4855	1974	PhD	Together	30351.00000	1	0	19	14	0	24	3	3
9	5899	1950	PhD	Together	5648.00000	1	1	68	28	0	6	1	1
10	1994	1983	Graduation	Married	NaN	1	0	11	5	5	6	0	2
11	387	1976	Basic	Married	7500.00000	0	0	59	6	16	11	11	1
12	2125	1959	Graduation	Divorced	63033.00000	0	0	82	194	61	480	225	112
13	8180	1952	Master	Divorced	59354.00000	1	1	53	233	2	53	3	5
14	2569	1987	Graduation	Married	17323.00000	0	0	38	3	14	17	6	1

Рисунок 3.1 – Фактори впливу на споживацьку активність

Джерело: Розроблено автором на основі [7]

MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5
88	3	8	10	4	7	0	0	0
6	2	1	1	2	5	0	0	0
42	1	8	2	10	4	0	0	0
5	2	2	0	4	6	0	0	0
15	5	5	3	6	5	0	0	0
14	2	6	4	10	6	0	0	0
27	4	7	3	7	6	0	0	0
23	2	4	0	4	8	0	0	0
2	1	3	0	2	9	0	0	0
13	1	1	0	0	20	1	0	0
1	1	1	0	2	7	0	0	0
16	1	2	0	3	8	0	0	0
30	1	3	4	8	2	0	0	0
14	3	6	1	5	6	0	0	0
5	1	1	0	3	8	0	0	0

Рисунок 3.1 – Фактори впливу на споживацьку активність (продовження)

Джерело: Розроблено автором на основі [7]

Отже, маємо 26 змінних та 2240 екземплярів в датасеті.

	count	mean	std	min	25%	50%	75%	max
ID	2240.00000	5592.15982	3246.66220	0.00000	2828.25000	5458.50000	8427.75000	11191.00000
Year_Birth	2240.00000	1968.80580	11.98407	1893.00000	1959.00000	1970.00000	1977.00000	1996.00000
Income	2216.00000	52247.25135	25173.07666	1730.00000	35303.00000	51381.50000	68522.00000	666666.00000
Kidhome	2240.00000	0.44420	0.53840	0.00000	0.00000	0.00000	1.00000	2.00000
Teenhome	2240.00000	0.50625	0.54454	0.00000	0.00000	0.00000	1.00000	2.00000
Recency	2240.00000	49.10938	28.96245	0.00000	24.00000	49.00000	74.00000	99.00000
MntWines	2240.00000	303.93571	336.59739	0.00000	23.75000	173.50000	504.25000	1493.00000
MntFruits	2240.00000	26.30223	39.77343	0.00000	1.00000	8.00000	33.00000	199.00000

Рисунок 3.2 – Описова статистика датасету

Джерело: Розроблено автором на основі [7]

На основі даних, наведених вище (рис. 3.2), можемо на даному етапі вже зробити деякі висновки про типового споживача: за віком наближається до 50 років, має приблизний дохід в 52 тис. дол. на рік, маленьких дітей скоріше не має, проте може мати дитину-підлітка. В середньому раз на півтора місяці купляє товар в магазині, і найбільше коштів витрачає на купівлю вина та м'яса, найменше – а купівлю фруктів. Найбільше покупок здійснює в торгових точках, і трохи менше – замовляє на сайті.

df.isnull().sum()	df.dtypes
ID 0	ID int64
Year_Birth 0	Year_Birth int64
Education 0	Education object
Marital_Status 0	Marital_Status object
Income 24	Income float64
Kidhome 0	Kidhome int64
Teenhome 0	Teenhome int64
Recency 0	Recency int64
MntWines 0	MntWines int64
MntFruits 0	MntFruits int64
MntMeatProducts 0	MntMeatProducts int64
MntFishProducts 0	MntFishProducts int64
MntSweetProducts 0	MntSweetProducts int64
MntGoldProds 0	MntGoldProds int64
NumDealsPurchases 0	NumDealsPurchases int64
NumWebPurchases 0	NumWebPurchases int64
NumCatalogPurchases 0	NumCatalogPurchases int64
NumStorePurchases 0	NumStorePurchases int64
NumWebVisitsMonth 0	NumWebVisitsMonth int64
AcceptedCmp3 0	AcceptedCmp3 int64
AcceptedCmp4 0	AcceptedCmp4 int64
AcceptedCmp5 0	AcceptedCmp5 int64
AcceptedCmp1 0	AcceptedCmp1 int64
AcceptedCmp2 0	AcceptedCmp2 int64
Complain 0	Complain int64
Response 0	Response int64
dtype: int64	dtype: object

Рисунок 3.3 – Типи даних змінних датасету

Джерело: Розроблено автором на основі [7]

Маємо більшість значень в форматі int64 / float64 (рис. 3.3), що позначають змінні числових типів, та дві змінні типу object, що позначають строковий тип даних. При подальшій роботі проведемо перетворення категоріальних змінних – «Education» та «Marital_Status» - для їх використання в аналізі.

Крім того, можемо помітити наявність порожніх значень у змінній «Income», тож проведемо більш ґрунтовну роботу по підготовці її до дослідження. З метою заповнення відсутніх значень зробимо припущення, що категоріальна змінна «Education» може бути одним з ключових факторів для пояснення змінної «Income». Тому, виконавши групування «Income» за «Education» та визначивши середнє значення, заповнимо порожні дані в «Income» відповідно до належності конкретного екземпляра до тієї чи іншої групи «Education» (рис. 3.4).

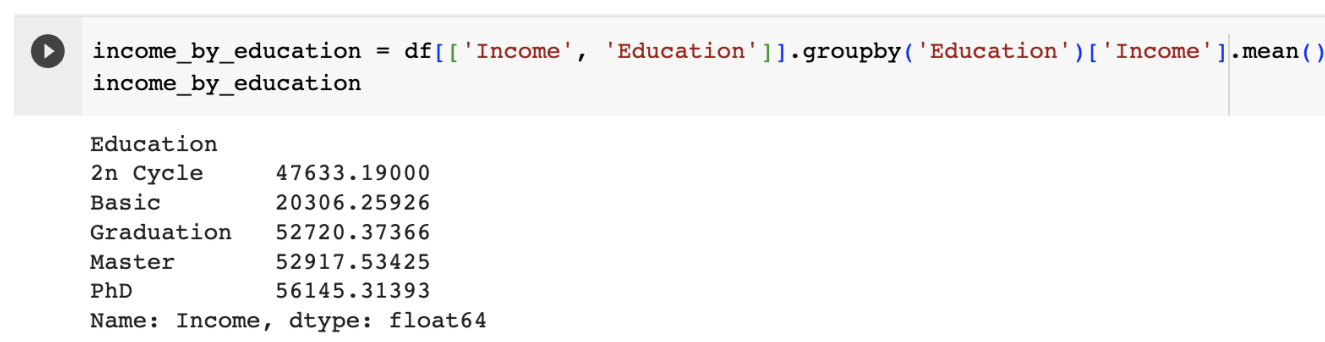


Рисунок 3.4 – Алгоритм заповнення пропущених значень датасету

Джерело: Розроблено автором на основі [7]

Проведемо перевірку та впевнимися, що не маємо більше порожніх значень в датасеті (рис. 3.5).

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 26 columns):
 #   Column                Non-Null Count  Dtype
---  -
 0   ID                    2240 non-null   int64
 1   Year_Birth            2240 non-null   int64
 2   Education             2240 non-null   object
 3   Marital_Status        2240 non-null   object
 4   Income                2240 non-null   float64
 5   Kidhome               2240 non-null   int64
 6   Teenhome              2240 non-null   int64
 7   Recency               2240 non-null   int64
 8   MntWines              2240 non-null   int64
 9   MntFruits             2240 non-null   int64
10   MntMeatProducts       2240 non-null   int64
11   MntFishProducts       2240 non-null   int64
12   MntSweetProducts      2240 non-null   int64
```

Рисунок 3.5 – Перевірка на відсутність пропущених значень

Джерело: Розроблено автором на основі [7]

Крім того, для зручності проведення розрахунків, створимо нову змінну «Age» замість існуючої «Year_Birth» (рис. 3.6).



```
df["Age"] = 2023 - df["Year_Birth"]
df = df.drop(["Year_Birth"], axis=1)
```

Рисунок 3.6 – Алгоритм створення нової змінної «Age»

Джерело: Розроблено автором на основі [7]

Використання в подальшому дослідженні змінної «Age» вважатимемо більш доцільним, адже частина року народження «19..» є надлишковою і збільшує значення віку на константу 1900 для всіх спостережень.

Наступним кроком проведемо роботу по усуненню викидів. Для цього створимо дві функції - outlier_thresholds() та replace_outliers() (рис. 3.7). Перша з них визначатиме показник найбільшого та найменшого значення для змінної, які вважатимуться нормальними для даного показника, та у разі їх перевищення значення буде визначене як викид. Межі визначатимемо за формулами [2]:

$$\text{lower inner fence: } Q1 - 1.5 * IQ \quad (3.1)$$

$$\text{upper inner fence: } Q3 + 1.5 * IQ, \quad (3.2)$$

де $Q1$ – перший квартиль;

$Q3$ – третій квартиль;

IQ – міжквартильний розмах.

Тож, за нижню межу матимемо значення першого квартилю, зменшеного на величину в півтора міжквартильного розмаху, за верхню межу - значення третього квартилю, збільшеного на величину в півтора міжквартильного розмаху, відповідно.

```

def outlier_thresholds(dataframe: pd.DataFrame, variable: str) -> tuple:
    quartile1 = dataframe[variable].quantile(0.25)
    quartile3 = dataframe[variable].quantile(0.75)
    interquartile_range = quartile3 - quartile1
    up_limit = quartile3 + 1.5 * interquartile_range
    low_limit = quartile1 - 1.5 * interquartile_range
    return (low_limit, up_limit)

