

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ТЕЛЕКОМУНІКАЦІЙ
КАФЕДРА СИСТЕМНОГО АНАЛІЗУ

КВАЛІФІКАЦІЙНА РОБОТА

на тему: «ВИКОРИСТАННЯ АНАЛІТИКИ ДАНИХ ДЛЯ
ПРОГНОЗУВАННЯ СПОЖИВЧОГО ПОПИТУ
НА ТОВАРИ ТА ПОСЛУГИ»

на здобуття освітнього ступеня бакалавра

зі спеціальності 124 Системний аналіз
(код, найменування спеціальності)

освітньо-професійної програми Системний аналіз
(назва)

Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

(підпис) Артем ФІЛІППОВ
Ім'я, ПРІЗВИЩЕ здобувача

Виконав: здобувач вищої освіти гр. САД-41
Артем ФІЛІППОВ
Ім'я, ПРІЗВИЩЕ

Керівник: Сергій КОЗАЧЕНКО
ст викладач каф. СА. Ім'я, ПРІЗВИЩЕ

Рецензент: _____
науковий ступінь, Ім'я, ПРІЗВИЩЕ
вчене звання

Київ 2024

**ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ТЕЛЕКОМУНІКАЦІЙ**

Кафедра Системного аналізу

Ступінь вищої освіти Бакалавр

Спеціальність 124 Системний аналіз

Освітньо-професійна програма Системний аналіз

ЗАТВЕРДЖУЮ

Завідувач кафедри Ровіл НАФСЄВ

Ровіл НАФСЄВ

« ____ » _____ 2024 р.

**ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ**

Філіппов Артем Олексійович

(прізвище, ім'я, по батькові здобувача)

1. Тема кваліфікаційної роботи: Використання аналітики даних для прогнозування
споживчого попиту на товари та послуги

керівник кваліфікаційної роботи Сергій КОЗАЧЕНКО, к.е.н., доцент
(Ім'я, ПРІЗВИЩЕ, науковий ступінь, вчене звання)

затверджені наказом Державного університету інформаційно-комунікаційних технологій
від « ____ » _____ 20__ р. № ____

2. Строк подання кваліфікаційної роботи « ____ » _____ 20__ р.

3. Вихідні дані до кваліфікаційної роботи: інформаційно-аналітичний метод,
інтернет-ресурси стосовно аналітики аних для прогнозування споживчого попиту на товари
та послуги
Науково-технічна література з питань, пов'язаних з аналітикою даних

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)

1. Методи аналітики даних для прогнозування споживчого попиту
2. Аналіз поведінки клієнтів за допомогою системи аналітики даних
3. Методологія аналізу даних

5. Перелік ілюстративного матеріалу: *презентація*

6. Дата видачі завдання « ____ » _____ 20__ р.

КАЛЕНДАРНИЙ ПЛАН

№ № з/п	Назва етапів кваліфікаційної роботи	Строк вико- нання етапів ро- боти	Примі- тка
1	Підбір науково-технічної літератури		
2	Методи аналітики даних для прогнозу- вання споживчого попиту		
3	Аналіз поведінки клієнтів за допомогою системи аналітики даних		
4	Методологія аналізу даних		
5	Вступ, висновки, реферат		
6	Розробка обов'язкових матеріалів		
7	Попередній захист роботи		

Здобувач вищої освіти

(підпис)

Артем ФІЛІППОВ

(Ім'я, ПРІЗВИЩЕ)

Керівник
кваліфікаційної роботи

(підпис)

Сергій КОЗАЧЕНКО

(Ім'я, ПРІЗВИЩЕ)

**ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ТЕЛЕКОМУНІКАЦІЙ**

**ПОДАННЯ
ГОЛОВІ ЕКЗАМЕНАЦІЙНОЇ КОМІСІЇ
ЩОДО ЗАХИСТУ КВАЛІФІКАЦІЙНОЇ РОБОТИ
на здобуття освітнього ступеня бакалавра**

Направляється здобувач Філіппов А.О. до захисту кваліфікаційної роботи
(*прізвище та ініціали*)
за спеціальністю 124 Системний аналіз
(*код, найменування спеціальності*)

освітньо-професійної програми Системний аналіз
(*назва*)

на тему: «Використання аналітики даних для прогнозування споживчого по-
питу на товари та послуги».

Кваліфікаційна робота і рецензія додаються.

Директор ННІТ

(*підпис*)

Владислав КРАВЧЕНКО

(*Ім'я, ПРІЗВИЩЕ*)

Висновок керівника кваліфікаційної роботи

Здобувач Філіппов Артем Олексійович під час написання кваліфікаційної роботи показав гарну теоретичну підготовку, володіння необхідними знаннями з системного аналізу, користуватися навчальною, довідковою і науково-технічною літературою. Працюючи над завданнями, які доручались керівником, проявив ініціативність, сумлінність та хист до інженерної роботи.

Все це дозволяє оцінити виконану кваліфікаційну роботу здобувача Філіппова Артема Олексійовича оцінку ««відмінно» та присвоїти йому кваліфікацію «Бакалавр з системного аналізу».

Керівник кваліфікаційної роботи _____

(*підпис*)

Сергій КОЗАЧЕНКО

(*Ім'я, ПРІЗВИЩЕ*)

«_____» ____ 20 ____ року

Висновок кафедри про кваліфікаційну роботу

Кваліфікаційна робота розглянута. Здобувач Філіппов А. О. допускається до захисту даної роботи в Експертній комісії.

Завідувач кафедри СА
(*назва*)

(*підпис*)

Ровіл НАФЄСВ

(*Ім'я, ПРІЗВИЩЕ*)

«_____» ____ 20 ____ року

ВІДГУК РЕЦЕНЗЕНТА

на кваліфікаційну бакалаврську роботу

здобувача вищої освіти Філіппова Артема Олексійовича
на тему «Використання аналітики даних для прогнозування споживчого попиту на товари та послуги»

Актуальність.

Аналітика даних стає все більш важливою для підприємств у сучасному світі. Прогнозування споживчого попиту на товари та послуги дозволяє підприємствам оптимально планувати виробництво, запаси, маркетингові кампанії та інші стратегічні рішення. Оскільки в умовах постійних змін і конкуренції на ринку підприємствам потрібно швидко реагувати на зміни в споживчому попиті, щоб забезпечити стабільний розвиток бізнесу, тема кваліфікаційної роботи є досить актуальною.

Позитивні сторони.

1. Докладно описано методи аналітики даних для прогнозування споживчого попиту.
2. Зосереджено увагу на прогнозуванні поведінки клієнтів за допомогою системи аналітики даних.
3. Представлено ключові деталі методології, яка використовується для аналізу даних.

Недоліки.

1. В роботі мають місце деякі неточності в перекладі технічних термінів.
2. В роботі не завадило би надати більше розрахунків щодо прогнозування споживчого попиту.

Відзначені зауваження не впливають на загальну позитивну оцінку кваліфікаційної роботи

Висновок: *кваліфікаційна робота на здобуття ступеня магістра заслуговує оцінку "_____", а здобувач Філіппов Артем Олексійович повністю заслуговує присвоєння кваліфікації: «Бакалавр з системного аналізу»*

Рецензент:
науковий ступінь, вчене звання

_____ підпис

_____ Ім'я, ПРІЗВИЩЕ

РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття освітнього ступеня бакалавра: Фідіппова Артема 59 стор., 34 рис., 2 табл., 23 джерел.

Мета роботи – покращення ефективності управління споживчим попитом підприємства шляхом використання сучасних методів аналітики даних.

Об'єкт дослідження – методи аналітики даних для прогнозування споживчого попиту на товари та послуги.

Предмет дослідження – споживчий попит на товари та послуги.

Короткий зміст роботи:

Докладно описано методи аналітики даних для прогнозування споживчого попиту. Зосереджено увагу на прогнозуванні поведінки клієнтів за допомогою системи аналітики даних. Представлено ключові деталі методології, яка використовується для аналізу даних.

КЛЮЧОВІ СЛОВА:

СИСТЕМНИЙ АНАЛІЗ, АНАЛІТИКА ДАНИХ, ПРОГНОЗУВАННЯ СПОЖИВЧОГО ПОПИТУ, МЕТОДИ СИСТЕМНОГО АНАЛІЗУ, МЕТОДОЛОГІЯ АНАЛІТИКИ ДАНИХ, ПОВЕДІНКА КЛІЄНТІВ

ЗМІСТ

ВСТУП.....	8
РОЗДІЛ 1 МЕТОДИ АНАЛІТИКИ ДАНИХ ДЛЯ ПРОГНОЗУВАННЯ СПОЖИВЧОГО ПОПИТУ.....	10
1.1 Процес прийняття рішення про розробку продукту.....	10
1.2 Дані про споживача та споживчий попит	12
1.3 Сутність поняття аналітики даних.....	13
1.4 Аналіз даних, штучний інтелект і машинне навчання.....	14
1.4.1 Визначення поточного клієнта компанії.....	17
1.4.2 Діагностика недоліків поточної ринкової ситуації.....	18
1.4.3 Дедукція потенційних концепцій продукту.....	19
1.4.4 Етап проектування.....	20
1.4.5 Прийняття рішення щодо вибору продукту.....	21
РОЗДІЛ 2 ПРОГНОЗУВАННЯ ПОВЕДІНКИ КЛІЄНТІВ ЗА ДОПОМОГОЮ СИСТЕМИ АНАЛІТИКИ ДАНИХ.....	24
2.1 Аналіз існуючих систем аналітики даних.....	24
2.2 Порядок реалізації запропонованої системи аналітики даних.....	26
2.3 Генерація гіпотези	27
2.4 Імпорт набору даних.....	27
2.5 Попередня обробка даних.....	28
2.6 Вилучення ознак	29
2.7 Застосований алгоритм.....	31
2.7.1 Алгоритм лінійної регресії	32
2.7.2 Алгоритм випадкового лісу.....	33
РОЗДІЛ 3 МЕТОДОЛОГІЯ АНАЛІЗУ ДАНИХ.....	36
3.1 Служба бізнес-аналітики Power BI	36
3.1.1 R-Загальна концепція.....	37
3.1.2 Бібліотеки, що використовуються в мові програмування R	37
3.2 Опис набору даних.....	38
3.3 Попередня обробка даних.....	39
3.4 Аналіз RFM.....	42
3.4.1 Розрахунок значення актуальності.....	42
3.4.2 Розрахунок значення частоти.....	43
3.4.3 Розрахунок грошової вартості.....	43
3.5 Розрахунок балів RFM.....	44
3.6 Пояснення сегментів RFM.....	45
3.7 Результати дослідження та візуалізація.....	50
3.8 Аналіз RFM.....	52
ВИСНОВКИ.....	55
ПЕРЕЛІК ПОСИЛАНЬ.....	57
ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ (Презентація).....	59

ВСТУП

В епоху цифровізації експоненціальне зростання кількості даних суттєво змінює спосіб роботи підприємств. З появою технологій великих даних ритейлери отримали доступ до величезних обсягів даних про моделі поведінки, вподобання та покупки своїх клієнтів. Ці дані можуть бути перетворені на цінну інформацію за допомогою процесу клієнтської аналітики, яка може допомогти ритейлерам покращити свої маркетингові стратегії та сприйняття клієнтів і збільшити свій прибуток.

Традиційні виробники повинні також мати справу з руйнівними змінами в їхньому бізнесі в умовах зростання нестабільності зовнішніх впливів, скорочення життєвих циклів продукту, збільшення глобальної конкуренції та експоненціального технологічного прогресу.

Виклик для цих компаній полягає в роботі з новими даними та застосуванні правильних методів аналізу потреб клієнта в потрібний час у процесі розробки продукту.

Тільки шляхом цілеспрямованого використання правильних методів традиційні виробники можуть продовжити аналіз даних про використання клієнтами продукту.

Отже, новий продукт або пропозиції послуг можна швидко розробити та пристосувати до потреб клієнтів, надаючи підприємствам конкурентну перевагу . рішень.

Таким чином, інновації і рішення можуть бути розроблені для успішних і відповідних для клієнтів продуктів. Мета роботи полягає в тому, щоб продемонструвати, як різні області методології та аналізу можуть вплинути на вибір продукту споживачем.

Аналітика споживчого попиту передбачає використання передових методів інтелектуального аналізу даних для вилучення значущої інформації з величезних обсягів доступних даних. Це дозволяє ритейлерам сегментувати своїх клієнтів на основі їхньої поведінки, уподобань і моделей покупок. Сегментація клієнтів

дозволяє ритейлерам орієнтуватися на певні групи клієнтів за допомогою персоналізованих маркетингових стратегій, адаптованих до їхніх потреб і вподобань.

Сегментація клієнтів має важливе значення для ідентифікації та утримання цінних клієнтів. У той час як залучення нових клієнтів є життєво важливим, утримання існуючих клієнтів є не менш важливим. Втрата клієнта означає втрату довічної цінності покупок, яка може суттєво вплинути на прибуток ритейлера. Таким чином, ідентифікація клієнтів, яка є можливістю та впровадження ефективних стратегій утримання має вирішальне значення для виживання та зростання роздрібного бізнесу.

Ця робота починається з поглибленого пояснення та аналізу ключових термінів, пов'язаних з великими даними, клієнтською аналітикою, інтелектуальним аналізом даних та сегментацією клієнтів, через посилення на численні приклади їх застосування в сфері роздрібної торгівлі. У другому розділі дослідження зосереджено та представлено детальний опис набору даних, який буде використовуватися для аналізу, та інструменти, які можуть полегшати процес аналізу.

Крім різних технік та методів обробки даних, представлено попередню обробку та сегментацію клієнтів за допомогою RFM-аналізу.

1 МЕТОДИ АНАЛІТИКИ ДАНИХ ДЛЯ ПРОГНОЗУВАННЯ СПОЖИВЧОГО ПОПИТУ

1.1 Процес прийняття рішення про розробку продукту

На відміну від популярних уявлень про процес розробки продукту, який являє собою серію дій, які корпорація здійснює для створення, проектування та маркетингу продукту, чіткий опис процесу прийняття рішення про те, який продукт розробити, рідко зустрічається в науковій літературі.

Однак завдяки оцифруванню продуктів компанії стають все більш здатними обробляти величезні обсяги даних про використання клієнтами.

Ці дані про використання відображають поведінку поточного клієнта компанії, яка не обов'язково прирівнюється до цільового клієнта цієї компанії.

Це створює взаємодію між відділами, які можуть захотіти вирішити, які продукти розробляти, виходячи з уявного цільового клієнта, або, які хочуть прийняти рішення на основі поточного клієнта.

Тому поряд з іншими важливими корпоративними функціями в компанії, такими як якість або обслуговування клієнтів, особливо в рамках функції прийняття рішень щодо продукту, слід враховувати дані про споживчий попит.

На відміну від PDP, чисті процеси прийняття рішень були детально проаналізовані в академічних роботах, деякі з них налічують понад 35 років.

В цілому, визначення процесу прийняття рішення, який є сприятливим для успішного прийняття рішення в різних типах організаційних структур, пов'язані з закономірностями в процесах прийняття рішень

Стверджується, що процес прийняття рішень щодо продуктів повинен бути зосереджений на повноті, що означає, що повинен бути описаний кожен необхідний крок. Крім того, він повинен бути зосереджений на якості, що означає, що виконання етапів процесу є якісним. Зрештою, такий процес має бути

зосереджений на важливому, наприклад, на усуненні слабких сторін продукту. Тому було проаналізовано наукові підходи до процесів прийняття рішень в цілому і зроблено перехід до процесу прийняття рішень щодо продуктів.

До початку процесу розробки продукту починається PDP і має на меті визначення характеристик поточного клієнта, діагностування його потреби в продукті, на основі чого слід розробити відповідний проект.

Таким чином, було визначено загальний процес, що складається з наступних етапів:

- визначення;
- діагностика;
- вилучення;
- проектування;
- прийняття рішення.

Виконання всіх кроків може бути виконано кваліфіковано, оскільки вони характеризуються потенціалом збору даних про використання клієнтів на кожному з цих п'яти кроків та їх аналізу, щоб зробити висновки про основну поведінку клієнтів.

Врешті-решт, процес зосереджується на важливому – усуненні слабких місць поточних продуктів, що розкривається аналізом даних про використання клієнтами.

Насправді більш широкий процес прийняття рішень відбувається до конкретного процесу розробки. Іноді обидва процеси також можуть перетинатися, але рішення про категорію, форму або специфікацію продукту, який є бажаним для клієнта, має початися до того, як почнеться будь-яка розробка.

Насправді, цей процес навіть не обов'язково повинен бути виділений в дослідженнях і розробках, але також може здійснюватися в маркетингу або корпоративній стратегії..

1.2 Дані про споживача та споживчий попит

Дані про клієнта або споживача визначаються як будь-який запис особистої, поведінкової, психографічної та демографічної інформації споживачів, зібраної компанією. Можна припустити випадок, коли компанія збирає тільки дані від споживачів інструментів або продуктів, які компанія надає безпосередньо споживачеві, який тим самим стає клієнтом.

Випадок, коли компанія купує дані про споживачів, які можуть бути потенційними клієнтами у третьої сторони. Різниця між цими двома випадками частково полягає в меті аналізу компанії, тобто у виявленні основної поведінки, з якою конкретно клієнти компанії використовують продукцію компанії, порівняно з виявленням потенційних нових клієнтів з групи споживачів.

Отже, дані про використання клієнтів – це запис активності з посиланням на конкретний інструмент або продукт. Таким чином, це є виведенням поведінки клієнта в ситуаціях, коли використовуються обидва продукти, і при запис є технічно можливим.

Таким чином, деякі автори посилаються на поведінкові дані в тому ж контексті, в той час як інші використовують поняття даних про залученість. Це означає, що значуща інформація може бути отримана з фактичних даних споживачів за допомогою усталених статистичних і математичних підходів, а також машинного навчання або штучного інтелекту.

У недавньому минулому наукова розробка методів аналізу даних про використання клієнтами зазнавала стійкого зростання в різних сферах бізнесу, таких як маркетинг або дослідження та розробки. Сфера запису, збору та аналізу даних про використання клієнтами є всеосяжною, і тут можна окреслити багато концепцій.

1.3 Сутність поняття аналітики даних

Термін «аналітика даних» є загальновживаним поняттям, яке буде визначено для цілей цього огляду як багаторівнева, математична та обчислювальна обробка даних. Що стосується впливу аналітики даних на прийняття рішень людиною, то підійде ієрархія описової, прогностичної та прескриптивної аналітики. Описовий аналіз є основним видом аналізу даних про використання клієнтів. Він передбачає узагальнення та опис закономірностей знань за допомогою фундаментальних статистичних методів, таких як середнє значення, медіана, мода, стандартне відхилення, дисперсія та вимірювання частоти певних подій у великих потоках даних.

За своєю суттю предиктивна аналітика полягає у виявленні закономірностей і реєстрації взаємозв'язків у даних. У цьому контексті вона включає прогнози та статистичні моделі, які визначають майбутні події на основі історичних та поточних даних. Крім того, прескриптивна аналітика поєднує прогнози з атрибутами призначення населення.

У рамках PDP, наприклад, минулі потреби клієнта можуть бути визначені за допомогою описової аналітики, тоді як майбутні потреби можуть бути прогнозовані за допомогою предиктивної аналітики. Зрештою, методи з області прескриптивної аналітики можуть бути використані для визначення атрибутів продукту, які, швидше за все, можуть задовільнити майбутні потреби клієнта.

Для цього огляду вкрай важливо відрізнити аналіз даних від штучного інтелекту, оскільки ці дві концепції представляють дві різні групи методологій у концепції аналізу даних. Як описано в наступному підрозділі, аналіз даних – це просто вивчення даних, на відміну від багаторівневої, математичної та обчислювальної обробки даних, яка відбувається на кожному рівні (описовому, прогностичному та прескриптивному) аналізі даних.

Ця диференціація є ключовою, оскільки у випадку PDP усі рівні аналізу даних відіграватимуть важливу роль в аналізі даних про використання клієнтами за

допомогою методів з обох галузей: аналізу даних та штучного інтелекту.

На рис. 1.1 представлена таксономія аналізу даних з трьома рівнями: описовим, прогностичним і прескриптивним, з використанням методів з обох областей: аналізу даних і штучного інтелекту.

Першочергове значення має визначення того, які рівні аналізу даних можуть бути застосовані до яких етапів PDP, оскільки це відображення виявить прогалину в сучасних академічних дослідженнях.

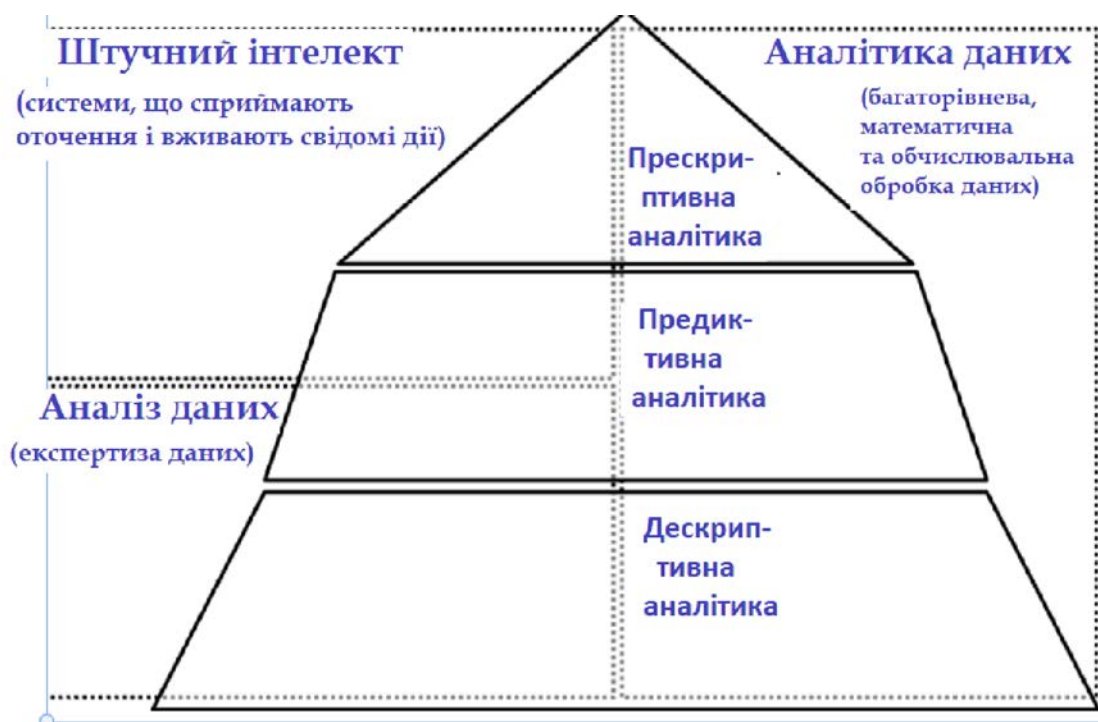


Рисунок 1.1 – Таксономія аналізу даних з трьома рівнями: описовим, прогностичним і прескриптивним

1.4 Аналіз даних, штучний інтелект і машинне навчання

В описовій статистиці дані аналізуються шляхом характеристики або опису спостережуваних даних, які можуть бути отримані або не отримані з більшої сукупності. На противагу цьому, індуктивна статистика робить висновок про припущення вибірки щодо більшої сукупності. Спостережувані дані, які є вибіркою з більшої генеральної сукупності, є невід'ємним аспектом індуктивної статистики.