def replace_outliers(dataframe: pd.DataFrame, variable: str) -> None:
    low_limit, up_limit = outlier_thresholds(dataframe, variable)
    df[variable] = np.where(df[variable] > up_limit, up_limit, df[variable])
    df[variable] = np.where(df[variable] < low_limit, low_limit, df[variable])

```

Рисунок 3.7 – Алгоритм визначення аномальних значень змінних датасету

Джерело: Розроблено автором на основі [7]

Друга функція `replace_outliers()` викликатиме попередню функцію `outlier_thresholds()` та, в разі виявлення, що дане спостереження є викидом за поточною змінною, змінюватиме його на найближчу межу – мінімальну чи максимальну, яка була визначена в `outlier_thresholds()`.

Для прикладу, проведемо дослідження на викиди для новоствореної змінної «Age» з використанням гістограми та боксплоту (рис. 3.8).

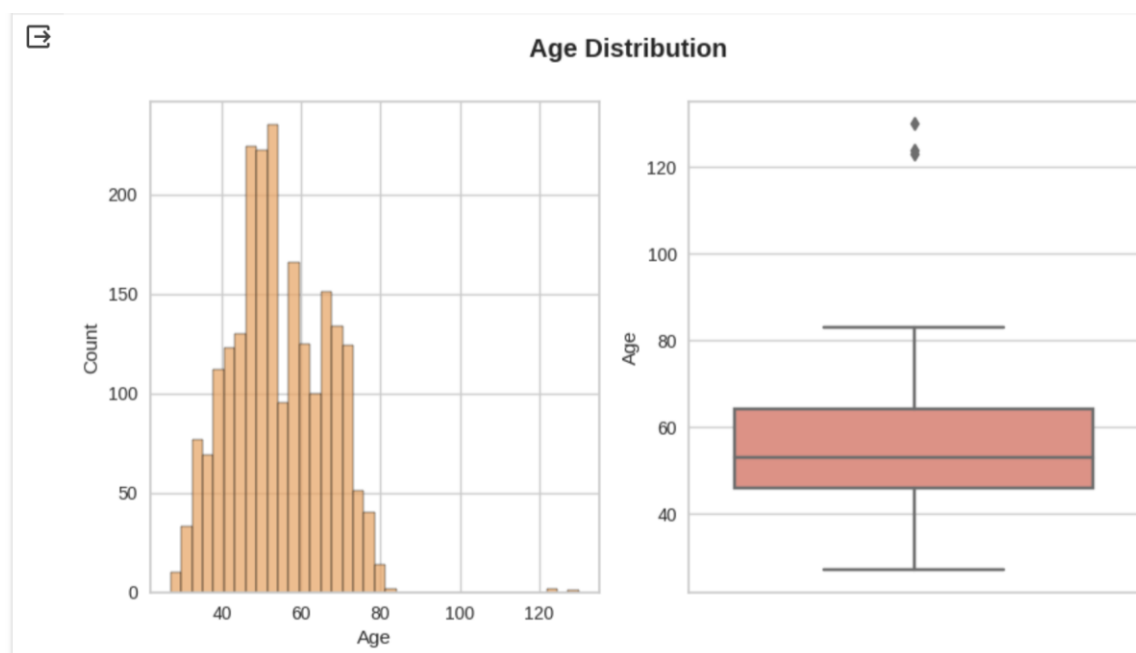


Рисунок 3.8 – Розподіл змінної «Age»

Джерело: Розроблено автором на основі [7]

Можемо спостерігати декілька викидів між третім та четвертим кuartилями. Логічно визначивши, що ймовірність того, що вік покупців приймає значення більші за 120, є мінімальною, застосуємо `replace_outliers()` до цієї змінної.

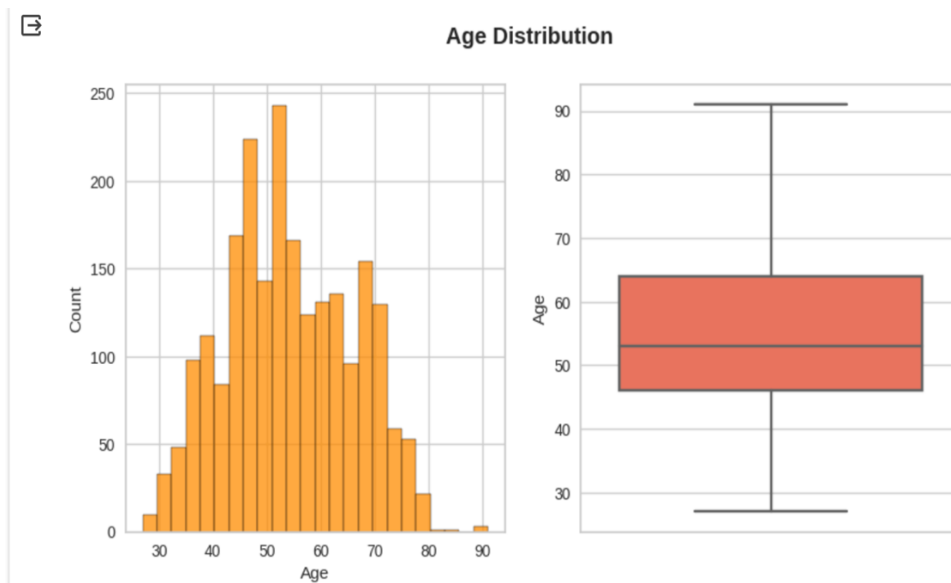


Рисунок 3.9 – Розподіл змінної «Age» після заміни аномальних значень

Джерело: Розроблено автором на основі [7]

Можемо впевнитися в коректній роботі функції `replace_outliers()` за рис. 3.9, а саме за відсутністю викидів на графіку для змінної «Age». Проведемо подібну обробку змінної «Income». На початку маємо такі дані і відповідні візуалізації (рис. 3.10).

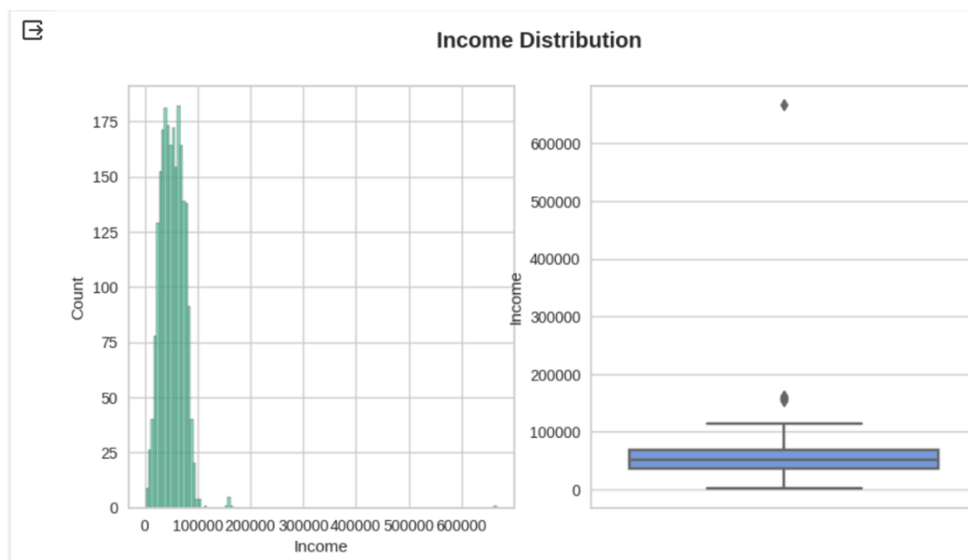


Рисунок 3.10 – Розподіл змінної «Income»

Джерело: Розроблено автором на основі [7]

За аналогією зі змінною «Age», існують декілька викидів між 3 та 4 кuartилями, один з яких, припускаємо, був введений до датасету з помилкою, адже відрізняється від середнього значення у більше ніж 10 разів. Застосуємо `replace_outliers()` до всіх змінних числового типу і проспостерігаємо як зміниться розподіл змінної «Income» (рис. 3.11).

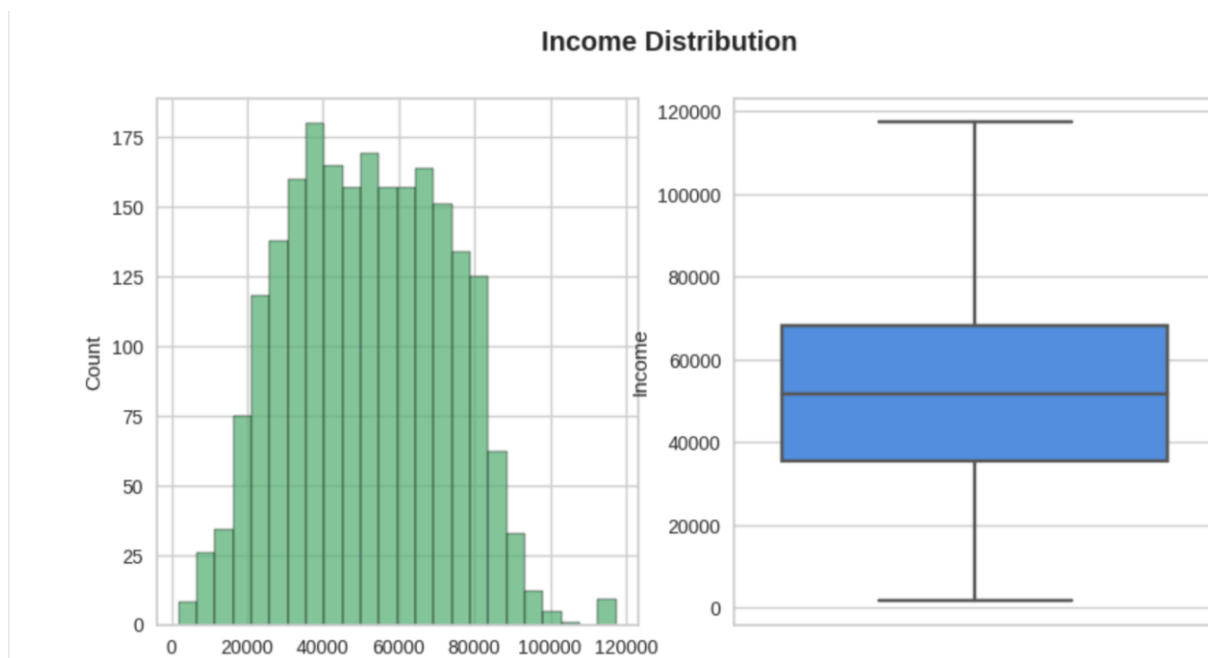


Рисунок 3.11 – Розподіл змінної «Income» після заміни аномальних значень

Джерело: Розроблено автором на основі [7]

Можемо спостерігати відсутність викидів за змінною «Income».

Продовжимо роботу з категоріальними змінними, а саме застосуємо `LabelEncoder` для переведення унікальних значень змінних у відповідні числові значення.

```
df["Education"].unique()

array(['Graduation', 'PhD', 'Master', 'Basic', '2n Cycle'], dtype=object)

df["Marital_Status"].unique()

array(['Single', 'Together', 'Married', 'Divorced', 'Widow', 'Alone',
       'Absurd', 'YOLO'], dtype=object)
```

Рисунок 3.12 – Унікальні категорії категоріальних змінних

Джерело: Розроблено автором на основі [7]

Наразі маємо 6 унікальних значень для поля «Education» (рис. 3.12): «Graduation», «PhD», «Master», «Basic», «2n Cycle», які будуть перетворені у «1», «2», ... , «6» відповідно. Для 8 унікальних значень змінної «Marital_Status» проводимо таку ж саму процедуру заміни значень з типу «object» на «int64» (рис. 3.13).

```
label_encoding = LabelEncoder()

for column in df.columns:
    if df[column].dtype == "object":
        df[column] = label_encoding.fit_transform(df[column])
```

Рисунок 3.13 – Алгоритм перетворення категоріальних змінних

Джерело: Розроблено автором на основі [7]

Розглянемо датасет із заміненними значеннями в категоріальних змінних (рис. 3.14).

df.head()

	ID	Education	Marital_Status	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits
0	5524.00000	2	4	58138.00000	0.00000	0.00000	58.00000	635.00000	81.00000
1	2174.00000	2	4	46344.00000	1.00000	1.00000	38.00000	11.00000	1.00000
2	4141.00000	2	5	71613.00000	0.00000	0.00000	26.00000	426.00000	49.00000
3	6182.00000	2	5	26646.00000	1.00000	0.00000	26.00000	11.00000	4.00000
4	5324.00000	4	3	58293.00000	1.00000	0.00000	94.00000	173.00000	43.00000

Рисунок 3.14 – Датасет після первинної підготовки даних

Джерело: Розроблено автором на основі [7]

Наразі маємо всі змінні числових типів. Для нормалізації даних використаємо інструмент `sklearn.preprocessing.RobustScaler` та проведемо її одразу для всіх змінних датасету (рис. 3.15).

```
scaled_data.head(10)
```

	Education	Marital_Status	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts
0	0.00000	0.00000	0.19934	0.00000	0.00000	0.18000	0.96046	2.28125	2.21759	2.30851
1	0.00000	0.00000	-0.16077	1.00000	1.00000	-0.22000	-0.33819	-0.21875	-0.28241	-0.21277
2	0.00000	0.50000	0.61078	0.00000	0.00000	-0.46000	0.52549	1.28125	0.27778	2.10638
3	0.00000	0.50000	-0.76222	1.00000	0.00000	-0.46000	-0.33819	-0.12500	-0.21759	-0.04255
4	2.00000	-0.50000	0.20407	1.00000	0.00000	0.90000	-0.00104	1.09375	0.23611	0.72340
5	1.00000	0.50000	0.33292	0.00000	1.00000	-0.66000	0.72112	1.06250	0.14352	-0.25532
6	0.00000	-1.00000	0.12291	0.00000	1.00000	-0.30000	0.12799	1.78125	0.44907	0.80851
7	2.00000	-0.50000	-0.55435	1.00000	0.00000	-0.34000	-0.20291	0.06250	-0.05093	-0.19149
8	2.00000	0.50000	-0.64909	1.00000	0.00000	-0.60000	-0.33195	-0.25000	-0.19907	-0.19149
9	2.00000	0.50000	-1.40336	1.00000	1.00000	0.38000	-0.30281	-0.25000	-0.28241	-0.23404

Рисунок 3.15 – Нормалізація показників датасету

Джерело: Розроблено автором на основі [7]

3.2. Аналіз результатів сегментування споживачів та визначення латентних факторів їхньої поведінки

Оберемо метод PCA (Principal component analysis) для зменшення розмірності досліджуваного датасету та виділення основних компонент. З метою визначення оптимальної кількості основних компонент визначимо `n_components=None` при оригінальному оголошенні моделі (рис. 3.16).

```
deco_pca = PCA(n_components=None)
deco_pca.fit(scaled_data)
data_pca = deco_pca.transform(scaled_data)
```

Рисунок 3.16 – Побудова моделі зменшення розмірності

Джерело: Розроблено автором на основі [7]

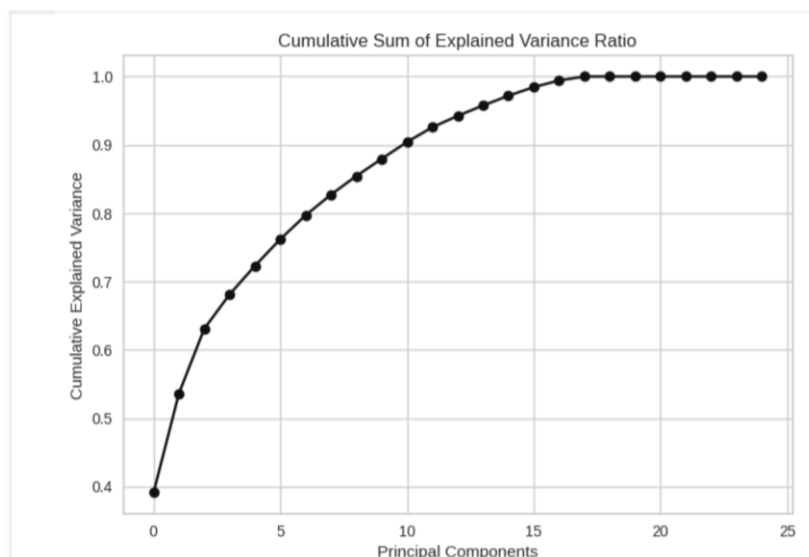


Рисунок 3.17 – Розподіл кумулятивної пояснювальної варіації за різної кількості
ГОЛОВНИХ КОМПОНЕНТ

Джерело: Розроблено автором на основі [7]

На рис. 3.17 можемо спостерігати показник кумулятивного пояснювального рівня варіації, іншими словами – яку частку варіації оригінального датасету пояснює даний вибір кількості основних компонент. Так як кількість змінних, які приймають участь в аналізі основних компонент – 25 (оригінальний датасет, за виключенням змінної «ID»), максимально можливий показник основних компонент сягає 24 – випадок, коли 23 змінні розцінюватимуться як головні компоненти і ще 2 змінні – як єдина компонента.

На графіка також можемо помітити, що на рівні 17 компонент показник кумулятивної варіації досягає 1 і перестає рости, тому обрана нами в перспективі оптимальна кількість основних компонент на даному етапі вже може бути обмежена 17-тьма.

До прикладу, розглядаючи більш детально виділення, наприклад, 11-ти основних компонент, можемо вже бути впевненими у тому, що показник кумулятивної варіації перевищує 0.9, і що вже можемо вважати достатньо точним результатом (рис. 3.18).