Індуктивна статистика використовується для отримання ймовірності майбутньої події, а методи з цієї області мають справу з оцінками та узагальненнями щодо вибіркової сукупності. Описова та індуктивна статистики іноді перетинаються в аналітиці даних, наприклад, в розвідувальному аналізі даних. Як характеристики спостережуваних даних, так і передбачення з вивченої моделі можуть бути проаналізовані або досліджені, наприклад, за допомогою візуалізації даних. ML часто використовується щоразу, коли вивчена статистична модель (рис. 1.2) використовується обчислювально для висновків. Це підмножина ШІ, яка визначається як система, яка сприймає навколишнє середовище та здійснює навмисні дії. Як правило, машинне навчання відноситься до будь-якої діяльності, яку машина виконує без явного програмування, і робить це за допомогою алгоритмів для ітеративного навчання або навчання статистичної моделі. Контрольоване навчання відноситься до висновку з певної вибіркової форми (рис. 1.2), яка містить як вхідні, так і бажані результати. Зіставлення конкретно призначених характеристик поведінки споживачів з конкретними характеристиками використання є багатим ґрунтом для контрольованого навчання і називається «міченими навчальними даними». На відміну від навчання з учителем, у сфері навчання без учителя не пропонуються вихідні характеристики, де правила та заздалегідь визначені цілі визначають правильність передбачення. Штучні нейронні мережі – це широко використовувана архітектура вивченої моделі в аналізі даних та ШІ, що складається з окремих обчислювальних блоків, згрупованих у шари. Зрештою, використання додаткових прихованих шарів, які дозволяють потенційно здійснити перестановку атрибутів з нижчих шарів, сформувало концепцію глибокого навчання, яка відноситься до глибокої архітектури штучної нейронної мережі. Таким чином, глибинне навчання може бути як контрольованим, так і неконтрольованим. На рис.1.2 показані відповідні концепції з області аналізу даних та штучного інтелекту, з ML та контрольованими та неконтрольованими методами навчання як підмножинами. Глибоке навчання уособлює клас в рамках ML і може бути застосоване до обох методів машинного навчання, описаних тут, а ML використовує індуктивну статистику для висновків і

прогнозування генеральної сукупності. Зрештою, ця систематика не відображає всіх характеристик кожного показаного рівня і не має універсальної мети, а слугує компіляцією тих понять, які здаються актуальними.

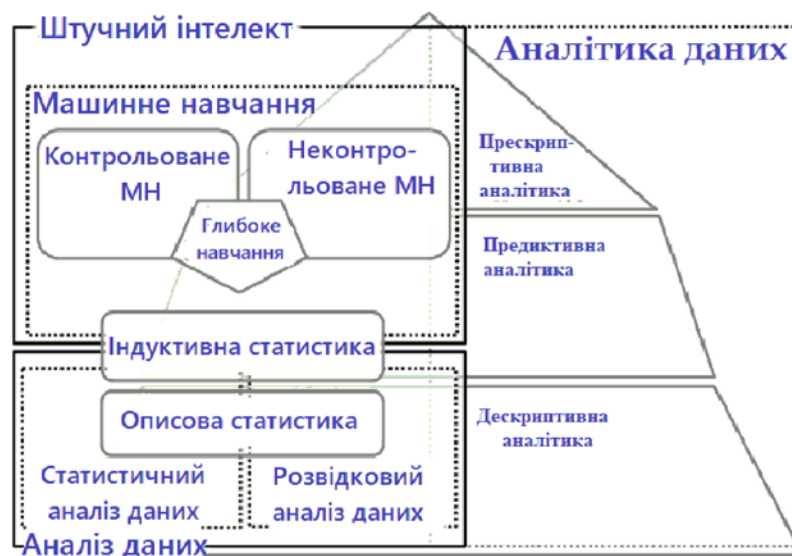


Рисунок 1.2 – Таксономія машинного навчання і аналітики даних

Однак таксономія наголошує на контрольованих і неконтрольованих підходах до машинного навчання, які є корисними у встановленні PDP та розміщенні їх у контексті, скажімо, класу алгоритмів, таких як глибоке навчання.

У цьому підрозділі детально дано опис кожного кроку, описано поточну ситуацію та її виклики, а також те, як зараз приймаються рішення щодо продукту. Описуючи, як різні рівні аналізу даних можуть бути застосовані до кожного кроку, цей розділ слугуватиме спробою заповнити прогалину, виявлену в цьому огляді. Далі кожен крок PDP, як визначено вище, детально описаний і доповнений реальними прикладами та застосуваннями.

При огляді літератури, присвяченої аналітиці даних про використання клієнтами, було виявлено, що більшість авторів використовують методи тільки з одного рівня аналізу даних або поєднують конкретні методи з двох або трьох рівнів в повторюваних шаблонах. Вонг і Вей об'єднали інтелектуальний аналіз даних з описовою аналітики з прогностичною моделлю. Рідко всі три рівні аналітики даних поєднуються; наприклад, Янк, Дьолле і Шух створили систему підтримки прийняття рішень для продуктових портфелів. Нагадаємо, що в процесі створення

продукту PDP вирішує, який продукт розробити, виходячи з визначеного клієнта з діагностованою потребою в продукті, відповідний дизайн якого повинен бути виведений з прототипування. Як визначено вище, процес складається з п'яти етапів. PDP характеризується потенціалом збору даних про використання клієнтів на кожному з цих п'яти кроків та їх аналізу, щоб зробити висновки про основну поведінку клієнтів.

1.4.1 Визначення поточного клієнта компанії

Перший етап процесу PDP передбачає формальне визначення поточного клієнта компанії шляхом аналізу даних про її використання. Матриця була інтенсивною, описова аналітика могла бути використана для представлення поточних клієнтів, описуючи демографічні показники, соціально-економічну конституцію або поведінку під час використання продукту. Предиктивна аналітика може бути використана для формулювання припущень про майбутній розвиток цих факторів. На противагу цьому, прескриптивна аналітика використовується рідше, оскільки в більшості випадків неможливо вплинути на фактори, які визначають клієнта на основі даних про використання.

Проблеми: В даний час визначення потенційних нових клієнтів часто базується на даних споживачів і зазвичай знаменує собою початок управління взаємовідносинами з клієнтами. Однак було виявлено, що дані про використання клієнтів не використовуються широко під час формального визначення поточних клієнтів. Можна побачити проблему множинного, розбіжного визначення клієнтів, що існує паралельно в компанії, наприклад, в маркетингу, з одного боку, і досліджень і розробок, з іншого. Описовий аналіз поведінки існуючих клієнтів, що надходять, наприклад, в результаті досліджень і розробок, може перешкодити відділам маркетингу скласти нереалістичне визначення потенційних нових клієнтів на основі даних про споживачів, яке, наприклад, описує поведінку групи споживачів в Інтернеті. Багатообіцяючі ідеї також можна знайти в поєднанні описової та прогностичної аналітики, наприклад, при прогнозуванні

майбутньої поведінки певної групи клієнтів, як це буде описано нижче.

Яскравий приклад описової та предиктивної аналітики на етапі визначення поточного клієнта можна знайти у сфері сегментації. Вонг і Вей створили інтегровану модель сегментації клієнтів, перш ніж прогнозувати індивідуальні пропозиції послуг. Зокрема, автори проаналізували дані про купівельну поведінку клієнтів у зв'язці з закупівельними цінами онлайн-тур агентств. Система підтримки прийняття рішень використовувала інтелектуальний аналіз даних для отримання даних про ціни та характеристики клієнтів, прогножуючи потенційні подальші покупки користувача.

1.4.2 Діагностика недоліків поточної ринкової ситуації

Під час діагностики недоліки поточної ринкової ситуації будуть виявлені на основі заздалегідь визначеного клієнта. На цьому етапі буде вирішено питання про те, чи задоволені, виконані або перевищені конкретні потреби клієнта. Описова аналітика може бути використана для опису ринкової ситуації, наприклад, щодо сезонності або тенденцій основних часових рядів. Крім того, може бути використана предиктивна аналітика, оскільки фактичні потреби клієнта не можуть бути відомі, але їх потрібно визначити. Подібно до кроку 1, на основні фактори, що лежать в основі кроку 2, у більшості випадків можна не впливати, тому прескриптивні методи аналітики не є дуже перспективними.

Виклики: Діагностика поточної ситуації на ринку є ключовим компонентом існуючих відділів корпоративної стратегії. Аналізуються макроекономічні показники, і навіть показники споживчих настроїв обчислюються методами як описової, так і прогностичної аналітики, наприклад, для прогнозування того, чи відвернуться клієнти від компанії, що називається відтоком, чи ні. Однак рідко зустрічаються приклади застосування цих методів до даних в рамках PDP.

Насправді, використання клієнтами відкриває великий потенціал для етапу діагностики, оскільки можна виявити кореляцію між поведінкою використання та ринковими даними, такими як використання електромобілів під час високих цін на

викопні енергоносії. Недоліки в рівновазі попиту та пропозиції часто аналізуються за допомогою сучасних методів, таких як аналіз настроїв покупців на основі відгуків про товари в Інтернеті. Серед іншого, Айрленд та Лю застосували описову та прогностичну аналітику для аналізу настроїв онлайн-оглядів продуктів. У їхньому дослідженні використовувалися дані, отримані за допомогою аналізу настроїв з онлайн-оглядів продуктів, щоб визначити потреби клієнтів. Автори детально розглядають описову аналітику з використанням методів максимальної ентропії або прогностичної аналітики, таких як класифікація, основана на машині опорного вектора.

1.4.3 Дедукція потенційних концепцій продукту

На третьому кроці PDP виводяться потенційні концепції продукту. Дедукція є найскладнішим етапом PDP, оскільки вона вимагає створення точних рішень передбачуваних проблем, які, за прогнозами, виникнуть у клієнта. При цьому можна використовувати описову, прогностичну та прескриптивну аналітику для отримання потенційних рішень у спосіб, орієнтований на дані. Після описового аналізу, наприклад, існуючих концепцій продукту, може бути змодельований потенційний обсяг нового продукту цієї концепції і, після чого методи прескриптивної аналітики можуть підтримати позиціонування нового продукту в портфелі компанії.

Виклики: Безумовно, перетворення етапу дедукції в процес, орієнтований на дані, було б на сьогоднішній день найскладнішим перенесенням, поряд з простим визначенням ринкових прогалів і, як наслідок, їх заповненням, інноваційні ідеї часто вимагають винахідницького кроку на додаток до цього. Однак нещодавні досягнення, особливо в галузі штучного інтелекту, досягли прогресу в цьому відношенні. Янк, Дьолле та Шух інтегрували всі рівні аналітики для створення системи підтримки прийняття рішень для портфелів продуктів. Примітно, що автори спиралися на свій директивний метод на модель класифікації на основі нейронних мереж. Вони розробили свій попередній конвеєр очищення даних,

використовуючи кореляційний аналіз і прогнозування часових рядів для прогнозування волатильних ринкових потреб.

1.4.4 Етап проектування

Етап проектування PDP має справу виключно з даними, отриманими в результаті використання однієї або декількох концепцій або прототипів. Він не прирівнюється до остаточної специфікації продукту, наприклад, його геометрії або концепції виробництва. Насправді, безумовно, прототипування є ітераційним процесом, що вимагає ретельного аналізу даних про використання прототипу. Однак цей ітеративний процес має основну мету ідеальної побудови прототипу, а не надання даних про використання клієнтів для PDP. На цьому етапі предиктивна аналітика може моделювати використання прототипу клієнтами та порівнювати результати за допомогою описової аналітики. Прескриптивні методи аналітики є менш адекватними на етапі проектування PDP, оскільки прототип вже побудований на цьому етапі, і завдання полягає в тому, щоб використовувати дані, отримані в результаті його тестового використання, а не повторно їх реконструювати.

Виклики: Сьогодні багато рішень щодо продуктів ґрунтуються на аналізі даних прототипів. Таким чином, четвертий крок PDP вже є невід'ємною частиною багатьох компаній і розробників, які детально розробили описові методи аналізу даних про використання прототипів. Однак в останні десятиліття методи аналізу даних були збагачені областю штучного інтелекту, як описано вище. Основна проблема сьогодні полягає в тому, щоб оцінити, де, наприклад, методи, основані на машинному навчанні, можуть бути застосовані на етапі проектування. У прототипуванні значною тенденцією є цифровізація апаратного забезпечення та симуляція тестування. Тому першорядний інтерес представляє, як методи аналізу даних і машинного навчання можуть бути інтегровані в цей четвертий крок PDP. У 2021 році Райхе та Тімінгер розробили технологічну модель для моделювання життєвого циклу продукту різних прототипів. Вони використовували методи описової та прогнозовної аналітики для створення цифрового двійника прототипу.