```
[102] deco_pca_11 = PCA(n_components=11)
      deco_pca_11.fit(scaled_data)
      deco_pca_11.transform(scaled_data)

      len(deco_pca_11.explained_variance_ratio_), sum(deco_pca_11.explained_variance_ratio_)

(11, 0.904130656795651)
```

Рисунок 3.18 – Модель основних компонент з кількістю компонент n=11

Джерело: Розроблено автором на основі [7]

Дослідимо детальніше пояснювальні варіації з кількістю компонент до 11 (рис. 3.19).

```
[0.39094547 0.14504635 0.09477142 0.05036244 0.04176278 0.0390534
 0.03492715 0.03028826 0.02719264 0.02511884 0.02466191]
```

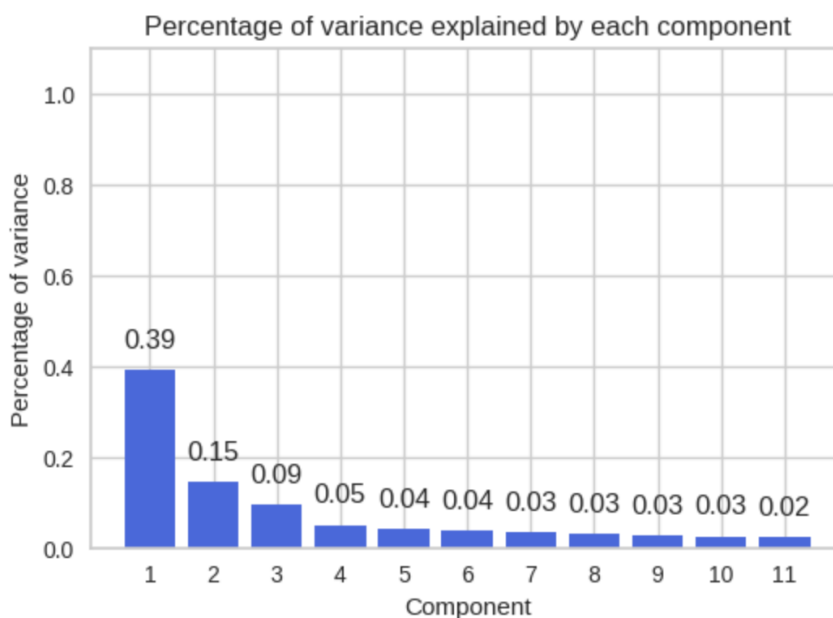


Рисунок 3.19– Відсоток пояснювальної варіації основних компонент

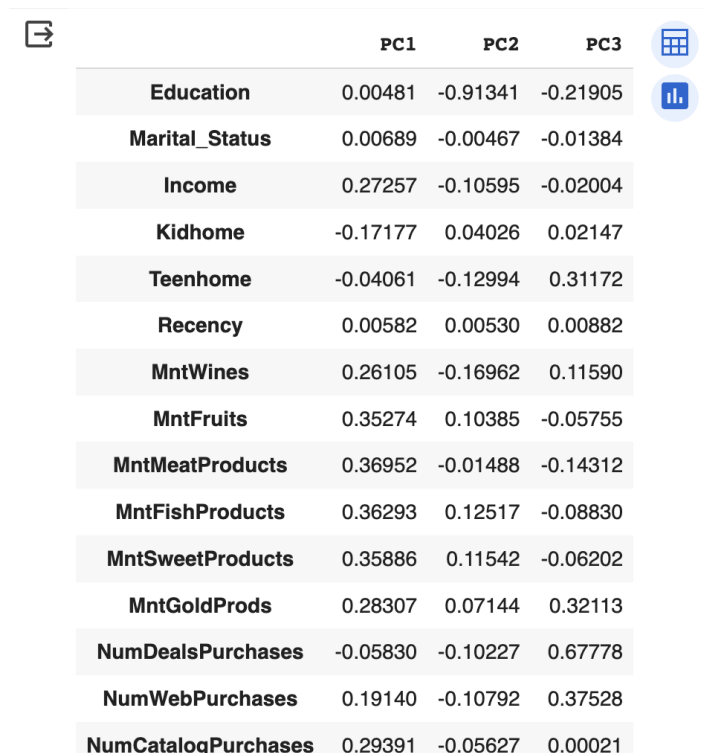
Джерело: Розроблено автором на основі [7]

З даного графіку та наведених показників відсотків пояснювальної варіації можемо зробити наступні висновки: найбільше інформативності в собі несе розподіл на 1 – 3 компоненти, після 4 компоненти включно показник пояснювальної варіації зупиняє активних рух та зупиняється на позначці 0.5-0.2 протягом наступних 8 компонент. Отже, вважатимемо найбільш доречним та обґрунтованим показником пояснювальної варіації вибір трьох головних компонент.

```
pca = PCA(n_components=3, random_state = 0)
pca.fit(scaled_data)

PCA_dataset = pd.DataFrame(pca.transform(scaled_data), columns=([ "PC1", "PC2", "PC3"])))
```

Рисунок 3.20 – Модель основних компонент з кількістю компонент n=3

Джерело: Розроблено автором на основі [7]


	PC1	PC2	PC3
Education	0.00481	-0.91341	-0.21905
Marital_Status	0.00689	-0.00467	-0.01384
Income	0.27257	-0.10595	-0.02004
Kidhome	-0.17177	0.04026	0.02147
Teenhome	-0.04061	-0.12994	0.31172
Recency	0.00582	0.00530	0.00882
MntWines	0.26105	-0.16962	0.11590
MntFruits	0.35274	0.10385	-0.05755
MntMeatProducts	0.36952	-0.01488	-0.14312
MntFishProducts	0.36293	0.12517	-0.08830
MntSweetProducts	0.35886	0.11542	-0.06202
MntGoldProds	0.28307	0.07144	0.32113
NumDealsPurchases	-0.05830	-0.10227	0.67778
NumWebPurchases	0.19140	-0.10792	0.37528
NumCatalogPurchases	0.29391	-0.05627	0.00021

Рисунок 3.21 – Критерії приналежності змінних до головних компонент

Джерело: Розроблено автором на основі [7]

До основної компоненти 1 можемо віднести такі змінні, як «MntFruits», «MntMeatProducts», «MntFishProducts», «MntSweetProducts» та «Income» (рис. 3.20-3.21). Даній головній компоненті дамо назву «Витрати та доходи», адже перелічені вище змінні, які є найбільш корельованими з першою основною компонентною, всі відображають напрямки витрат споживачів. Другій основній компоненті дамо назву «Освіта» за змінною, яка найбільше корелює з даною компонентною, інші змінні мають не значні ступені кореляції. Третій основній компоненті можемо дати назву «Кількість покупок», адже найбільший зв'язок вона має зі змінними «NumDealsPurchases» та «NumWebPurchases».

Інші змінні, що не представлені на малюнку, мали незначні відношення до визначених трьох компонент.

Розглянемо розподіл досліджуваних об'єктів на трьохмірному малюнку, кожна з осей якого відповідає одній з визначених компонент (рис. 3.22).

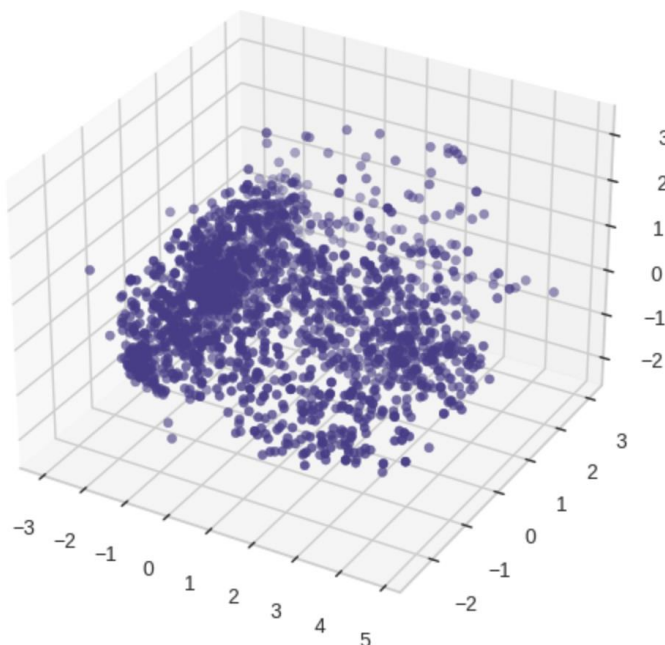


Рисунок 3.22 – Розподіл спостережень за визначеними головними компонентами

Джерело: Розроблено автором на основі [7]

За даним зображенням можемо наперед побачити скупчення схожих спостережень ближче до площини X-Z, що в контексті головних компонент відповідає площині PC1-PC3.

Для початку, скористаємось методом К-середніх для визначення відповідних кластерів серед споживачів. Для визначення можливого оптимального значення кількості кластерів скористаємось `yellowbrick.cluster.KElbowVisualizer` (рис. 3.23).

```
model = KMeans()
visualizer = KElbowVisualizer(model, k=10)

visualizer.fit(PCA_dataset)
visualizer.show()
```

Рисунок 3.23 – Модель кластеризації методом К-середніх

Джерело: Розроблено автором на основі [7]

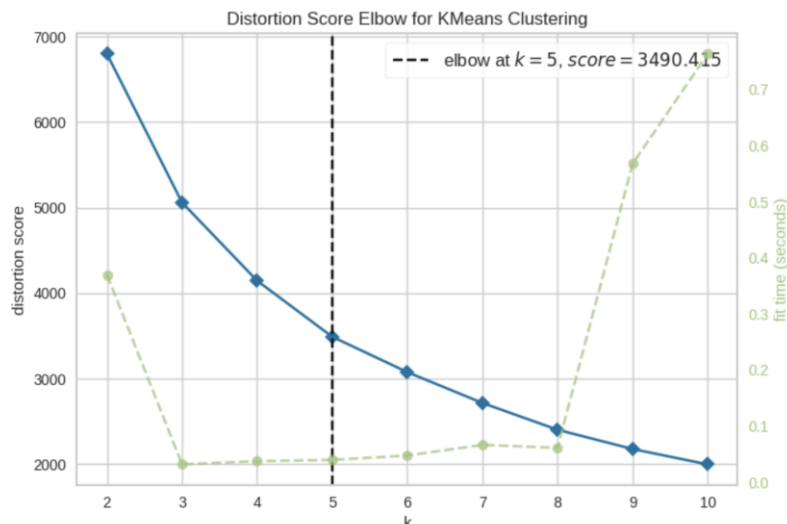


Рисунок 3.24 – Показник викривлення для різної кількості кластерів

Джерело: Розроблено автором на основі [7]

Для визначення оптимальної кількості кластерів всередині KelbowVisualizer можна використовувати різні оцінні метрики [2]. За замовчуванням метод використовує «distortion score», або показник викривлення, який розраховується як сума квадратів відстаней від кожної точки до відповідного центру кластера. За даним показником бачимо оптимальну кількість кластерів визначену як 5 (рис. 3.24). Використаємо як метрики для оцінки показник «silhouette» та «calinski_harabasz» (рис. 3.25-3.26).

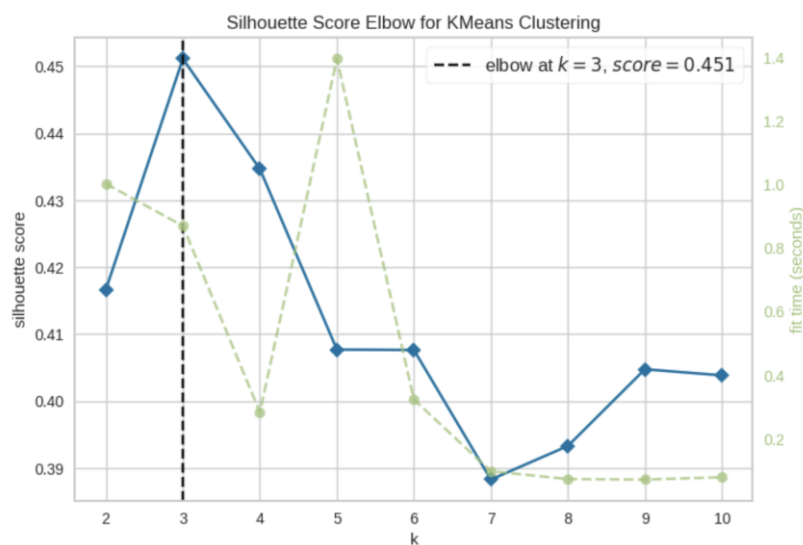


Рисунок 3.25 – Показник силуету для різної кількості кластерів

Джерело: Розроблено автором на основі [7]

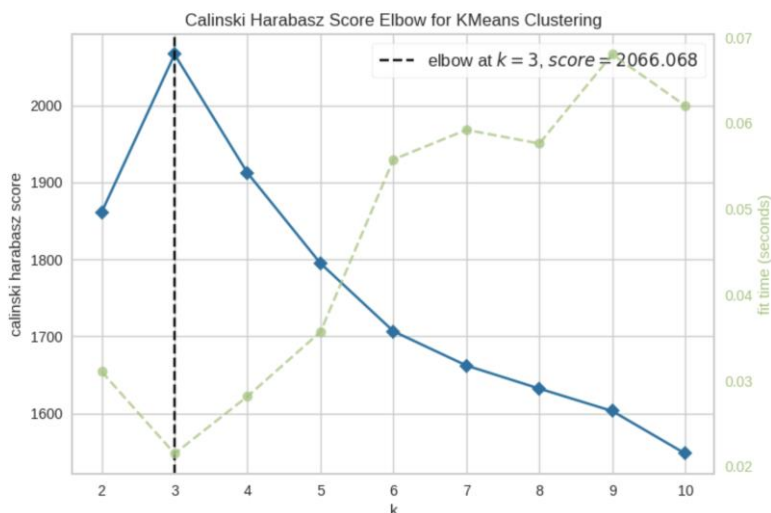


Рисунок 3.26 – Індекс Калінські-Харабаша для різної кількості кластерів

Джерело: Розроблено автором на основі [7]

За цими двома показниками отримаємо оптимальну кількість кластерів – 3. Тож, маючи результати розрахунків оптимальної кількості кластерів за методами «distortion score», «silhouette» та «calinski_harabasz» продовжуватимемо дослідження, порівнюючи результати обчислень для 3-ох та 5-ти кластерів.

Побудуємо модель К-середніх для кількості кластерів 3 та розрахуємо показник `explained_variance_ratio` – показник пояснювальної варіації (рис. 2.27).