Широко використовуються методи описової аналітики, такі як частотний аналіз, і методи прогнозової аналітики, такі як прогнозування прийняття або настроїв клієнтів за допомогою нейронних мереж.

1.4.5 Прийняття рішення щодо вибору продукту

Останнім кроком PDP є вибір продукту або одного з раніше розроблених прототипів продукту. Описова аналітика лежить в основі цього рішення на основі даних у PDP. Інформація, зібрана на попередньому кроці, може бути передана замовнику. Методи предиктивної аналітики можуть бути застосовані, наприклад, для повторної оцінки прогнозів потенційного обсягу, отриманих раніше в PDP. Врешті-решт, рішення про продукт може бути підкріплене прескриптивною аналітикою, наприклад, вивченням суми всіх раніше зібраних даних і прописуванням найбільш підходящої ціни.

Виклики: Сьогодні результати описових та прогностичних моделей вже використовуються для підтвердження рішень щодо продукту. Описові методи особливо поширені для того, щоб зобразити, які результати минулих опитувань або аналізу ринку були отримані. Іноді прогнозні моделі також використовуються, коли справа доходить до оцінки того, який продукт може бути найбільш адекватним, враховуючи оцінки, наприклад, майбутнього потенціалу обсягу, доходу або збільшення частки ринку. Головною перешкодою для застосування прогностичних моделей на цьому етапі PDP може бути сприйняття достовірності прогнозів. Ця проблема нікуди не зникає, коли мова заходить про методи прескриптивної аналітики. Насправді це може навіть посилитися. Таким чином, вкрай важливо зруйнувати ці бар'єри, що може бути можливим за допомогою методів з області Explainable AI (XAI) або Interactive ML (IML). Ключовий етап PDP пов'язаний з прийняттям рішення щодо правильного продукту, враховуючи результати всіх попередніх кроків. Наприклад, можна було б призначити оперативні рішення та оцінити, чи перевершують вони базові показники. Будучи третім кроком дедукції, цей крок знову може використовувати всі рівні аналітики даних для опису,

прогнозування і, в кінцевому підсумку, визначення наслідків і продуктивності людських рішень. У 2019 році Берцімас і Каллус об'єднали концепції ML з дослідженням операцій і наукою про управління (OR–MS) для прописування оптимальних рішень у задачах OR–MS. Автори використовували статистичну теорію прийняття рішень для вибору оптимальних оцінок для приписів на основі даних двоетапного планування відвантаження, на які вони націлені. На описовій основі вони оцінювали різні k найближчих сусідів (KNN), ядра та локальні лінійні методи, а також дерева та ансамблі. Зокрема, вони використовували алгоритми найближчого сусіда для визначення кластерів у даних по поведінку клієнтів, перш ніж застосувати регресію KNN для прогнозування поведінки клієнтів. Примітно, що запропонований ними коефіцієнт прескриптивності, який забезпечує міру точності в двоступеневій задачі лінійного прийняття рішень, з якою вони стикаються, вносить свій вклад в область прескриптивної аналітики на останньому етапі PDP.

PDP можна відрізнити від процесів розробки продукту, оскільки утилізація продукту та кінцеве виробництво не відіграють ключової ролі. Однак PDP завершується етапом підтримки прийняття рішень, на якому всі дані, отримані та оцінені в попередніх процесах, використовуються для покращення прийняття рішень людиною. Примітно, що ця процедура не відображає весь процес розробки продукту, що спричинило б за собою безліч інших процесів, таких як тестування. На рис. 1.3 зображено аналітичну структуру поряд із загальним PDP.



Рисунок 1.3 – Аналітична основа разом із загальним процесом розробки продукту.

Як зазначалося вище, сучасний аналіз даних про використання клієнтами не має чіткого визначення PDP. Таким чином, у цьому огляді визначено процес PDP, який демонструє, які області аналізу можуть бути розгорнуті на кожному етапі процесу прийняття рішення про продукт. Для кожного з п'яти кроків PDP були визначені проблеми та описані приклади того, як можуть бути використані методи аналізу даних про використання клієнтами. У зв'язку з поширенням даних і постійно зростаючими можливостями використання даних про використання клієнтами необхідно вирішити, які методи аналізу можуть бути застосовані до якого результату. Існує безліч методів оцінки та аналізу даних про використання клієнтів, але для вибору продукту не потрібно представляти результати аналізу даних про клієнтів. Цей огляд сприяє тому, що ділиться підходами, які використовуються для отримання конкретних результатів. Крім того, він показує, які області аналізу можуть бути використані на кожному етапі PDP і які аналізи повинні бути використані на якому рівні PDP. Результати, отримані в результаті цього огляду, можуть бути використані на практиці для ілюстрації зростаючої важливості даних про використання клієнтами в усіх сферах. У зв'язку з різким збільшенням даних про клієнтів можна представити уявлення про аналіз і його методи. Цей огляд приносить користь традиційним компаніям, яким доводиться мати справу з новим інформаційним каналом даних про використання клієнтів. З огляду на виклики та приклади PDP, було продемонстровано, як методи аналізу даних про використання клієнтів можуть бути використані для їх вирішення. Він також демонструє, які процедури підходять для різних типів аналізів і які методи можуть дати найкращі результати. Ілюстрація глибини даних може бути використана для дослідження та надає практичні поради щодо використання правильних даних та методологій. Висновки цього огляду ілюструють, які аналітичні дані та результати можуть бути використані в PDP зі збільшенням обсягу даних про використання клієнтами. Це дозволяє створювати нові інсайти про клієнтів, а компанії можуть краще враховувати потреби клієнтів у розробці інновацій. Інтереси замовника можуть бути інтегровані в процес розробки, а процес прийняття рішень може бути вдосконалений.

2 АНАЛІЗ ПОВЕДІНКИ КЛІЄНТІВ ЗА ДОПОМОГОЮ СИСТЕМИ АНАЛІТИКИ ДАНИХ

З величезного набору даних для підвищення точності та ефективності прогнозування, методи інтелектуального аналізу даних є дуже ефективними інструментами для вилучення прихованих знань. Інтелектуальна аналітична система прийняття рішень потребує інтеграції аналізу та прогнозування рішень. Точність прогнозу продажів забезпечує величезний вплив на бізнес. Багато бізнес-організацій в основному покладаються на базу знань і прогнозування попиту на тенденції продажів. У цьому розділі проведено детальне вивчення та аналіз зрозумілих прогностичних моделей для покращення прогнозів майбутніх продажів. Ці проблеми можна подолати за допомогою різних методів інтелектуального аналізу даних. У цьому розділі коротко проаналізовано поняття даних про продажі та прогноз продажів. Визначено різні методи та заходи для прогнозування продажів. На основі оцінки ефективності запропоновано відповідну прогностичну модель для прогнозу тенденцій продажів.

2.1 Аналіз існуючих систем аналітики даних

Прогнозування поведінки клієнтів – це щось на кшталт створення математичної моделі для зображення загальної поведінки, що спостерігається серед певних груп клієнтів, щоб передбачити, як схожі клієнти будуть поводитися в схожих ситуаціях. Існуючі моделі поведінки клієнтів залежать від методів інтелектуального аналізу даних, і кожна модель призначена для відповіді на одне запитання одночасно. Наприклад, ця модель може бути використана для прогнозування поведінки певної групи споживачів з ефектом маркетингової стратегії. Якщо ця модель була ефективною, то маркетолог буде слідувати тій же стратегії, щоб залучати все більше і більше груп клієнтів. Але існуюча система була складною і дорогою, тому що математичні методи та інструменти, які використовували експерти, були дуже складними і дуже дорогими. Навіть після

побудови дорогої моделі, яка була дуже дорогою в маніпулюванні та обробці, щоб з'ясувати, що саме маркетолог повинен зробити, щоб привести клієнтів до своєї компанії. Але, з іншого боку, багато моделей були дуже простими і передбачуваними, тому що вони пропускали деякі функції, які були для них перешкодою.

З наведеного вище огляду літератури можна побачити, що на основі алгоритму машинного навчання в якості комбінації були прийняті узагальнені лінійні моделі, дерево рішень, дерево градієнтного прискорення. Для картографування та групування даних було використано комбінацію алгоритму класифікації C4.5 та моделі зменшення карти. Застосовувались такі великі концепції, як apache, hadoop і HDFS. Також у них була функція візуалізації даних за допомогою D3.js. Їх ефективність і масштабованість були порівняно хорошими. Також застосовувалась техніка інтелектуального аналізу даних разом із правилом асоціацій. Також використовувалась техніка CRM, тобто техніка управління взаємовідносинами з клієнтами, а також деякі методи інтелектуального аналізу даних, які можуть бути використані для отримання знань з великої кількості наборів даних. Для виконання проекту, поряд з CRM, використовувався процес введення правил на кластеризованих даних з бази даних клієнтів на основі їх запитів. Вони надавали важливого значення утриманню даних, дистиляції даних, запитам. За допомогою відповідей на ці запитання було передбачено багато гіпотез. У наступних роботах використовували RFM разом з алгоритмом k-середніх. Цей алгоритм дуже ефективний у наборі даних, що має неоднакові змінні різної величини. Після цього з'ясували, що грошова вартість має деяку різницю в давності та частоті. Наступна робота була присвячена набору даних електронної комерції та використовувала м'яку кластеризацію. Використовувався метод м'якої кластеризації, який застосовує латентний підхід кластеризації змішаних класів для класифікації клієнтів на основі категорій. Основна відмінність у цьому полягала в тому, що він повністю залежав від Інтернету. Також використовувався метод Дріхле, який полягає в сегментації результатів за багатьма категоріями, і результати були кращими, ніж жорстка кластеризація та модель скінченної суміші.

2.2 Порядок реалізації запропонованої системи аналітики даних

Щоб дещо відрізнитися від існуючої системи, у цій роботі пропонується нова комбінація алгоритмів. Тут в основному був реалізований алгоритм лінійної регресії та випадкового лісу. Порівнюються результати цих методологій, щоб з'ясувати високу точність. До основних модулів цього процесу належать (рис.2.1):

- генерація гіпотези;
- імпорт набору даних;
- попередня обробка даних;
- вилучення ознак;
- застосування алгоритму;
- прогнозування результату;
- побудова та візуалізація найкращої моделі.

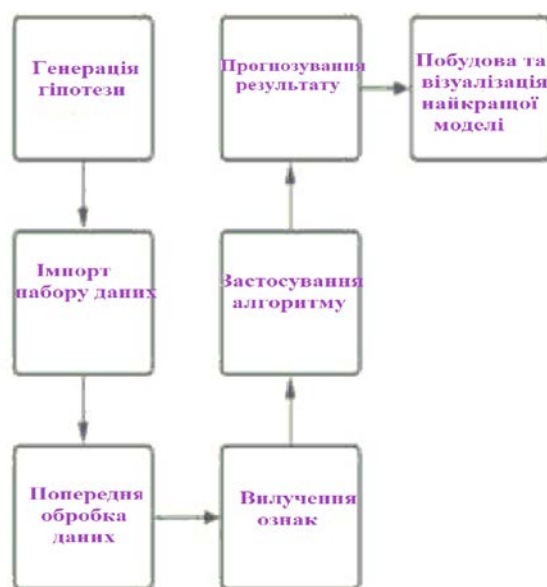


Рисунок 2.1 – Модулі системи аналітики даних

2.3 Генерація гіпотези

Це перший етап для продовження проекту, на цьому етапі ретельно аналізується постановка проблеми і генеруються різні гіпотези. Гіпотеза - це те, що показує взаємозв'язок між різними даними. Отже, ми повинні створити гіпотезу,

поставивши багато питань всередині нас, і ми повинні звернутися до багатьох джерел і сформулювати цю гіпотезу, після чого ми повинні уточнити її та знайти для неї рішення. У нашому випадку завдання полягає в тому, щоб спрогнозувати найбільш підходящий метод, щоб зрозуміти поведінку клієнтів і з'ясувати, як вони будуть вести себе в різних обставинах. Розглядаючи нашу проблему, ми повинні враховувати деяку інформацію, наприклад, з якого типу міста продається товар, якщо це міська зона, то продажі будуть високими. Тоді ми враховується місткість магазину, так само великі магазини матимуть більше продажів. Якби маркетингові стратегії були кращими і не було конкуруючих магазинів, то можна було б очікувати хороших продажів. Якщо компанія дає привабливу рекламу і знижки, то продажі будуть високими. Це різні гіпотези, згенеровані для нашого проекту.