```
kmeans = KMeans(n_clusters=3, random_state = 0, n_init=10)

Kmeans_clusters = kmeans.fit_predict(PCA_dataset)
PCA_dataset['Clusters'] = Kmeans_clusters

df['Clusters'] = Kmeans_clusters

total_var = pca.explained_variance_ratio_.sum() * 100
total_var

63.07632350190121
```

Рисунок 3.27 – Модель кластеризації методом К-середніх з кількістю кластерів $p=3$

Джерело: Розроблено автором на основі [7]

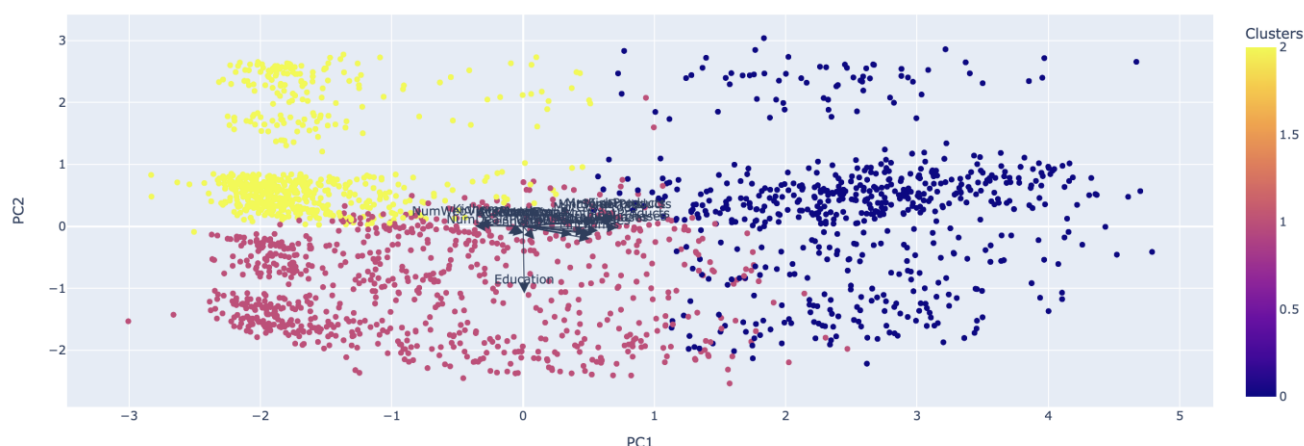


Рисунок 3.28 – Розподіл спостережень за кластерами $n=3$ в контексті першої та другої
ГОЛОВНИХ КОМПОНЕНТ

Джерело: Розроблено автором на основі [7]

При показнику `explained_variance_ratio` рівному 63%, побудували візуалізацію визначених кластерів за осями першої та другої основних компонент – «Витрати та доходи» та «Освіта» (рис. 3.28). На зображенні можемо бачити виділені напрямки основних компонент та відносну перпендикулярність компонент 1 та 2 в даній площині, а також візуально чіткий розподіл спостережень на кластери. Вдаючи до більш детальної характеристики споживачів, можемо сказати, що перша частина споживачів характеризується більш низьким рівнем доходів та витрат та більш низьким рівнем освіти (спостереження рожевого кольору), друга частина – так само відносно невеликим рівнем доходів та витрат і високим рівнем освіти (спостереження жовтого кольору). Остання група споживачів характеризується високим рівнем доходів та витрат і не має переважного розподілу за освітою – розкидана за всіма рівнями.

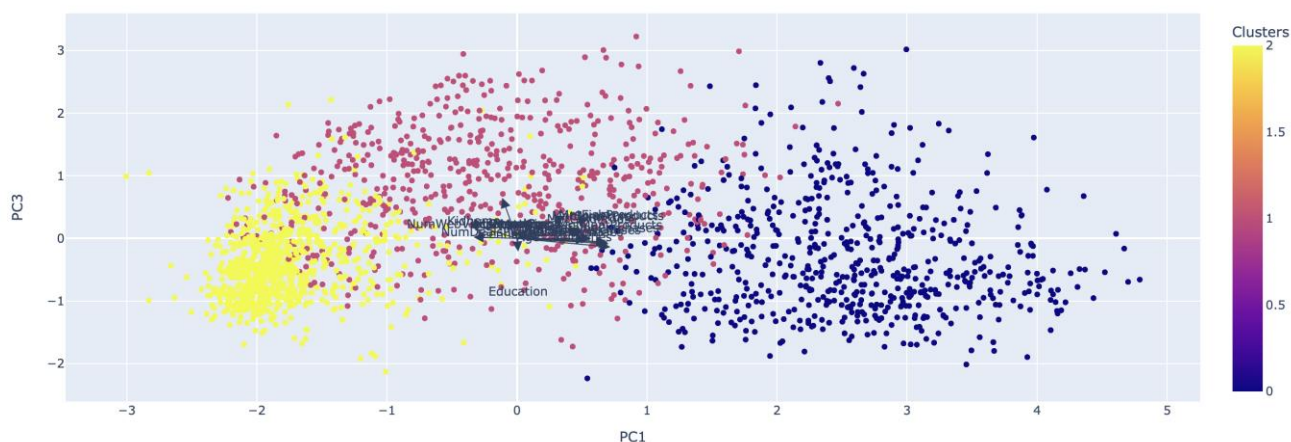


Рисунок 3.29 – Розподіл спостережень за кластерами $n=3$ в контексті першої та третьої
ГОЛОВНИХ КОМПОНЕНТ

Джерело: Розроблено автором на основі [7]

Візуалізувавши розподіл споживачів в системі координат основних компонент 1 та 3 (рис. 3.29), можемо так само створити характеристичні портрети покупців різних груп. До першої групи відносимо споживачів, що виділені на зображенні жовтим кольором, та мають відносно нижчий рівень доходів та витрат і відповідно нижчий рівень кількості покупок. Другий кластер характеризується відносно середнім рівнем доходів та витрат, і вищим рівнем кількості покупок в порівнянні з першою групою. І остання група покупців характеризується високим рівнем доходів та витрат, і низьким рівнем кількості покупок, на базі чого можемо зробити припущення про їх характер поведінки при здійсненні покупок – більш якісні товари з високими цінінками та невелика кількість купівель.

Повторимо розрахунки для кількості кластерів рівною 5, яка була визначеною оптимальною за показником викривлення (distortion score). В результаті отримуємо такий самий рівень `explained_variance_ratio` = 63%, і наступний візуальний розподіл за кластерами відносно головних компонент (рис.3.30-3.31).

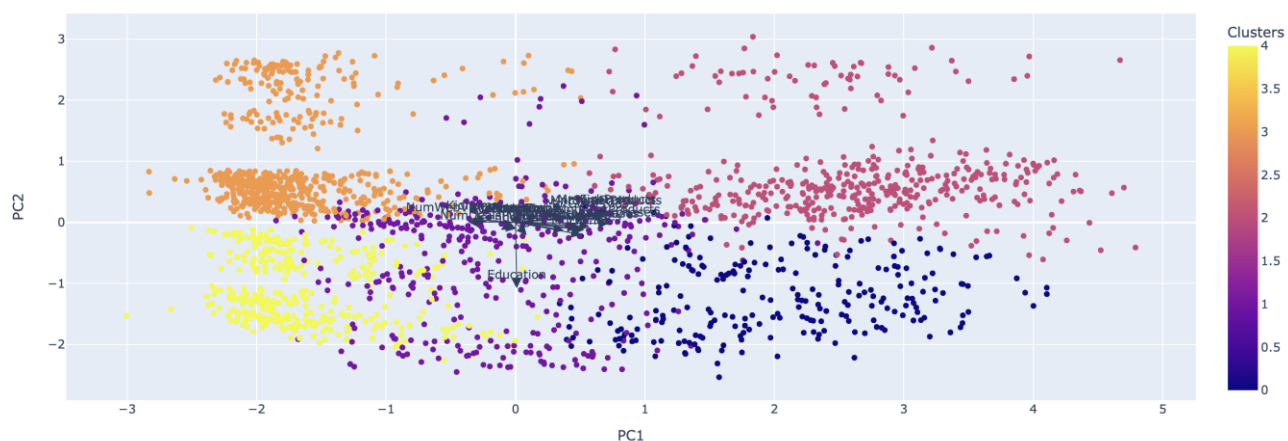


Рисунок 3.30 – Розподіл спостережень за кластерами $n=5$ в контексті першої та другої
ГОЛОВНИХ КОМПОНЕНТ

Джерело: Розроблено автором на основі [7]

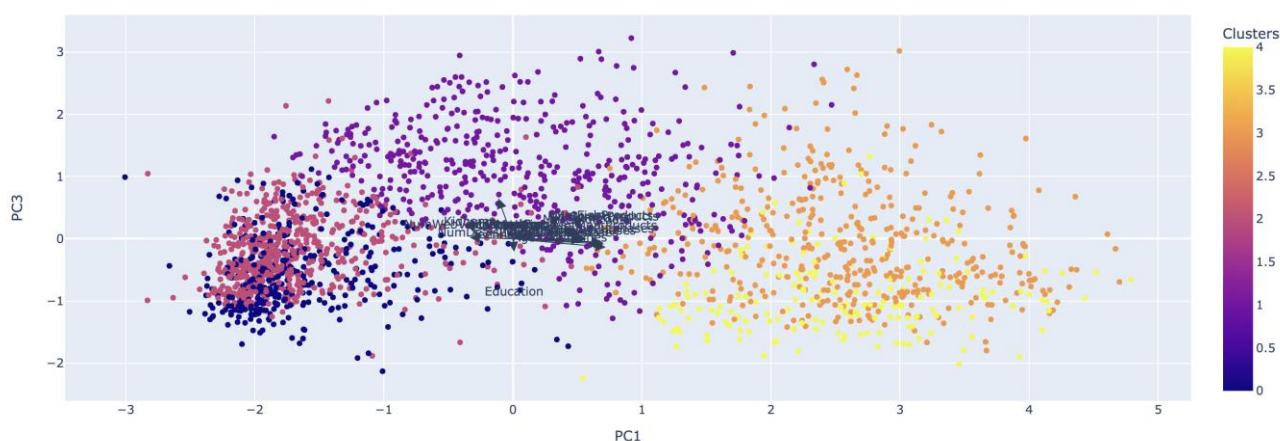


Рисунок 3.31 – Розподіл спостережень за кластерами $n=5$ в контексті першої та третьої
ГОЛОВНИХ КОМПОНЕНТ

Джерело: Розроблено автором на основі [7]

Можемо побачити на візуалізаціях за головними компонентами PC1-PC3, що кластери 0 та 2, а також 3 та 4 є достатньо близькими за характеристиками віднесених до них спостережень. Однак за графіком розсіювання на площині PC1-PC2 можна помітити чіткий розподіл на рівні $PC2=0$.

Проведемо додаткові дослідження за показниками силуету, індексу Девіса-Болдіна та індексу Калінскі-Харабаша для значень кількості кластерів між 2-ма та 5-ма. Для складання найбільш повної картини дослідження скористаємося не

тільки методом К-середніх, але і ієрархічною кластеризацією. Отримуємо наступні результати (табл. 3.1-3.2).

Таблиця 3.1 – Показники точності кластеризації методом К-середніх для кількості кластерів 2-5

Кластеризація К-середніх			
Кількість кластерів	Показник силуету	Індекс Девіса-Болдіна	Індекс Калінські-Харабаша
2	0.416652	0.998637	1861.010245
3	0.451137	0.884798	2066.067715
4	0.434678	0.911392	1912.087615
5	0.407701	1.002047	1794.778036

Джерело: Розроблено автором на основі [7]

Таблиця 3.2 – Показники точності ієрархічної кластеризації для кількості кластерів 2-5

Ієрархічна кластеризація			
Кількість кластерів	Показник силуету	Індекс Девіса-Болдіна	Індекс Калінські-Харабаша
2	0.417373	0.971674	1820.509389
3	0.451070	0.885368	2065.536414
4	0.417143	0.945028	1777.906116
5	0.373309	1.122873	1643.1894

Джерело: Розроблено автором на основі [7]

Порівнюючи значення отриманих результатів, можемо дійти висновку, що кількість кластерів рівна 3-ьом є найбільш оптимальною, адже має найбільші значення показників силуету та індексу Калінські-Харабаша, а також найнижчі значення індекса Девіса-Болдіна. Звернемо увагу, що у відносних величинах були отримані однакові результати за обома методами кластеризації – методом К-середніх та ієрархічною кластеризацією: вони обидва показали оптимальну кількість кластерів рівну 3-ьом, хоча абсолютні показники індексів незначно

відрізняються. Найменш оптимальною кількістю кластерів, за отриманими результатами, можемо вважати 5, хоч вона і була визначена як оптимальна за показником викривлення. Робимо припущення, що показник викривлення (distortion score) не є оптимальною метрикою для оцінки кластеризації даного датасету.

Отже, провівши кластерний аналіз з використання двох методів моделювання – К-середніх та ієрархічної кластеризації – можемо дійти до висновку, що найбільш оптимальною кількістю кластерів за показниками силуєту, індексу Девіса-Болдіна та індексу Калінскі-Харабаша є 3.

Говорячи про розподіл спостережень за кластерами в рамках оригінальних змінних датасету, можемо побачити наступне (рис. 3.32).