2.4 Імпорт набору даних

Набір даних, обраний тут, був набором даних, отриманим з різних магазинів, які містять усі продукти, що використовуються в домашній діяльності. Він був зібраний у 2022 році і налічує близько 1560 товарів з 10 міст. Було завантажено цей набір даних з веб-сайту Kaggle. У ньому згадувалися атрибути цих продуктів. Атрибути включають-Код товару, який позначає унікальну ідентичність товару, потім вагу - позначає вагу продукту, рівень жиру - це ще один атрибут, з якого ми можемо отримати вміст жиру в товарі, видимість продукту показує % від загальної площі відображення, виділеної цьому товару в магазині. Тип товару позначає категорію, до якої цей товар належить, MRP дає вартість товару. Outlet показує унікальний ідентифікатор магазину, рік показує рік, в якому був заснований магазин. Розмір торгової точки - це розмір магазину з точки зору охопленої площі. Розташування показує тип міста, в якому розташований магазин, будь то міський / сільський. Тип Outlet вказує на те, чи є торгова точка просто продуктовим магазином або свого роду універмагом. Продажі показують продажі товару в конкретному магазині. Було використано програму python для програмування, і за допомогою цих бібліотек були імпортовані набори даних.

2.5 Попередня обробка даних

Набір даних — це сукупність об'єктів даних, які також називають записами, точками, векторами, шаблонами, подіями, випадками, зразками, спостереженнями або сутностями. Попередня обробка даних - це той крок, на якому дані перетворюються, щоб привести їх до стану, який тепер машина може легко їх проаналізувати. Особливості даних тепер можуть бути легко інтерпретовані алгоритмом. Ознаки можуть бути як категоріальними, так і числовими.

Категоріальний відноситься до об'єктів, які мають визначений набір значень. Числовий відноситься до об'єктів, які включають числові значення, які можуть бути неперервними або дискретними. Етапами, пов'язаними з попередньою обробкою даних, є очищення даних, перетворення даних, скорочення даних. Тут ми виконуємо процес очищення даних, щоб знайти та замінити всі нульові значення, доступні в наборі даних. У цьому очищенні даних виявляються відсутні значення і вони підраховуються за допомогою функції суми, а техніка заміни виконується середнім значенням тієї категорії, яка має відсутнє значення. У нашій задачі ми змогли знайти нульові значення в атрибуті продажів і вазі, розміру Outlet. Ці нульові значення були замінені середніми значеннями. Після цього кроку було знайдено чіткий підрахунок усіх предметів. На основі кожної категорії, як-от товар, тип торгової точки, розмір, вага, тип товару, було розраховано їх кількість.

2.6 Вилучення ознак

Ознака - це властивість, спільна для всіх незалежних значень, на основі яких можна зробити прогнозування. Будь-яка властивість може бути функцією, якщо вона корисна для моделі. Інженерія функцій — це процес використання знань предметної області для вилучення функцій із необроблених даних за допомогою методів інтелектуального аналізу даних. Ці функції можуть бути використані для підвищення продуктивності проекту. Проектування характеристик можна вважати важливим етапом будь-якого проекту. Було розглянуто деякі нюанси в наборі даних

у розділі вилучення ознак. На рис.2.2 представлено розподіл сукупності за ознаками.

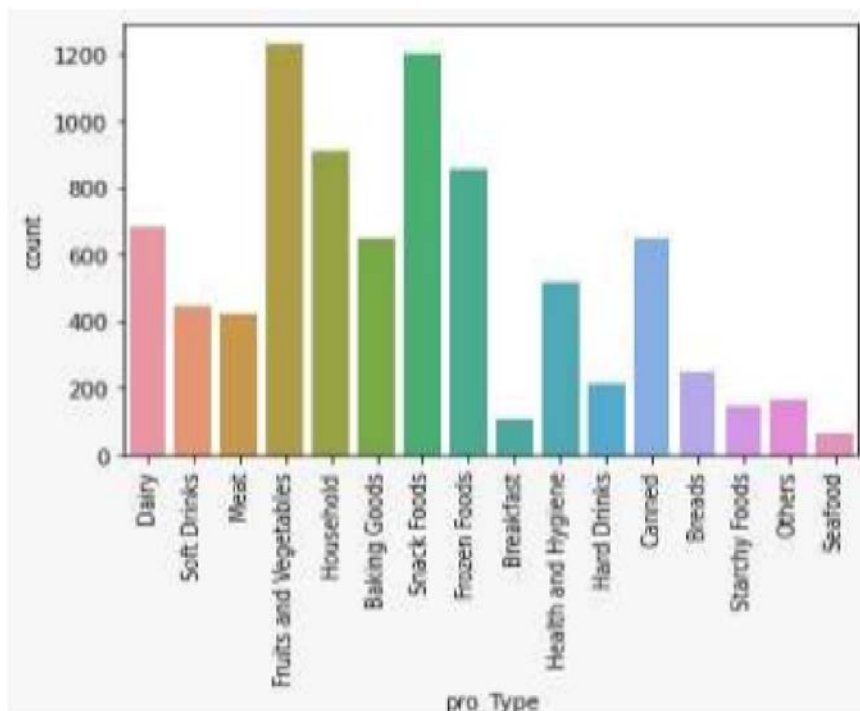


Рисунок 2.2– Розподіл сукупності за ознаками.

Було виявлено, що тип торгової точки має деяку плутанину, тому тут і тип супермаркету 2, і тип супермаркету 3 об'єднані разом. Після цієї комбінації в цьому є якась явна суттєва різниця. І кількість типів розеток розраховується в кожній категорії. Тоді враховується атрибут видимості, оскільки вони мають вищу важливість характеристик, якщо товар має кращу видимість і зберігається на великій площі, то це матиме хороші продажі. За типом продукції у нас було 16 видів продуктів, їх поділили на три основні типи: продукти харчування, неспоживані продукти та напої. Тоді ми розглядаємо атрибут рівня жиру, спочатку для харчових продуктів дотримується вміст жиру і ділиться на їжу з низьким вмістом жиру і звичайну жирну їжу. Їх кількість підраховується. Потім модифікуються подальші категорії, тому що у нас є продукти, які не належать до продуктів харчування. Так до модифікованих категорій відносяться знежирені, звичайні жирні, неїстівні продукти. Графік розподілу продуктів за рівнем жиру представлено на рис. 2.3.

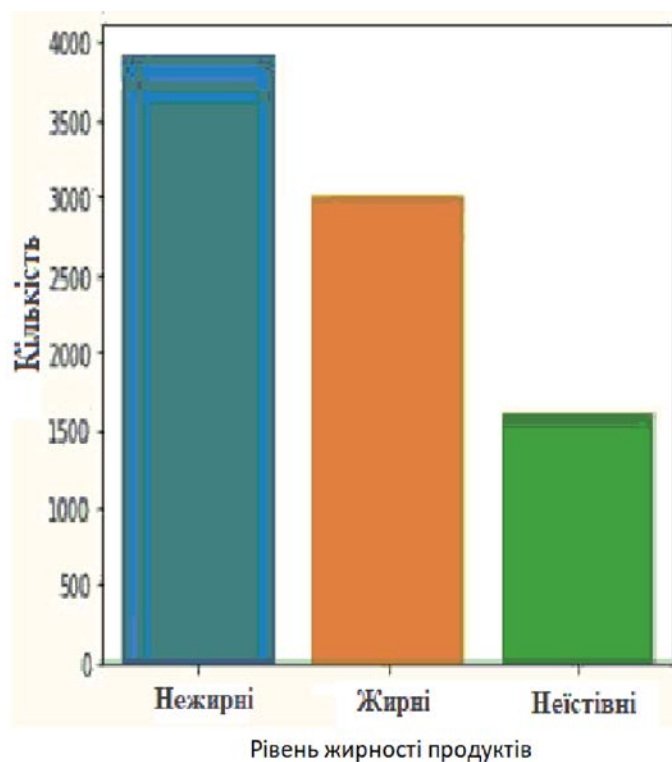


Рисунок 2.3— Графік розподілу продуктів за рівнем жирності

Потім наступною характеристикою був розмір торгової точки. У цю категорію входять магазини середнього, малого і високого рівня. Якби торговою точкою товару був дуже великий магазин, то продажі були б порівняно нижчими, оскільки є всі класи людей, що належать до різних місць, таких як міські, сільські та невеликі райони, які були віднесені до рівня 1, рівня 2 та рівня 3. Графік розподілу продуктів за розміром торгових точок представлено на рис. 2.4.

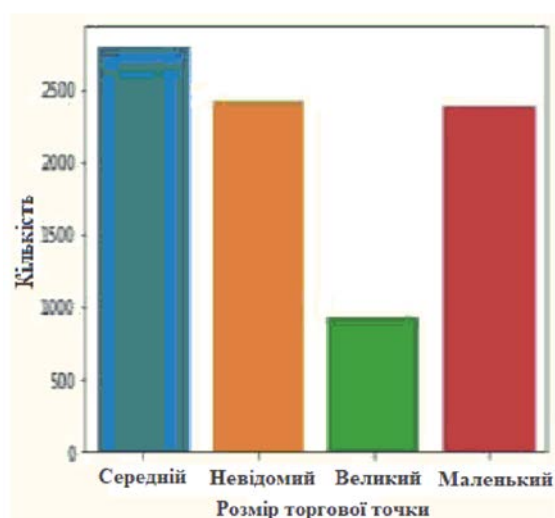


Рисунок 2.4 – Графік розподілу продуктів за розміром торгових точок

Як зазначалося вище, на наступному графіку (рис 2.5) показано кількість товарів залежно від типу локації.

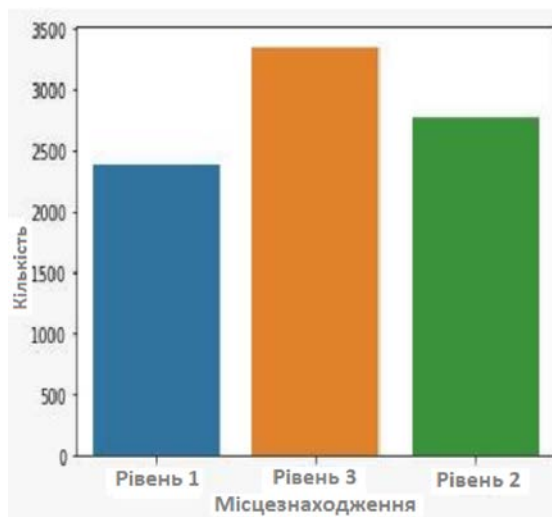


Рисунок 2.5 – Графік розподілу продуктів за типом населеного пункту

Потім було побудовано графік на основі типу торгової точки (рис 2.6).

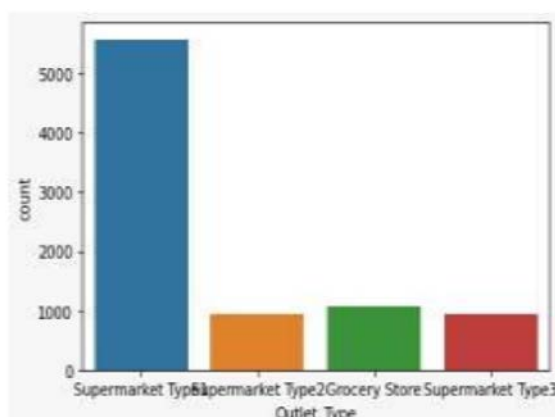


Рисунок 2.6 – Графік розподілу продуктів за типом торгової точки

Отже, цей етап вилучення ознак завершено, і всі ці вилучені об'єкти будуть включені до наступного кроку застосування алгоритму. Ця особливість даних вплине на кінцеві результати. Маючи цю технологію, можна створити максимально точну структуру даних, а отже, буде простіше створити найкращу модель. Зменшення кількості функцій за рахунок вилучення функцій гарантує, що навчання моделі вимагатиме менше пам'яті та обчислювальної потужності, що обумовлює

скорочення часу навчання, а також допоможе зменшити ймовірність перенавчання. Спрощення навчальних даних зробить модель легкою для інтерпретації, що може бути важливим при обґрунтуванні реальних проблем в результаті виходів моделі.