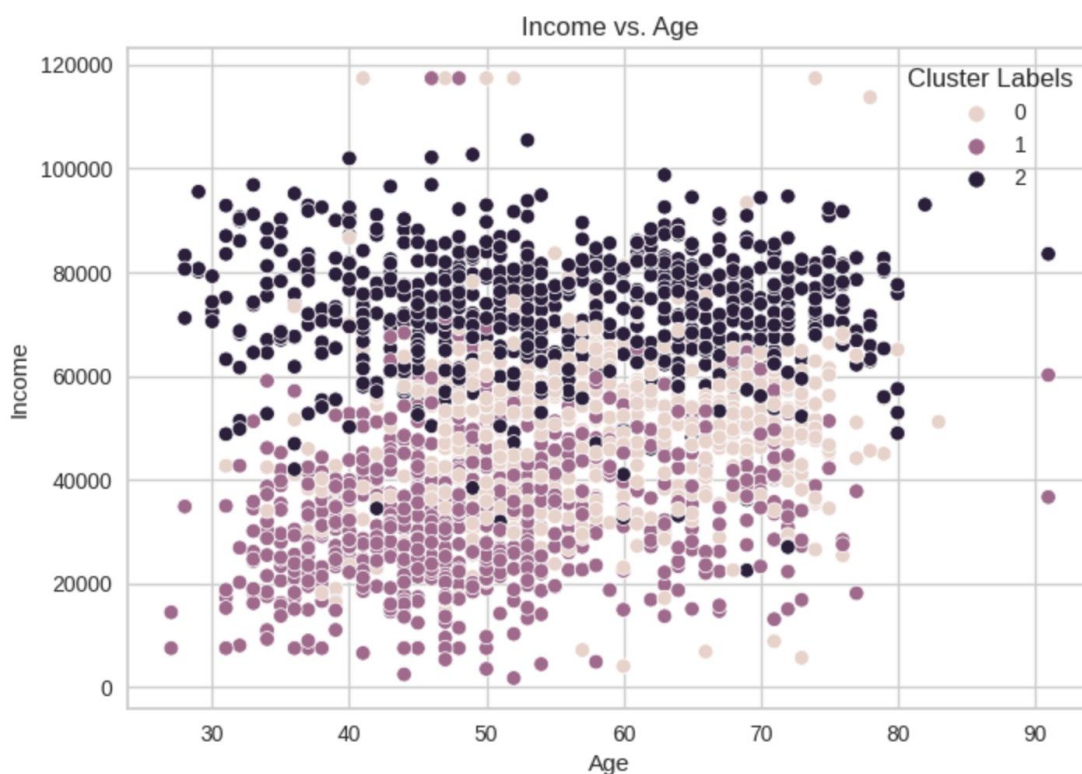


Рисунок 3.32 – Розподіл спостережень за кластерами в контексті факторів віку та доходу

Джерело: Розроблено автором на основі [7]

При розподілі спостережень за віссю координат Вік-Дохід, можемо зробити наступні висновки про групи споживачів:

- перша група має дохід переважно в рамках 10 тис. дол. – 45 тис. дол., і є більш молодшою порівняно з іншими споживачами – 30-55 років;
- друга група має дохід близько 45-70 тис. дол. та віком сягає від 45 до 70 років;
- третя група споживачів має найбільший дохід відносно інших груп – 60 тис. дол. – 90 тис. дол. та за віком є розподіленою переважно в рамках 40 – 75 років.

Цікавим є розподіл покупців в рамках «Повертаємість» (кількість днів після останнього придбання товару) – «Витрати» (рис. 3.33).

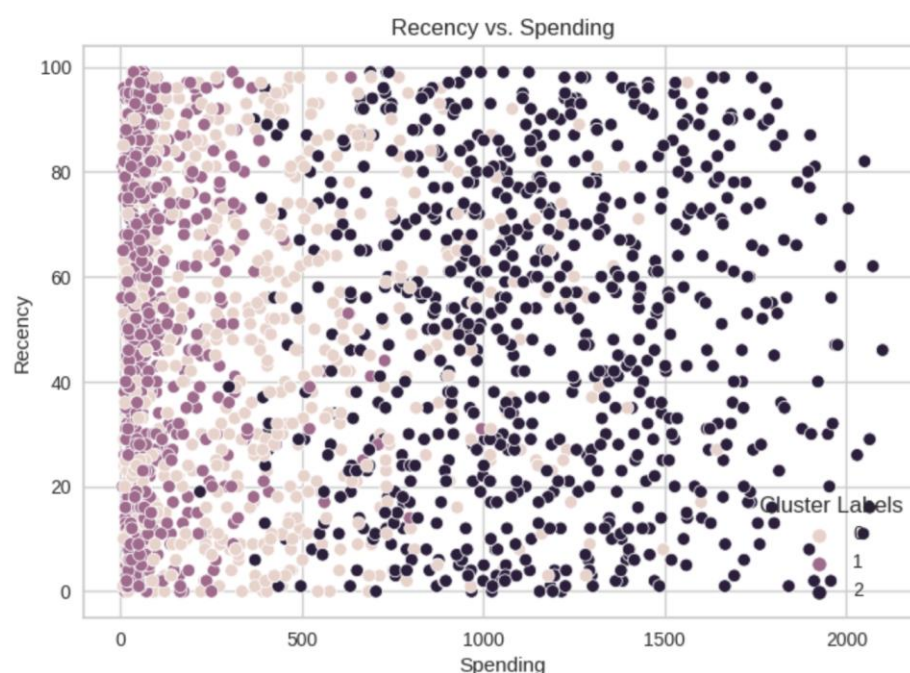


Рисунок 3.33 – Розподіл спостережень за кластерами в контексті факторів витрат та часу здійснення останньої покупки

Джерело: Розроблено автором на основі [7]

Як бачимо, всі три групи відрізняються за значенням витрат:

- перша група споживачів витрачає переважно 0-250\$
- друга група – 250-750\$
- третя група – 750-2000\$

Однак кількість днів з останньої покупки не є різною для всіх трьох груп, тому робимо висновок, що покупці, що витрачають більше коштів в даній торговій

мережі не здійснюють покупки частіше, а скоріше мають тенденцію до придбання товарів з більш високим чеком. Даний висновок може бути корисним при побудові наступної маркетингової кампанії, або при оперативному коригуванні поточної.

3.3. Аналіз результатів моделювання поведінки цільової аудиторії на базі інструментів RFM та машинного навчання

Проведемо сегментацію споживачів за показниками RFM, з метою подальшого визначення необхідних маркетингових дій для кожної з категорій покупців.

Підрахуємо основні показники для встановлення загального RFM_SCORE, а саме - frequency_score та monetary_score, враховуючи те, що третій необхідний показник вже наявний в оригінальному датасеті – змінна «Recency».

Показник «monetary» - суму загальних витрат клієнта – підраховуватимемо шляхом складання всіх змінних з «Mnt», а саме: «MntWines», «MntFruits», «MntMeatProducts», «MntFishProducts», «MntSweetProducts», «MntGoldProds», та додамо нову змінну «monetary» у датафрейм.

Показник «frequency» – частоту здійснюваних покупок – визначатимемо шляхом складання значень всіх змінних, що містять «Purchases», а саме: «NumDealsPurchases», «NumWebPurchases», «NumCatalogPurchases», «NumStorePurchases», та додамо нову змінну до датасету.

Виокремлюючи в новий датафрейм всі необхідні змінні для знаходження загального RFM_SCORE, отримуємо наступні результати (рис. 3.34-3.35).

	ID	Recency	frequency	monetary
0	5524.00000	58.00000	25.00000	1551.50000
1	2174.00000	38.00000	6.00000	27.00000
2	4141.00000	26.00000	21.00000	776.00000
3	6182.00000	26.00000	8.00000	53.00000
4	5324.00000	94.00000	19.00000	422.00000
5	7446.00000	16.00000	22.00000	716.00000
6	965.00000	34.00000	21.00000	590.00000
7	6177.00000	32.00000	10.00000	169.00000
8	4855.00000	19.00000	6.00000	46.00000
9	5899.00000	68.00000	2.00000	49.00000

Рисунок 3.34 – Абсолютні величини показників RFM для відповідних спостережень

Джерело: Розроблено автором на основі [7]

	count	mean	std	min	25%	50%	75%	max
ID	2240.00000	5592.15982	3246.66220	0.00000	2828.25000	5458.50000	8427.75000	11191.00000
Recency	2240.00000	49.10938	28.96245	0.00000	24.00000	49.00000	74.00000	99.00000
frequency	2240.00000	14.69063	7.47210	0.00000	8.00000	15.00000	21.00000	35.00000
monetary	2240.00000	568.20670	547.03674	5.00000	68.75000	395.00000	990.50000	2099.50000

Рисунок 3.35 – Описова статистика показників RFM

Джерело: Розроблено автором на основі [7]

За таблицею характеристик можемо бачити, що, до прикладу, середня кількість замовлень, здійснювана типовим покупцем наближається до 14, а середня кількість коштів, що витрачається клієнтом знаходиться на рівні 568 дол. Крім того, показник *monetary* є найменш концентрованим, адже його стандартне відхилення в абсолютній величині дорівнює значенню середнього. На противагу цьому, *frequency* є більш концентрованим показником, адже його стандартне відхилення є в половину меншим за абсолютне значення середнього. Показник *recency* займає посереднє значення та не відрізняється критично розкиданими або концентрованими спостереженнями. Крім того, цікавим є спостереження, що показник *monetary*, в порівнянні з іншими показниками, має відносно велику

різницю між середнім значенням та медіаною, що говорить про його асиметричність. Показники recency та frequency є симетричними.

Виконаємо дискретизацію трьох визначених показників та отримаємо наступні результати (рис. 3.36).

	ID	Recency	frequency	monetary	recency_score	frequency_score	monetary_score
0	5524.00000	58.00000	25.00000	1551.50000	3	5	5
1	2174.00000	38.00000	6.00000	27.00000	4	1	1
2	4141.00000	26.00000	21.00000	776.00000	4	4	4
3	6182.00000	26.00000	8.00000	53.00000	4	2	1
4	5324.00000	94.00000	19.00000	422.00000	1	4	3

Рисунок 3.36 Дискретні показники recency – frequency - monetary

Джерело: Розроблено автором на основі [7]

Наразі маємо в датасеті абсолютні значення показників recency, frequency та monetary, і відповідні їм дискретні значення – чим більший показник, тим більшим є його дискретний аналог, окрім показника recency – в даному випадку чим більше днів з останньої покупки пройшло, тим меншим буде показник score. Класичний показник RFM ми розраховували б шляхом об'єднання трьох попередньо визначених показників, проте на даному етапі, для подальшого розділення спостережень на групи за ознакою необхідності здійснення дій компанією, використаємо перші два фактори як ті, що характеризують дії покупців в часових рамках.

```
seg_map = {
    r'[1-2][1-2]': 'hibernating',
    r'[1-2][3-4]': 'at_risk',
    r'[1-2]5': 'cant_loose',
    r'3[1-2]': 'about_to_sleep',
    r'33': 'need_attention',
    r'[3-4][4-5]': 'loyal_customers',
    r'41': 'promising',
    r'51': 'new_customers',
    r'[4-5][2-3]': 'potential_loyalists',
    r'5[4-5]': 'champions'
}
```

Рисунок 3.37 – Алгоритм розподілу споживачів за показниками RFM

Джерело: Розроблено автором на основі [7]

Розподілятимемо покупців за правилом, описаним на рис. 3.37. Отже, ті покупці, що мають дискретне значення *resency* та *frequency* в межах 1-2 будуть визначені як «hibernating» - сплячі, тобто такі, від яких компанія не може очікувати активності найближчим часом. Ті покупці, що мають показники *resency* в межах 1-2 та *frequency* – 3-4 будуть визначені як група «at_risk», тобто найближчим часом компанія може розраховувати на те, що, за відсутності відповідних дій, покупці перейдуть у «сплячий» стан, тобто також знизять свою активність. Покупці з рівнем *resency* в межах 1-2 та *frequency* рівним 5, тобто такі які зазвичай приходять часто, але останнім часом мають паузу у відвідуваності, будуть віднесені до групи «cant_loose», тобто таких, які є цінними для компанії з точки зору активності, і необхідно обов’язково здійснити заходи з метою запобігання їх переходу в категорії «at_risk» та «hibernating». Покупці, зі значенням *resency* рівним 3 та *frequency* рівним 1-2 будуть розцінені як ті, що найближчим часом можуть перейти в групу «hibernating», а саме «about_to_sleep», бо мають відповідне значення останньої здійсненої покупки на середньому рівні відносно інших, проте частота здійснення – на мінімальних рівнях. До категорії покупців «need_attention» будуть віднесені ті, що мають середні значення за обома показниками, тому при заохочувальних діях можуть стати цінними покупцями для торгової мережі, однак при відсутності дій – можуть легко перейти до менш активних груп споживачів. До групи «loyal_customers» будуть віднесені ті споживачі, які мають *resency* на рівні 3-4, тобто нещодавно здійснювали покупку в торговій мережі, і *frequency* на рівні 4-5, тобто мають відносно високий рівень відвідуваності в часовому контексті. Покупці, що характеризуються значенням *resency* на рівні 4 та *frequency* на рівні 1, можуть бути класифіковані як «promising», тобто такі, що нещодавно здійснили покупку, але ще не здійснюють їх так часто, щоб вважатися найбільш цінними. Близькими за характеристиками є покупці групи «new_customers», які мають *resency* на рівні 5 та *frequency* на рівні 1, і тільки нещодавно здійснювали покупку, проте ще не мають напрацьованої досвідом тенденції до частого відвідування. Клієнти, що мають високі показник *resency* – 4-5 та відносно середній показник *frequency* – 2-3, будуть віднесені до групи «potential_loyalists», адже мають

хороший потенціал для того, щоб, при збільшенні частоти відвідуваності, стати найбільш цінними покупцями. І до групи, що є найбільш цінною для компанії, та яка має показники recency та frequency на максимальних рівнях, відносяться покупці з умовною назвою «champions». Відповідно до описаного правила розподілу віднесемо кожного з клієнтів до відповідних груп.

Матимемо наступний результат (рис. 3.38-3.39).

	ID	Recency	frequency	monetary	recency_score	frequency_score	monetary_score	RFM_SCORE	segment
0	5524.00000	58.00000	25.00000	1551.50000	3	5	5	35	loyal_customers
1	2174.00000	38.00000	6.00000	27.00000	4	1	1	41	promising
2	4141.00000	26.00000	21.00000	776.00000	4	4	4	44	loyal_customers
3	6182.00000	26.00000	8.00000	53.00000	4	2	1	42	potential_loyalists
4	5324.00000	94.00000	19.00000	422.00000	1	4	3	14	at_risk
5	7446.00000	16.00000	22.00000	716.00000	5	4	4	54	champions
6	965.00000	34.00000	21.00000	590.00000	4	4	3	44	loyal_customers
7	6177.00000	32.00000	10.00000	169.00000	4	2	2	42	potential_loyalists
8	4855.00000	19.00000	6.00000	46.00000	5	1	1	51	new_customers
9	5899.00000	68.00000	2.00000	49.00000	2	1	1	21	hibernating

Рисунок 3.38 – Групування цільової аудиторії за показниками RFM

Джерело: Розроблено автором на основі [7]

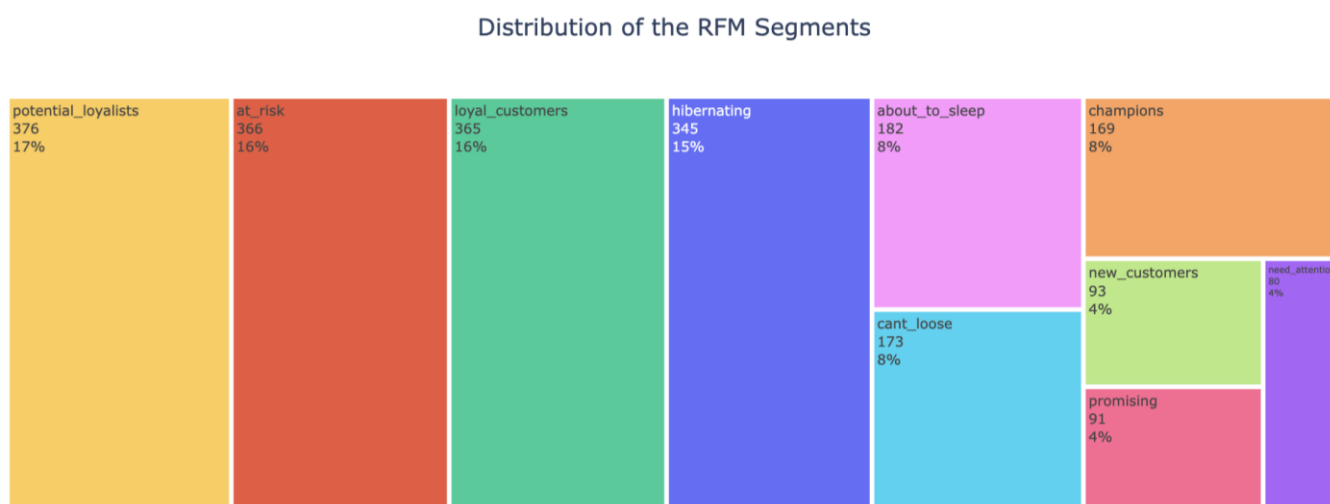


Рисунок 3.39 – Групування цільової аудиторії за показниками RFM

Джерело: Розроблено автором на основі [7]

З метою побудови найбільш вдалої маркетингової кампанії, що базуватиметься на розумінні характеристик основних груп споживачів, та приносить найбільшу вигоду підприємству, виконаємо регресійний аналіз однієї

зі змінних RFM, а саме `monetary`, тобто спробуємо побудувати найкращу модель, яка прогнозувала би яку кількість коштів покупець буде готовий заплатити за товари. Для цього скористаємося декількома видами моделей – рідж регресія, ласо регресія, еластікнет, поліноміальна регресія та класична лінійна регресія.

Так як цільова змінна є сумою таких змінних як «`MntWines`», «`MntFruits`», «`MntMeatProducts`», «`MntFishProducts`», «`MntSweetProducts`», «`MntGoldProds`», усунемо їх з датасету та дослідимо можливість побудови моделі з іншими змінними як пояснювальними показниками. Спершу нормалізуємо наші змінні та виконаємо розподіл на тренувальну та тестову вибірки з відношенням 2:1.

Почнемо з побудови `ridge` регресії з декількома значенням `alpha` з метою встановлення оптимального (рис.3.40).

```
for alpha in range(1, 500, 50):
    ridge = RidgeCV(alphas=[alpha/1000,], cv=5)
    ridge.fit(train_x, train_y)
    rmse_train = rmse_score(ridge, train_x, train_y)
    rmse_test = rmse_score(ridge, test_x, test_y)
    print(alpha/1000, ": R2 train -", ridge.score(train_x, train_y), ", RMSE:", rmse_train)
    print(alpha/1000, ": R2 test -", ridge.score(test_x, test_y), ", RMSE:", rmse_test)
```

0.001 : R2 train - 0.7970065098980567 , RMSE: 0.006673074048877308
 0.001 : R2 test - 0.7787733688054123 , RMSE: 0.006736172321372632
 0.051 : R2 train - 0.5349574908512691 , RMSE: 0.010100228881085022
 0.051 : R2 test - 0.6262565000502747 , RMSE: 0.008755507203124463
 0.101 : R2 train - 0.4584641034869704 , RMSE: 0.010899297669635728
 0.101 : R2 test - 0.5144941769624716 , RMSE: 0.009979107637525061
 0.151 : R2 train - 0.39865109142628374 , RMSE: 0.011485453619068303
 0.151 : R2 test - 0.43586460220101064 , RMSE: 0.010756875971131851
 0.201 : R2 train - 0.351412422177659 , RMSE: 0.011928043165757478
 0.201 : R2 test - 0.37766142507378797 , RMSE: 0.011298163259849312
 0.251 : R2 train - 0.3136106929232497 , RMSE: 0.012270722746208584
 0.251 : R2 test - 0.33295583627913716 , RMSE: 0.011696927009777744
 0.301 : R2 train - 0.28286354118145196 , RMSE: 0.012542548527696108
 0.301 : R2 test - 0.2975955200891922 , RMSE: 0.012002953453402375
 0.351 : R2 train - 0.25745018421758825 , RMSE: 0.012762850620315527
 0.351 : R2 test - 0.26895413240221744 , RMSE: 0.012245225826106243
 0.401 : R2 train - 0.2361357057733472 , RMSE: 0.012944729923899369
 0.401 : R2 test - 0.24529653728487744 , RMSE: 0.012441783985325512
 0.451 : R2 train - 0.21802488868826964 , RMSE: 0.013097287296402454
 0.451 : R2 test - 0.22543354626561996 , RMSE: 0.012604447900139497

Рисунок 3.40 – Коефіцієнти детермінації моделей рідж регресії з різними показниками `alpha`

Джерело: Розроблено автором на основі [7]

Аналізуючи отримані показники, можемо зробити цікаве спостереження, що для більшої частини значень `alpha` коефіцієнт детермінації є вищим для тестової вибірки, ніж для тренувальної. Однак дане спостереження не розповсюджується на найменше значення `alpha` з вибраних нами. Показники середньоквадратичної

похибки є вищими скоріше для тренувальних вибірок. Можемо зауважити також, що зі збільшенням α коефіцієнт детермінації стрімко спадає, і найкращу модель отримуємо при найменшому значенні α – 0.001: коефіцієнти детермінації на тренувальній та тестовій вибірках – 0.7970 та 0.7787 відповідно, показник середньоквадратичної похибки – 0.0066 та 0.0067.

```
ridge.coef_
```

```
array([-0.10511167, -0.19509079,  3.37546355, -0.07158368, -0.00468902,
       -1.36067357, -0.40993426, -0.04520935, -0.4516448 ,  0.01473681,
       -0.04920439,  0.          ,  0.          ,  0.          ,  0.          ,
        0.          ,  0.          ,  0.          , -2.25850329,  1.40088639,
       -0.89205161])
```

Рисунок 3.41 – Параметри моделі рідж регресії

Джерело: Розроблено автором на основі [7]

Побудувавши моделі та визначивши параметри змінних для найкращої з них (рис. 3.41), можемо спостерігати, що змінним «AcceptedCmp1», «AcceptedCmp2», «AcceptedCmp3», «AcceptedCmp4», «AcceptedCmp5», «Complain» та «Response» були присвоєні параметри рівні нулю. Крім того, якщо порівнювати отримані параметри змінних між собою в абсолютних величинах, можемо помітити, що максимальний з них є рівним 3.37, мінімальний – 0.004. Робимо припущення, що наступні змінна Teenhome були визначені моделлю як така, що найменше впливають на цільову змінну.

Отже, можемо попередньо рекомендувати дану модель як одну з оптимальних для здійснення прогнозів показника monetary.

Наступним кроком спробуємо застосувати модель ласо регресії до досліджуваного датасету. Використовуючи вже нормалізовані та розділені на тренувальну та тестову вибірку дані, побудуємо ласо регресію з автоматичним підбором значення α .

Отримуємо коефіцієнт детермінації рівний 0.9303 на тренувальних даних та 0.8865 на тестових даних. Дана модель, як і рідж регресія при найменшому значенні α , має більший коефіцієнт детермінації на тренувальних даних, ніж на тестових, що і було очікувано. Автоматично була вибрана оптимальна модель з

показником alpha рівним 1.0809432430795118e-07. Вибрані моделлю пояснювальні змінні є наступними:

```
lasso.coef_
array([ -0.          , -0.          , 13.54812656,  0.          ,
        -0.          , -0.29424242, -0.          , -0.          ,
        -0.          , -0.          , -0.          ,  0.          ,
         0.          ,  0.          ,  0.          ,  0.          ,
         0.          ,  0.          , -0.97949707,  2.08969895,
       -12.28142641])
```

Рисунок 3.42 – Параметри моделі ласо регресії

Джерело: Розроблено автором на основі [7]

Можемо спостерігати, що моделлю були вибрані лише такі пояснювальні змінні як Age, Spending, Frequency, Income, Recency (рис.3.42). Найбільший абсолютний вплив на цільову змінну мають Income та Frequency.

Порівнюючи між собою моделі рідж та ласо, можемо зауважити спільні риси в тому, що обидві моделі прирівняли до нуля параметри змінних AcceptedCmp, що говорили про факт здійснення або не здійснення покупки споживачем під час відповідної рекламної кампанії. Головна відмінність отриманих результатів за цими двома моделями полягає в кількості обраних змінних – під час побудови ласо регресії вагомі коефіцієнти були присвоєні п'ятьом змінним, рідж регресії – 14. Обидві моделі визначили високі коефіцієнти для змінних Income, Age, Spending, Frequency та Income.

Продовжимо дослідження побудовою моделі еластичнет. Використовуючи ті ж самі нормалізовані та розділені дані, отримуємо наступні результати (рис. 3.43).