2.7 Застосований алгоритм

На відміну від інших реалізацій, було вирішено застосувати лінійну регресію та алгоритм випадкового лісу. Багато дослідників використовували кілька алгоритмів, таких як узагальнена лінійна модель, дерево рішень, градієнтне прискорення, алгоритм класифікації C4.5 і модель зменшення карти, техніку інтелектуального аналізу даних разом із правилом асоціацій, техніку CRM, правило процесу індукції, техніку RFM, K - алгоритм.

2.7.1 Алгоритм лінійної регресії

Лінійна регресія використовується для моделей прогнозного аналізу. Це техніка, яка пояснює ступінь зв'язку між двома або більше змінними (множинна регресія) за допомогою лінії найкращої відповідності. Проста лінійна регресія використовується, коли у нас є одна незалежна змінна і одна залежна змінна. Діаграма розсіювання використовується для розміщення однієї лінії в цій техніці. Проста форма лінійної регресії з однією залежною і однією незалежною змінною виражається рівнянням:

$$Y = aX + b \quad (2.1)$$

Це рівняння відоме як рівняння лінійної регресії, де Y — цільова змінна, X — вхідна змінна, « a » — нахил, а « b » — перехоплення.

Та лінія, яка може краще пояснити відносини, називається найкращою лінією. З іншого боку, лінія найкращої апроксимації поверне найточніше значення Y на основі X , що спричиняє мінімальну різницю між фактичним та прогнозованим значенням Y . На рис. 2.7 представлено приклад простої лінійної регресії з однією незалежною змінною.

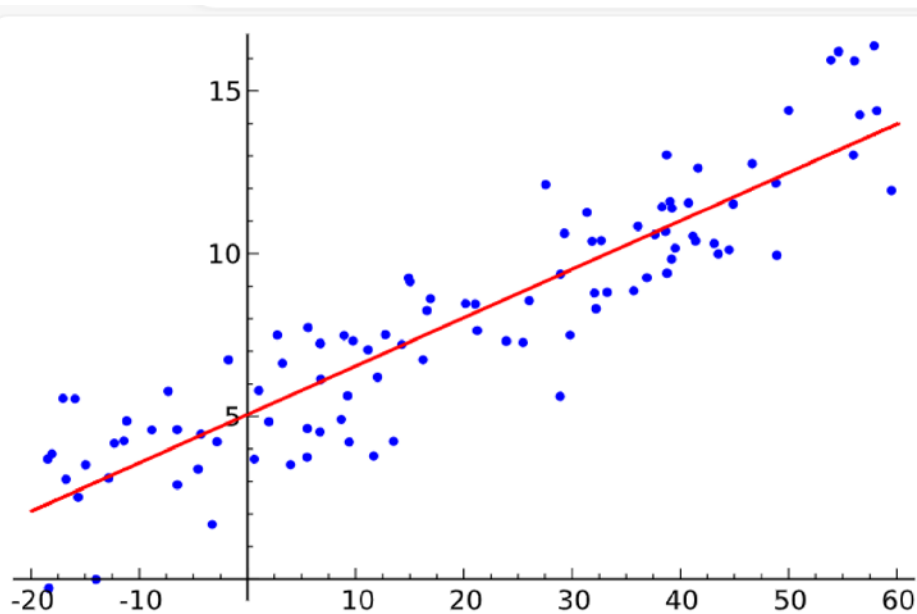


Рисунок 2.7 – Приклад лінійної регресії з однією незалежною змінною

На рис. 2.8 представлено графік оцінювання моделі прогнозного аналізу за допомогою застосування алгоритму лінійної регресії.

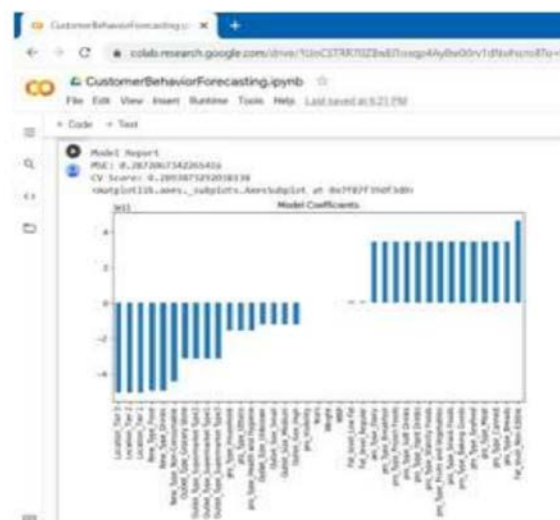


Рисунок 2.8 – Графік оцінювання моделі прогнозного аналізу за допомогою застосування алгоритму лінійної регресії.

З цього результату ми могли побачити, що середнє квадратичне значення похибки становить приблизно 0,287 (приблизно), а оцінка перехресної варіації становить 0,289 (приблизно).

2.6.1 Алгоритм випадкового лісу

Алгоритм випадкового лісу - це дуже проста техніка, за допомогою якої ми можемо виміряти відносну важливість кожної функції, яку вилучили в процесі розробки об'єктів. Він вимірює об'єкт, обчислюючи, наскільки вузли дерев використовують цю функцію, щоб зменшити домішки на всіх деревах у лісі. Він також обчислює значення для кожної функції після тренування та виводить суму всіх важливих ознак. За важливістю цих функцій можна передбачити, які функції потрібні, а які непотрібні, тому функції, які не вносять, можна пропустити. Цей алгоритм має найкращу продуктивність і може краще працювати у всіх методах інтелектуального аналізу даних. На рис. 2.8 представлено приклад алгоритму випадкового лісу.

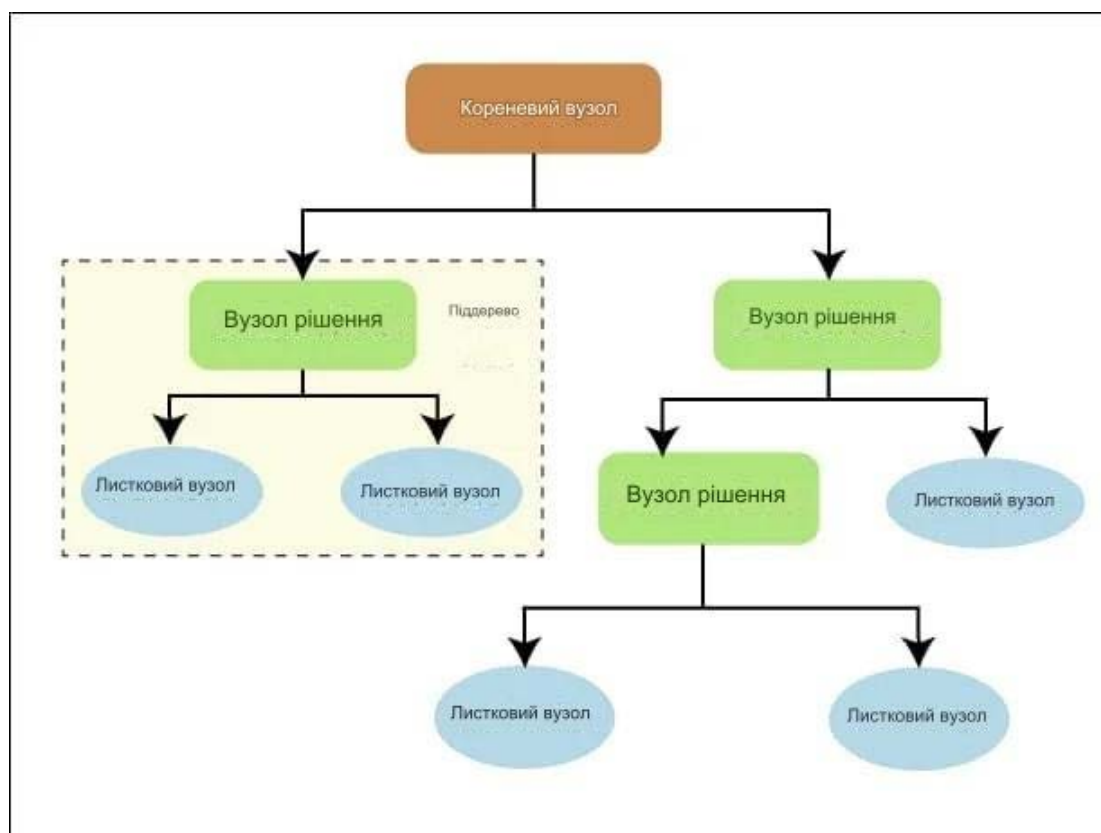


Рисунок 2.8 – Приклад алгоритму випадкового лісу

На рис. 2.9 представлено графік оцінювання моделі прогнозного аналізу за допомогою застосування алгоритму випадкового лісу.

3 МЕТОДОЛОГІЯ АНАЛІЗУ ДАНИХ

У цьому розділі представлено ключові деталі методології, яка використовується для аналізу даних разом із інформацією, включеною в набір даних. Крім того, проаналізовано методи первинної попередньої обробки, що використовуються для очищення та маніпулювання даними.

Основна мета цього дослідження полягає в тому, щоб визначити купівельну поведінку та переваги клієнтів. Для досягнення цієї мети в дослідницькій частині цієї дипломної роботи було застосовано чітку та добре структуровану методологію, а також кілька інструментів, підходів і методологій.

Дослідницький аналіз даних і ключові висновки щодо поведінки клієнтів були додатково показані на інформаційних панелях за допомогою Power BI, тоді як аналіз RFM було реалізовано за допомогою мови програмування R, використана версія R 4.2.2.

3.1 Служба бізнес-аналітики Power BI

Power BI – це служба бізнес-аналітики, розроблена корпорацією Майкрософт, яка забезпечує можливості діалогової візуалізації та бізнес-аналітики за достатньо простим інтерфейсом для кінцевих користувачів з можливістю створювати власні звіти та інформаційні панелі. Це дозволяє користувачам підключатися до широкого діапазону джерел даних, перетворювати та очищати дані, а також створювати звіти та інформаційні панелі за допомогою візуалізацій перетягування та запитів природною мовою. Power BI включає широкий вибір типів візуалізації, від базових діаграм і таблиць до більш розширених опцій, таких як теплові карти, карти дерев і спеціальні візуальні елементи. Він також надає функції співпраці та обміну даними, дозволяючи користувачам ділитися звітами та інформаційними панелями та спільно працювати над ними з іншими у своїй організації або навіть з зовнішніми зацікавленими сторонами. Доступ до Power BI можна отримати через веб-

інтерфейс, мобільні програми або інтегровані з іншими продуктами Microsoft, наприклад, Excel і SharePoint. Загалом Power BI є потужним інструментом для аналізу даних і візуалізації, що дозволяє організаціям знаходити тенденції та інформацію, цінні для прийняття рішень на основі їхніх даних.

3.1.1 R-Загальна концепція

Мова програмування R — це мова програмування з відкритим кодом і середовище для статистичних обчислень і графіків, які широко використовуються в аналізі даних, машинному навчанні та статистичному моделюванні. Це розроблений GNU, який має деякі схожості з мовою S, створеною лабораторіями Bell La. Він надає широкий спектр інструментів для обробки даних, візуалізації та статистичного аналізу, а його гнучкість і розширюваність роблять його найкращим вибором для дослідників, аналітиків та дослідників даних у різних галузях.

3.1.2 Бібліотеки, що використовуються в мові програмування R

Пакет бібліотеки *Lubridate*: цей пакет створено для полегшення роботи з датами та часом у мові програмування R. Він надає набір функцій, які дозволяють аналізувати, маніпулювати та форматовувати дати та час. Деякі приклади того, що можна зробити за допомогою пакету *lubridate*, включають вилучення компонентів дати (наприклад, року, місяця, дня), додавання або віднімання інтервалів часу та перетворення між різними форматами дати й часу.