```
elastic_net.coef_
array([-0.          , -0.          ,  9.20574089, -0.          , -0.          ,
       -0.7590223 , -0.32094154, -0.          , -0.37864155, -0.          ,
         0.          ,  0.          ,  0.          ,  0.          ,  0.          ,
         0.          ,  0.          ,  0.          , -2.35190525,  1.72984624,
       -2.55283608])
```

Рисунок 3.43 – Параметри моделі еластичнет

Джерело: Розроблено автором на основі [7]

Як бачимо, моделлю були обрані 7 змінних – Income, Recency, NumDealsPurchases, NumCatalogPurchases, Age, Spending, Frequency. Обрані змінні включають в себе всі ті змінні, що попередньо були обрані ласо регресією, а також не включають такі змінні, що були обрані рідж регресією. Автоматичний вибір оптимальної α визначив її рівною $2.0161825725735095e-07$. Коефіцієнти детермінації на тренувальній та тестовій вибірках рівні 0.8904 та 0.8305 відповідно.

Крім того, взявши до уваги значення α , автоматично вибрані моделями ласо та еластікнет, можемо продовжити дослідження побудовою декількох додаткових рідж регресій зі значеннями α , близькими до обраних попередніми моделями. Отримуємо такі результати (рис. 3.44).

```
for alpha in range(1, 11):
    ridge = RidgeCV(alphas=[alpha/10000000,], cv=5)
    ridge.fit(train_x, train_y)
    rmse_train = rmse_score(ridge, train_x, train_y)
    rmse_test = rmse_score(ridge, test_x, test_y)
    print(alpha/10000000, ": R2 train -", ridge.score(train_x, train_y), ", RMSE:", rmse_train)
    print(alpha/10000000, ": R2 test -", ridge.score(test_x, test_y), ", RMSE:", rmse_test)
```

```
1e-07 : R2 train - 0.9483673718762411 , RMSE: 0.003365479926417021
1e-07 : R2 test - 0.8976937596116079 , RMSE: 0.0045808444847681
2e-07 : R2 train - 0.9476808687323776 , RMSE: 0.0033877796185373898
2e-07 : R2 test - 0.9064217538468738 , RMSE: 0.004381087592075744
3e-07 : R2 train - 0.9470618504123203 , RMSE: 0.003407762089825355
3e-07 : R2 test - 0.9108051857731712 , RMSE: 0.004277246584443073
4e-07 : R2 train - 0.946531210085897 , RMSE: 0.003424798830643114
4e-07 : R2 test - 0.9132505130435448 , RMSE: 0.004218207534054572
5e-07 : R2 train - 0.9460753859237031 , RMSE: 0.003439366142151502
5e-07 : R2 test - 0.9146811509479664 , RMSE: 0.004183280437199639
6e-07 : R2 train - 0.9456785861097466 , RMSE: 0.003451997094472908
6e-07 : R2 test - 0.9155219837658872 , RMSE: 0.004162615902118338
7e-07 : R2 train - 0.9453279137222264 , RMSE: 0.0034631213696522506
7e-07 : R2 test - 0.915993984177366 , RMSE: 0.004150970810294745
8e-07 : R2 train - 0.9450135670139834 , RMSE: 0.0034730630103998847
8e-07 : R2 test - 0.9162227440358536 , RMSE: 0.004145315127046995
9e-07 : R2 train - 0.9447282048653244 , RMSE: 0.0034820633960082614
9e-07 : R2 test - 0.9162846231956708 , RMSE: 0.004143783947748495
1e-06 : R2 train - 0.9444663107775194 , RMSE: 0.0034903031692571046
1e-06 : R2 test - 0.9162285694865437 , RMSE: 0.004145171002317718
```

Рисунок 3.44 – Коефіцієнти детермінації моделей рідж регресії з різними показниками α в межах $1e-07$ – $1e-08$

Джерело: Розроблено автором на основі [7]

За новими розрахунками ми і отримуємо кращі значення коефіцієнта детермінації як на тренувальних даних, так і на тестових даних. Орієнтуватимемось на найкращий за тестовими даними та визначимо модель з $\alpha=9e-7$ як найбільш оптимальну.

```
ridge.coef_
```

```
array([-6.08904731e+00, -5.97823236e+00,  1.72605102e+01,  2.01271401e+01,
       -4.21533875e+00, -1.20482627e-02,  7.11519973e+00, -1.70428447e+01,
        2.69130470e+00, -7.04098754e+00, -7.65192871e+00,  0.00000000e+00,
        0.00000000e+00,  0.00000000e+00,  0.00000000e+00,  0.00000000e+00,
        0.00000000e+00,  0.00000000e+00,  1.14733106e+00,  2.26348187e+00,
       -1.42773278e+01])
```

Рисунок 3.45 – Параметри моделі рідж регресії при $\alpha=9e-7$

Джерело: Розроблено автором на основі [7]

Аналізуючи вибрані моделлю коефіцієнти (рис. 3.45), можемо дійти висновку, що були обрані ті самі змінні, що і попередньою ridge моделлю, різниця полягає в обчислених параметрах.

Обрахувавши RMSE для всіх трьох моделей, можемо скласти наступну порівняльну таблицю 3.3.

Таблиця 3.3 – Оцінка якості моделей рідж регресії, ласо регресії та еластичнет за показниками коефіцієнту детермінації та RMSE

Модель	alpha	Кількість включених змінних	Коефіцієнт детермінації		RMSE	
			train	test	train	test
ridge	9e-7	14	0.9447	0.9162	0.0034	0.0041
lasso	1.0809e-07	5	0.9303	0.8865	0.0039	0.0048
elasticnet	2.0161e-07	7	0.8904	0.8305	0.0049	0.0058

Джерело: Розроблено автором на основі [7]

За наведеною таблицею можемо зробити наступні висновки. Найбільше значення коефіцієнту детермінації та середньоквадратичної похибки на тестовій вибірці має рідж регресія з показником α рівним $9e-7$. Кількість змінних в даній моделі, коефіцієнти при яких не наближаються до нуля, сягає 14. Прирівняними до нуля були всі змінні AcceptedCmp та змінна Response, що позначали позитивну чи нейтральну реакцію споживача на відповідну рекламну кампанію, а також змінна

Complain, що відображає чи була подана покупцем скарга за весь час здійснення покупок.

Другою за результатами обчислень є модель ласо регресії з показником α рівним $1.0809e-07$, що визначила з-поміж всіх змінних 5 як найбільш значущі для цільової змінної. Показники коефіцієнта детермінації та середньоквадратичної помилки відрізняються від показників рідж регресії не на велику величину в абсолютному вимірі.

Модель еластікнет з показником α рівним $2.0161e-07$, показала гірші результати за інші дві моделі, однак також на незначну величину.

Модель рідж регресії з $\alpha=9e-7$ вважатимемо найбільш оптимальної для даного датасету.

Отже, на отриманих даних можемо зробити висновок, що лояльність споживачів дещо знизилась протягом останнього короткострокового періоду, однак більшість покупців можна класифікувати як лояльних.

ВИСНОВКИ

Споживацька активність є складним процесом, що включає в себе виявлення потреб споживача, збирання даних про товари, ознайомлення з варіантами, прийняття рішення про покупку та подальшу оцінку товару/послуги. Впродовж цих етапів споживачі приймають рішення залежно від особистих поглядів та цінностей, соціальних звичаїв, маркетингових кампаній, специфічних особливостей продукту, і навіть умов навколишнього середовища.

Сегментація цільової аудиторії має велике значення для бізнесу, адже допомагає в процесі створення нових продуктів під конкретні умови задоволення потреб, коригування існуючих товарів, а також розвитку маркетингових кампаній задля розширення кількості нових споживачів та заохочення до більшої активності вже «знайомих» клієнтів.

За результатами аналізу споживацької поведінки українців, проведеним компанією Deloitte, можемо зробити висновки, що більшість українців змінила свої споживацькі звички щодо частоти купівлі порівняно з 2021 роком – 60%, та 71% українців вважають, що змінилися їх споживацькі звички щодо витрат.

Аналізуючи показник частоти, можемо дійти висновку, що споживацька активність українців знизилась. Найбільшу різницю демонструє категорія відвідування закладів громадського харчування, найменшу – дитячі товари. В свою чергу, рівень витрат українців також зменшився в усіх досліджуваних категоріях, найбільші зміни зазнали косметичні товари та будівельні матеріали.

На основі аналізу маркетингових показників Google Merchandise Store засобами Google Analytics 4, можемо зробити припущення, що кількість додавання товарів в кошик та переглядів товарів є переважно сталими величинами, однак за проведення рекламних кампаній та додаткового заохочення користувачів до купівлі можна отримати відповідні показники популярності товару в 4-5 разі вищі за звичайні.

В ході моделювання, а саме застосування методу головних компонент, було виявлено 3 основні компоненти. До першої з них можемо віднести такі змінні, як «MntFruits», «MntMeatProducts», «MntFishProducts», «MntSweetProducts» та «Income». Дана головна компонента була названа «Витрати та доходи», адже перелічені змінні, які є найбільш корельованими з першою основною компонентною, всі відображають напрямки витрат споживачів. Другій основній компоненті дали назву «Освіта» за змінною, яка найбільше корелює з даною компонентною, інші змінні мають незначні ступені кореляції. Третя основна компонента була визначена як «Кількість покупок», адже найбільший зв'язок вона має зі змінними «NumDealsPurchases» та «NumWebPurchases».

В результаті проведення кластерного аналізу з використання двох методів моделювання – К-середніх та ієрархічної кластеризації – можемо дійти до висновку, що найбільш оптимальною кількістю кластерів за показниками силуєту, індексу Девіса-Болдіна та індексу Калінскі-Харабаша є 3 кластери.

При розподілі спостережень за віссю координат Вік-Дохід, можемо зробити наступні висновки про групи споживачів:

- перша група має дохід переважно в рамках 10 тис. дол. – 45 тис. дол., і є більш молодого порівняно з іншими споживачами – 30-55 років;
- друга група має дохід близько 45-70 тис. дол. та віком сягає від 45 до 70 років;
- третя група споживачів має найбільший дохід відносно інших груп – 60 тис. дол. – 90 тис. дол. та за віком є розподіленою переважно в рамках 40 – 75 років.

Три групи споживачів також відрізняються за значенням витрат:

- перша група споживачів витрачає переважно 0-250\$
- друга група – 250-750\$
- третя група – 750-2000\$

Однак кількість днів з останньої покупки не є різною для всіх трьох груп, тому робимо висновок, що покупці, які витрачають більше коштів, не здійснюють покупки частіше, а скоріше мають тенденцію до придбання товарів з більш

високим чеком. Даний висновок може бути корисним при побудові наступної маркетингової кампанії, або при оперативному коригуванні поточної.

За результатами регресійного аналізу можемо зробити наступні висновки. Найбільше значення коефіцієнту детермінації та середньоквадратичної похибки на тестовій вибірці має рідж регресія з показником α рівним $9e-7$. Кількість змінних в даній моделі, коефіцієнти при яких не наближаються до нуля, сягає 14. Прирівняними до нуля були всі змінні `AcceptedCmp` та змінна `Response`, що позначали позитивну чи нейтральну реакцію споживача на відповідну рекламну кампанію, а також змінна `Complain`, що відображає чи була подана покупцем скарга за весь час здійснення покупок.

Другою за результатами обчислень є модель ласо регресії з показником α рівним $1.0809e-07$, що визначила з-поміж всіх змінних п'ять найбільш значущих для цільової змінної. Показники коефіцієнта детермінації та середньоквадратичної помилки відрізняються від показників рідж регресії на незначну величину в абсолютному вимірі.

Модель еластікнет з показником α рівним $2.0161e-07$, показала гірші результати за інші дві моделі.

Отже, модель рідж регресії з $\alpha=9e-7$ вважатимемо оптимальною для даного датасету.

Розроблена в продовж дослідження система моделей може мати практичне застосування в маркетинговій діяльності підприємств для аналізу та прогнозування активності конкретної цільової аудиторії, вибору відповідних специфіці роботи підприємства факторів впливу на поведінку споживачів. Крім побудованих моделей можуть також бути використані програмні методи та алгоритми, застосовані в роботі. Перспективою подальшого дослідження може бути застосування розроблених моделей на даних конкретного підприємства та побудові відповідної маркетингової стратегії на основі отриманих результатів. Крім того, результати роботи можуть слугувати основою для подальших досліджень активності споживачів в контексті її зміни з часом і під впливом глобальних економічних процесів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Скригун, Н. П. Планування комунікаційного міксу як складової маркетингу / Н. П. Скригун, С. Б. Розумей, К. Ю. Семененко // Науковий журнал. – 2013. - № 2. – С. 301-307.
2. Кучма І.З. Еволюція дослідження споживацької поведінки: соціологічний підхід / І.З. Кучма // Соціальні технології: актуальні проблеми теорії та практики: збірник наукових праць. 2012. Вип. 53. С.253-255.
3. Radu V. Consumer behavior in marketing – patterns, types, segmentation [Digital resource] / V. Radu. – Access mode: <https://www.omniconvert.com/blog/consumer-behavior-in-marketing-patterns-types-segmentation/>
4. Споживацькі настрої українців у 2020 році [Електронний ресурс] / Deloitte. Київ: 2021. – Режим доступу: <https://www2.deloitte.com/content/dam/Deloitte/ua/Documents/Press-release/RWD%202020%20UA.pdf>
5. Споживацькі настрої українців після повномасштабного вторгнення [Електронний ресурс] / Deloitte. Київ: 2021. – Режим доступу: <https://www2.deloitte.com/content/dam/Deloitte/ua/Documents/consumer-business/Deloitte-Survey-2022-Consumer-sentiment-during-the-war.pdf>
6. Khajvand M. Estimating customer lifetime value based on RFM analysis of customer purchase behavior: case study [Digital resource] / M. Khajvand, K. Zolfaghar, S. Ashoori, S. Alizadeh. – Access mode: <https://www.sciencedirect.com/science/article/pii/S1877050910003868>
7. Customer Personality Analysis [Digital resource] / V. Radu. – Access mode: <https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data>
8. GA4 – Merchandise Store [Digital resource] / V. Radu. – Access mode: https://analytics.google.com/analytics/web/?utm_source=demoaccount#/p213025502/reports/reportinghub?params=_u..nav%3Dmaui&collectionId=6384423676

9. Бутенко Д. Використання воронки продаж для підвищення ефективності інтернет-торгівлі / Д. Бутенко, К. Немашко // ХНЕУ. – 2020. - № 2. – С. 301-302.
10. Методи аналізу даних: навчальний посібник для студентів / В.Є. Бахрушин. – Запоріжжя: КПУ, 2011. – 268 с.
11. Steven M. Holand – Principal component analysis (PCA) [Digital resource] / S. M. Holand. // Department of Geology, University of Georgia. – 2019. – Access mode: <http://stratigrafia.org/8370/handouts/pcaTutorial.pdf>
12. K-means clustering [Digital resource] / Geeksforgeeks – Access mode: <https://www.geeksforgeeks.org/k-means-clustering-introduction/>
13. Настенко, Є. А. Метод нечітких k-середніх з обмеженою масою робочої області формування кластерів довільної форми / Настенко Є. А., Уманець В. С. // Біомедична інженерія і технологія. – 2018. – № 1(1). – С. 65–70.
14. Ткаченко О.М. Метод кластеризації на основі послідовного запуску k-середніх з удосконаленням вибором кандидата на нову позицію вставки / О. М. Ткаченко, О. Ф. Грійо-Тукало, О. В. Дзісь, С. М. Лаховець // Наукові праці Вінницького національного технічного університету. – 2012. - №2. – С.323-331.
15. Стех Ю. В. Алгоритмічна оцінка оптимальності результатів кластеризації за критерієм відстані / Ю. В. Стех, Файсал М. Е. Сардіх, А. Б. Керницький, М. С. Домброва // Вісник Національного університету "Львівська політехніка". - Львів : Львів. політехніка, 2010. - № 685: Комп'ютерні системи. Проектування. Теорія і практика. - С. 131-134.
16. Індекс Девіса-Болдіна в кластерному аналізі [Електронний ресурс] // craftapppmobile. – 2018. – Режим доступу до ресурсу: <https://craftapppmobile.com/%D0%B8%D0%BD%D0%B4%D0%B5%D0%BA%D1%81-%D0%B4%D1%8D%D0%B2%D0%B8%D1%81%D0%B0-%D0%B1%D0%BE%D0%BB%D0%B4%D0%B8%D0%BD%D0%B0-%D0%B2->

[%D0%BA%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D0%BE%D0%BC-%D0%B0%D0%BD/](#).

17. Міркін Б. Г. Методи кластерного аналізу для прийняття рішень / Б. Г. Міркін; Національний дослідницький університет «Вища школа економіки»: 2011. – 88 с.
18. Trzciński M, Kowalski PA, Łukasik S. Clustering with Nature-Inspired Algorithm Based on Territorial Behavior of Predatory Animals. *Algorithms*. 2022; 15(2):43. <https://doi.org/10.3390/a15020043>
19. Машинне навчання і Data Science: Гребнева регресія [Електронний ресурс] // mql5. – 2023. – Режим доступу до ресурсу: <https://www.mql5.com/ru/articles/11735>.
20. Towards Data Science Anuja Nagpal. L1 and L2 Regularization Methods [Електронний ресурс] / Towards Data Science Anuja Nagpal // towardsdatascience. – 2017. – Режим доступу до ресурсу: <https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c>.
21. Aarshay Jain. Ridge and Lasso Regression in Python | Complete Tutorial [Електронний ресурс] / Aarshay Jain // analyticsvidhya. – 2016. – Режим доступу до ресурсу: [https://www.analyticsvidhya.com/blog/2016/01/ridge-lasso-regression-python-complete-tutorial/#:~:text=How%20Does%20Ridge%20Regression%20Work%3F,-As%20mentioned%20before&text=Here%2C%20%CE%B1%20\(alpha\)%20is,sum%20of%20squares%20of%20coefficients](https://www.analyticsvidhya.com/blog/2016/01/ridge-lasso-regression-python-complete-tutorial/#:~:text=How%20Does%20Ridge%20Regression%20Work%3F,-As%20mentioned%20before&text=Here%2C%20%CE%B1%20(alpha)%20is,sum%20of%20squares%20of%20coefficients).
22. Hastie, Trevor,, et al. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York, Springer, 2009.
23. Shruti Dhumne. Elastic Net Regression detailed guide [Електронний ресурс] / Shruti Dhumne // medium. – 2023. – Режим доступу до ресурсу: <https://medium.com/@shruti.dhumne/elastic-net-regression-detailed-guide-99dce30b8e6e>.

ДОДАТКИ

Додаток А

Програмні методи та алгоритми, використані в ході дослідження

```
import pandas as pd
import datetime as dt
# from sklearn.preprocessing import MinMaxScaler
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
pd.set_option('display.float_format', lambda x: '%.5f' % x)
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import missingno as msno
# import warnings
from sklearn.exceptions import ConvergenceWarning
from sklearn.preprocessing import LabelEncoder, RobustScaler, normalize
from sklearn.cluster import KMeans
from yellowbrick.cluster import KElbowVisualizer
import plotly.express as px
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
from sklearn.linear_model import RidgeCV, LassoCV, ElasticNetCV

df_ = pd.read_csv("marketing_campaign.csv", sep='\t')
df = df_.copy()
df.head(15)
df = df.drop(["Z_CostContact", "Z_Revenue", "Dt_Customer"], axis=1)
df.shape
df.dtypes
df.isnull().sum()

msno.matrix(df, color=(1, 0.38, 0.27))
plt.title("Null Value graph.", fontsize=20, weight="bold")
plt.show()
df.describe().T

income_by_education = df[['Income',
'Education']].groupby('Education')['Income'].mean()
income_by_education

income_by_education = income_by_education.to_dict()
df['Income'].fillna(df['Education'].map(income_by_education), inplace=True)
df.isnull().sum()

def outlier_thresholds(dataframe: pd.DataFrame, variable: str) -> tuple:
    quartile1 = dataframe[variable].quantile(0.25)
    quartile3 = dataframe[variable].quantile(0.75)
    interquartile_range = quartile3 - quartile1
```

```

up_limit = quartile3 + 1.5 * interquartile_range
low_limit = quartile1 - 1.5 * interquartile_range
return (low_limit, up_limit)

def replace_outliers(dataframe: pd.DataFrame, variable: str) -> None:
    low_limit, up_limit = outlier_thresholds(dataframe, variable)
    df[variable] = np.where(df[variable] > up_limit, up_limit, df[variable])
    df[variable] = np.where(df[variable] < low_limit, low_limit, df[variable])

df["Age"] = 2023-df["Year_Birth"]
df = df.drop(["Year_Birth"], axis=1)

fig, ax = plt.subplots(1, 2, figsize = (10,5))
sns.histplot(ax=ax[0], data=df, x="Age", color="sandybrown")
sns.boxplot(ax=ax[1], data=df, y="Age", color="salmon")
fig.suptitle('Age Distribution', weight="bold")
plt.show()

replace_outliers(df, "Age")

fig, ax = plt.subplots(1, 2, figsize = (10,5))
sns.histplot(ax=ax[0], data=df, x="Age", color="darkorange")
sns.boxplot(ax=ax[1], data=df, y="Age", color="tomato")
fig.suptitle('Age Distribution', weight="bold")
plt.show()

fig, ax = plt.subplots(1, 2, figsize = (10,5))
sns.histplot(ax=ax[0], data=df, x="Income", color="mediumaquamarine")
sns.boxplot(ax=ax[1], data=df, y="Income", color="cornflowerblue")
fig.suptitle('Income Distribution', weight="bold")
plt.show()

for column in df.columns:
    if df[column].dtype == "int64" or df[column].dtype == "float64":
        replace_outliers(df, column)

fig, ax = plt.subplots(1, 2, figsize = (10,5))
sns.histplot(ax=ax[0], data=df, x="Income", color="mediumseagreen")
sns.boxplot(ax=ax[1], data=df, y="Income", color="dodgerblue")
fig.suptitle('Income Distribution', weight="bold")
plt.show()

df["Education"].unique()

df["Marital_Status"].unique()

df_for_regress = df.copy()

label_encoding = LabelEncoder()

for column in df.columns:

```

```

if df[column].dtype == "object":
    df[column] = label_encoding.fit_transform(df[column])

df.head()

## sclaing
rob_scaling = RobustScaler() ## Create rob_scaling object
rob_scaling.fit(df) ## Compute the median and quantiles to be used for
scaling.

scaled_data = pd.DataFrame(rob_scaling.transform(df), columns= df.columns)
scaled_data = scaled_data.drop(["ID"], axis=1)

scaled_data.head(10)

scaled_data.isna().sum()

deco_pca = PCA(n_components=None)
deco_pca.fit(scaled_data)
data_pca = deco_pca.transform(scaled_data)

cumsum_pca = np.cumsum(deco_pca.explained_variance_ratio_)
plt.plot(cumsum_pca, linestyle='-', marker='o', color="k")
plt.title('Cumulative Sum of Explained Variance Ratio')
plt.xlabel('Principal Components')
plt.ylabel('Cumulative Explained Variance')
plt.show()

deco_pca_11 = PCA(n_components=11)
deco_pca_11.fit(scaled_data)
deco_pca_11.transform(scaled_data)

len(deco_pca_11.explained_variance_ratio_),
sum(deco_pca_11.explained_variance_ratio_)

print(deco_pca_11.explained_variance_ratio_)

fig, ax = plt.subplots(nrows=1, ncols=1, figsize=(6, 4))
ax.bar(
    x      = np.arange(deco_pca_11.n_components_) + 1,
    height = deco_pca_11.explained_variance_ratio_,
    color="royalblue",
)

sil_score = []
davies = []
for k in range(2, 6):
    agg_cluster = KMeans(n_clusters=k, random_state=0, n_init=50)
    agg_cluster.fit(PCA_dataset)
    agg_labels = KMeans(n_clusters=k).fit_predict(PCA_dataset)
    sil_score.append(silhouette_score(PCA_dataset, agg_labels))

```

```

    davies.append(davies_bouldin_score(PCA_dataset, agg_labels))

    print(f'{k=}, Silhouette score: ', silhouette_score(PCA_dataset, agg_labels))
    print(f'{k=}, Davies-Bouldin Score: ', davies_bouldin_score(PCA_dataset,
agg_labels))
    print(f'{k=}, calinski_harabasz_score: ',
calinski_harabasz_score(PCA_dataset, agg_labels))

for alpha in range(1, 100, 5):
    ridge = RidgeCV(alphas=[alpha/100,], cv=5)
    ridge.fit(train_x, train_y)
    rmse = rmse_score(ridge, test_x, test_y)
    print(alpha, ": R2 -", ridge.score(test_x, test_y), ", RMSE:", rmse)

ridge = RidgeCV(alphas=[alpha/100 for alpha in range(1, 100, 5)], cv=5)
ridge.fit(train_x, train_y)
ridge.score(test_x, test_y)

lasso = LassoCV(cv=5)
lasso.fit(train_x, train_y)
lasso.score(test_x, test_y)

elastic_net = ElasticNetCV(cv=5, l1_ratio=0.5)
elastic_net.fit(train_x, train_y)
elastic_net.score(test_x, test_y)

```