Пакет бібліотеки *Dplyr*: цей пакет містить набір інструментів для роботи з кадрами даних у мові програмування R. Він включає функції для фільтрації, сортування, узагальнення та модифікації даних, а також об'єднання кількох кадрів даних. Деякі з найбільш використовуваних функцій у пакеті *Dplyr* включають `filter()`, `mutate()`, `group_by()` і `summarize()`. *tidyverse*: це набір пакетів, включаючи *dplyr*, *ggplot2*, *tidyr* і кілька інших.

Пакет бібліотеки *Tidyverse* розроблено, щоб забезпечити набір інструментів

для обробки даних, візуалізації та аналізу, які добре працюють разом і дотримуються узгодженого набору принципів проектування. Деякі з ключових функцій і особливостей Tidyverse включають використання труб (`%>%`) для об'єднання кількох етапів обробки даних, використання принципів акуратних даних для організації даних у рядки та стовпці та використання ggplot2 для створення високоякісних візуалізацій даних.

Пакет бібліотеки *"Janitor"*: цей пакет розроблений, щоб допомогти виконувати завдання очищення та форматування даних у мові програмування R. Він надає набір функцій, які можуть допомогти видаляти непотрібні стовпці, перейменовувати стовпці і обробляти відсутні дані. Він також містить деякі корисні функції для візуалізації даних і роботи з категоріальними даними.

Пакет бібліотеки *"Skimr"*: цей пакет надає легкий для читання перелік набору даних, який може бути корисним, коли потрібен швидкий огляд даних. Підсумок містить основні статистичні дані, такі як середнє значення, медіана та стандартне відхилення, а також візуалізації, такі як гістограми та діаграми розсіювання.

3.2 Опис набору даних

Набір даних, який використовується для аналізу, містить інформацію про зроблені клієнтами покупки у звичайних магазинах із січня 2019 року по червень 2022 року. Набір даних було отримано з роздрібної компанії в секторі краси/косметики.

Включено такі змінні:

Receipt_id: номер рахунку-фактури. Унікальний номер кожної транзакції.

Customer: ідентифікатор клієнта – унікальний номер клієнта.

Subcategory: підкатегорія, до якої належить кожен продукт.

Brandname: бренд, якому належить кожен продукт.

Category: категорія, до якої належить кожен продукт.

Retailer:: різні роздрібні магазини, де були зроблені покупки.

City:: місто, де було здійснено кожну покупку.

Order_date: дата розміщення кожного замовлення.

Qty: кількість продуктів. Вона показує, скільки продуктів було продано в кожному рядку у наборі даних.

Finalprice: представляє загальну ціну, сплачену клієнтом за кожен рядок у наборі даних.

Vat:: податок на додану вартість – податок, який сплачується на продукти в кожному рядку в наборі даних.

3.3 Попередня обробка даних

Після встановлення та завантаження відповідних пакетів дані імпортуються в мову програмування R. Однак, щоб обмежити дані лише тими записами з 2019 року, пакет tidyverse використовується для створення нового кадру даних під назвою datanew. Цей новий кадр даних містить лише ті записи, які відповідають вищезазначеним критеріям.

Важливо мати перший погляд на те, як виглядають дані, щоб висвітлити будь-які проблеми в структурі та зафіксувати їх. Для подальшого пояснення початкової форми використовуються різні функції перед переходом до фази попередньої обробки.

На першому етапі перші шість рядків даних відображаються за замовчуванням за допомогою функції head(). Функція head() у мові програмування R — це швидкий і простий спосіб переглянути кілька перших рядків кадру даних, який може бути корисним для отримання швидкого огляду даних перед виконанням більш складного аналізу. Оскільки це компактне представлення даних, воно може бути корисним для виявлення будь-яких проблем або невідповідності у даних, наданих для аналізу. На рис. 3.1 представлено лістинг перегляду даних у мову програмування R.

Оскільки виконання статистичного аналізу та створення візуалізацій має велике значення для дослідження, символічні змінні, названі підкатегорією, брендом, категорією, роздрібним продавцем і містом, перетворюються на факторні змінні для кращого впорядкування та групування даних і розподілу на окремі категорії чи групи. Паралельно надається право змінній *order_date* створювати три нові змінні у кадрі даних, коли ми розділяємо місяць і рік, і створювати квартали зі змінної *order_date*. На рис. 3.3 представлено лістинг стану нових змінних (1/2), а на рис. 3.4 представлено лістинг стану нових змінних (2/2).

```
> datanew$brandname <- factor(datanew$brandname)
> datanew$subcategory <- factor(datanew$subcategory)
> datanew$category <- factor(datanew$category)
> datanew$retailer <- factor(datanew$retailer)
> datanew$city <- factor(datanew$city)
> |
```

Рисунок 3.3 – Лістинг стану нових змінних (1/2)

```
> #separate month - year - quarters + right format of full date
> datanew$order_date <- as.Date(datanew$order_date, format="%Y-%m-%d")
> datanew$month <- format(datanew$order_date, format = "%m")
> datanew$year <- format(datanew$order_date, format = "%Y")
> datanew$quarters<- quarters(datanew$order_date)
> |
```

Рисунок 3.4 – Лістинг стану нових змінних (2/2)

Оскільки ми маємо справу з великими даними, відсутні значення можуть суттєво вплинути на точність аналізу. Використана функція *drop_na()* є частиною пакету *tidyr*. Він працює з кадром даних і видаляє будь-який рядок, який містить принаймні одне відсутнє значення (NA). Очищення даних виконується, як показано нижче на рисунку 3.5.

```
> #cleaning_data
> ndata <- datanew %>%
+   drop_na()
```

Рисунок 3.5 – Лістинг очищення даних

3.4 Аналіз RFM

3.4.1 Розрахунок значення актуальності

Значення актуальності представляє час, що минув з моменту останнього здійснення клієнтом замовлення.

Спочатку створюється змінна під назвою *refDay* (Reference Day), де вона представляє максимальну дату набору даних.

Значення *Recency* обчислюються для кожного клієнта за допомогою нової змінної і пакету *dplyr* у мові програмування R. На рис. 3.6 представлено лістинг розрахунку змінної актуальності.

```
> refDay = max(ndata$order_date)
>
> Recency = ndata %>% group_by(customer) %>%
+   summarise(Recency = as.numeric(refDay)- as.numeric(max(order_date)))
> |
```

Рисунок 3.6 – Лістинг розрахунку змінної актуальності

3.4.2 Розрахунок значення частоти

Значення частоти вказує на те, як часто кожен клієнт робить покупку. За допомогою пакету *dplyr* його можна легко обчислити наступним чином, як показано нижче на рисунку 3.7.

```
> Frequency = ndata %>% group_by(customer) %>%
+   summarise(Frequency = n())
> |
```

Рисунок 3.7 – Лістинг розрахунку змінної частоти

3.4.3 Розрахунок грошової вартості

Оцінка RFM завершуються обчисленням грошової вартості, як показано на рис. 3.8.

```
> Monetary = ndata %>% group_by(customer) %>%
+   summarise(Monetary = sum(finalprice))
```

Рисунок 3.8– Лістинг обчислення грошової вартості

Ці три значення мають значення Recency, Frequency і Monetary поєднуються окремо. На рис. 3.9 представлено лістинг моделювання RFM.

```
> RFM = merge(Recency, Frequency, by = "customer")
> RFM = merge(RFM, Monetary, by = "customer")
> |
```

Рисунок 3.9– Лістинг моделювання RFM

3.5 Розрахунок балів RFM

Розрахунок балів за критерієм Recency для кожного клієнта представлено на рисунку 3.10, а розрахунок балів за критерієм Frequency для кожного клієнта представлено на рисунку 3.11.

```
> quantile(RFM$Recency)
0%  25%  50%  75% 100%
 0  146  374  883 1274
>
> Rscore = cut(RFM$Recency, breaks = c(0,150,380,650,890,1300))
> levels(Rscore) = c(5,4,3,2,1)
> |
```

Рисунок 3.10– Лістинг розрахунку балів за критерієм Recency для кожного клієнта

```

> quantile(RFM$Frequency)
 0%   25%   50%   75%  100%
  1     1     2     5   647
>
> Fscore = cut(RFM$Frequency, breaks = c(0,1,2,5,20,650))
> levels(Fscore) = c(1,2,3,4,5)
> |

```

Рисунок 3.11– Лістинг розрахунку балів за критерієм Frequency для кожного клієнта

Розрахунок балів за критерієм Monetary для кожного клієнта представлено на рисунку 3.12.

```

> quantile(RFM$Monetary)
      0%      25%      50%      75%     100%
  0.00    55.92   101.17   212.84 33185.47
> Mscore = cut(RFM$Monetary, breaks = c(0,56,102,215,220,34000))
> levels(Mscore) = c(1,2,3,4,5)
> |

```

Рисунок 3.12– Лістинг розрахунку балів за критерієм Monetary для кожного клієнта

Після обчислення балів R, F і M клієнтів ці три бали поєднуються з їхніми відповідними ідентифікаційними номерами, і створюється новий кадр даних під назвою RFM_Scores, як показано на рис. 3.13 – 3.14.

```

> RFM_Scores = cbind(RFM$customer, Rscore, Fscore, Mscore)
> head(RFM_Scores)
      Rscore Fscore Mscore
[1,] "1000618326783" "5"   "1"   "1"
[2,] "1000618425706" "5"   "1"   "3"
[3,] "1000621099384" "5"   "2"   "2"
[4,] "1000625319802" "5"   "3"   "3"
[5,] "1000626113114" "5"   "3"   "3"
[6,] "1000626115846" "4"   "3"   "2"

```

Рисунок 3.13– Лістинг розрахунку балів RFM для кожного клієнта

```

> RFM_Scores = as.data.frame(RFM_Scores)
> colnames(RFM_Scores) = c("customer", "R_Score", "F_Score", "M_Score")
> head(RFM_Scores)
  customer R_Score F_Score M_Score
1 1000618326783     5       1       1
2 1000618425706     5       1       3
3 1000621099384     5       2       2
4 1000625319802     5       3       3
5 1000626113114     5       3       3
6 1000626115846     4       3       2

```

Рисунок 3.13– Лістинг маніпулювання даними

3.6 Пояснення сегментів RFM

Не можна їх втратити: робив покупки регулярно, але давно.

Ризик: Подібно до попереднього сегменту, але клієнти характеризуються нижчим рівнем частоти.

Сплячий режим: найнижчі показники останнього часу та частоти.

Збирається спати: середній показник нещодавності та нижче середнього показник частоти.

Потрібна увага: середній рівень давності та частоти.

Постійні клієнти: активні та цінні клієнти. Середній і вищий за середній вік оцінки Resensy та високий рівень частоти-

Новий клієнт: цей тип клієнтів купив товари порівняно недавно, але купив на найнижчому рівні частоти.

Перспективний: Подібний до вищезазначеної групи клієнтів – єдина відмінність полягає в тому, що це не такий високий рівень останніх балів.

Потенційний лояльний: ці клієнти вже купували товари більше одного разу – середній віковий рівень частоти та характеризуються високими показниками

Чемпіони: це найпопулярніші клієнти з точки зору як балів давності, так і балів частоти.

Таблиця, використана для аналізу RFM, наведена на рис. 3.14.

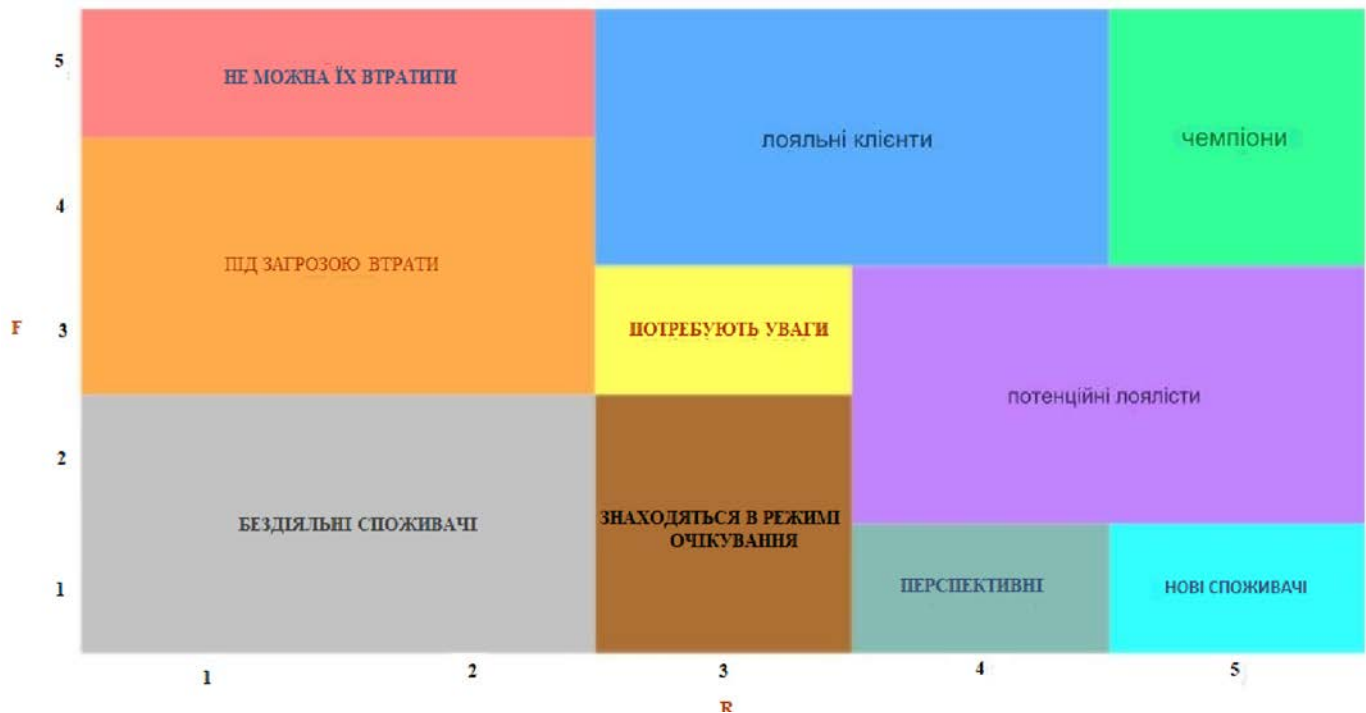


Рисунок 3.14– Таблиця аналізу RFM

Сегментація клієнтів була виконана на основі значень R і F. Значення RF об'єднуються та додаються як окремий стовпець, як показано на рис. 3.15.

```
> RF_scores = RFM_Scores$R_Score*10+RFM_Scores$F_Score
>
> RFM_Scores$RF_Scores = RF_scores
> head(RFM_Scores)
```

	customer	R_Score	F_Score	M_Score	RF_Scores
1	1000618326783	5	1	1	51
2	1000618425706	5	1	3	51
3	1000621099384	5	2	2	52
4	1000625319802	5	3	3	53
5	1000626113114	5	3	3	53
6	1000626115846	4	3	2	43

Рисунок 3.15– Додавання значень R, F

Наведені вище сегменти створено відповідно до наведеної вище таблиці балів, показано на рис. 3.16 – 3.19.

```

> Cant_lose = c(15,25)
> At_Risk = c(13,14,23,24)
> Hibernating = c(11,12,21,22)
> About_To_Sleep = c(31,32)
> Need_Attention = 33
> Loyal_Customers = c(34,35,44,45)
> Promising = 41
> New_Customer = 51
> Potential_Loyalists = c(42,43,52,53)
> Champions = c(54,55)
> |

```

Рисунок 3.16 – Створення сегментів

```

> #cant loose them segment
> for(i in 1:length(Cant_lose)){
+   for(k in 1:nrow(RFM_Scores)){
+     if(RFM_Scores$RF_Scores[k]==Cant_lose[i]){
+       RFM_Scores$RF_Scores[k] = "Cant'lose then"
+     }
+   }
+ }
>
> #at risk segment
> for(i in 1:length(At_Risk)){
+   for(k in 1:nrow(RFM_Scores)){
+     if(RFM_Scores$RF_Scores[k]==At_Risk[i]){
+       RFM_Scores$RF_Scores[k] = "At Risk"
+     }
+   }
+ }
>
> # hibernating segment
> for(i in 1:length(Hibernating)){
+   for(k in 1:nrow(RFM_Scores)){
+     if(RFM_Scores$RF_Scores[k]==Hibernating[i]){
+       RFM_Scores$RF_Scores[k] = "Hibernating"
+     }
+   }
+ }
>
> #About To Sleep segment
> for(i in 1:length(About_To_Sleep)){
+   for(k in 1:nrow(RFM_Scores)){
+     if(RFM_Scores$RF_Scores[k]==About_To_Sleep[i]){
+       RFM_Scores$RF_Scores[k] = "About To Sleep"
+     }
+   }
+ }
>

```

Рисунок 3.17– Розрахунок сегментів (1/3)

```

> #Need Attention segment
> RFM_Scores$RF_Scores = ifelse(RFM_Scores$RF_Scores == 33,
+                               "Need Attention", RFM_Scores$RF_Scores)
>
> #Promising segment
> RFM_Scores$RF_Scores = ifelse(RFM_Scores$RF_Scores == 41,
+                               "Promising", RFM_Scores$RF_Scores)
>
>
> #New Customer segment
> RFM_Scores$RF_Scores = ifelse(RFM_Scores$RF_Scores == 51,
+                               "New Customer", RFM_Scores$RF_Scores)
>
> #Loyal Customers segment
> Loyal_Customers = c(34,35,44,45)
> for(i in 1:length(Loyal_Customers)){
+   for(k in 1:nrow(RFM_Scores)){
+     if(RFM_Scores$RF_Scores[k]==Loyal_Customers[i]){
+       RFM_Scores$RF_Scores[k] = "Loyal Customers"
+     }
+   }
+ }
>
>

```

Рисунок 3.18 – Розрахунок сегментів (2/3)

```

> #Potential Loyalists segment
>
> Potential_Loyalists = c(42,43,52,53)
> for(i in 1:length(Potential_Loyalists)){
+   for(k in 1:nrow(RFM_Scores)){
+     if(RFM_Scores$RF_Scores[k]==Potential_Loyalists[i]){
+       RFM_Scores$RF_Scores[k] = "Potential Loyalists"
+     }
+   }
+ }
>
> #Champions segment
> Champions = c(54,55)
> for(i in 1:length(Champions)){
+   for(k in 1:nrow(RFM_Scores)){
+     if(RFM_Scores$RF_Scores[k]==Champions[i]){
+       RFM_Scores$RF_Scores[k] = "Champions"
+     }
+   }
+ }
> |

```

Рисунок 3.19 – Розрахунок сегментів (3/3)

Всі клієнти розділені на 10 сегментів. Створюється кадр даних під назвою newdata, який містить необхідну інформацію про клієнтів (Customer ID – Ідентифікатор клієнта) і сегменти, які належать кадру (рис. 3.20 – 3.21).

```

> newdata = data.frame(RFM_Scores$customer, RFM_Scores$RF_Scores)
> colnames(newdata) = c("Customer", "Segments")
> head(newdata)

```

	Customer	Segments
1	1000618326783	New Customer
2	1000618425706	New Customer
3	1000621099384	Potential Loyalists
4	1000625319802	Potential Loyalists
5	1000626113114	Potential Loyalists
6	1000626115846	Potential Loyalists

```

> |

```

Рисунок 3.20 – Створення кадру даних містить сегменти клієнтів

```

> Cant_lose = c(15,25)
> At_Risk = c(13,14,23,24)
> Hibernating = c(11,12,21,22)
> About_To_Sleep = c(31,32)
> Need_Attention = 33
> Loyal_Customers = c(34,35,44,45)
> Promising = 41
> New_Customer = 51
> Potential_Loyalists = c(42,43,52,53)
> Champions = c(54,55)
> |

```

Рисунок 3.21 – Результати кадру даних

Класифікація клієнтів на основі індивідуальних індексів нещодавності та частоти наведена в таблиці 3.1.

Таблиця 3.1 – Класифікація клієнтів на основі індивідуальних індексів нещодавності та частоти

Сегмент	Індекс R	Індекс F
Чемпіони	5	4,5
Лояльні	3,4	4,5
Потенційні лоялісти	4,5	2,3
Нові споживачі	5	1
Перспективні	4	1
Потребують уваги	3	3
Знаходяться в режимі Очікування	3	1,2
Не можна їх втратити	1,2	5
Під загрозою втрати	1,2	3,4
Бездіяльні споживачі	1,2	1,2

3.7 Результати дослідження та візуалізація

Після належного переформатування та очищення даних у R за допомогою DAX у Power BI створюється нова змінна з назвою Region для подальшої підтримки майбутнього аналізу даних та звітності. Конкретна змінна класифікує всі міста в наборі даних.

DAX — це мова формул, яка використовується в Power BI для створення спеціальних обчислень для аналізу даних. Формули DAX можна використовувати для створення нових обчислюваних стовпців, подібних до створеного.

Обчислювані стовпці — це стовпці, додані до таблиці, які містять формулу, яка обчислює нове значення для кожного рядка таблиці.

У Power BI розроблено інформаційну панель, яка містить шість окремих сторінок, кожна з яких містить чіткі показники та діаграми, які надають ключову інформацію про покупки, звички та переваги клієнтів. Перша сторінка містить інформацію про загальну купівельну поведінку клієнтів, зокрема дві картки, дві картки з кількома рядками, дві діаграми з областями, одну точкову діаграму та чотири кластеризовані гістограми як візуальні елементи та зрізувач за роками, який працює як фільтр і робить панель приладів динамічною.

Наданий кадр даних охоплює певний період та вказує на загальний отриманий дохід.

Друга сторінка звіту ілюструє аналіз частки ринку за категоріями та змінними бренду, які представлено на кругових діаграмах (рис. 3.22), як візуальні елементи. Він показує частку продажів і товарів, придбаних для кожної категорії та марки. Аналіз показує, що Марка 6 є найбільш домінуючим брендом з точки зору доходу та проданої продукції, займаючи 60% і 59% загальної частки ринку відповідно, і є найефективнішим брендом за всі роки окремо, потім Бренд 2. З іншого боку, Бренд 4 має найменшу частку ринку, займаючи лише 2% та 1% від загального доходу та кількості проданої продукції відповідно.

Категорія «Шкіра та волосся» приносить більше доходу, ніж дві інші категорії за вказаний період часу, займаючи 44% загальної частки ринку за всі роки. однак,

категорія макіяжу має найбільшу кількість проданих продуктів, займаючи 42% відсотки ринку від загальної кількості



Рисунок 3.22 – Частка ринку (категорія та бренд) Power BI

На третій сторінці звіту Power BI надаються візуальні представлення, включаючи гістограми двох кластер та карту, яка використовуються для відображення вподобань клієнтів на основі їхнього географічного розташування, включаючи місто та регіон. Звіт надає ключову інформацію про верхню частину 10 регіонів і міст, ранжованих за сумою, витраченою клієнтами.

Ці висновки проливають світло на відмінності в уподобаннях клієнтів у різних географічних місцях і можуть бути використані для розробки цільових маркетингових стратегій, спрямованих на конкретні регіони чи міста.

На сторінці 4 звіту можна знайти матрицю, яка надає інформацію щодо загального доходу від продажів і загальної кількості косметичних товарів, проданих за рік, з подальшим розподілом по місяцях. Результати показують, що листопад постійно є місяцем який генерує найвищий дохід (388 млн) за 2019, 2020 та 2021 роки, за яким слідує липень. Крім того, квітень – місяць із найбільшою кількістю продажів за 2022 рік, хоча записи доступні лише за першу половину цього року. У перерахунку на загальну суму кількості проданих товарів, листопад знову стає домінуючим місяцем у 2019 році 2021, тоді як липень і квітень є відповідними місяцями піку для 2020 і 2022 років.

Сезонний характер споживчої поведінки, який відображено у висновках,

пов'язаний з такими подіями, як Великодні канікули та Чорна п'ятниця, оскільки вони мають позитивний вплив на дохід і загальну суму покупок. І навпаки, продажі в літні місяці та під час пандемії Covid-19, коли більшість роздрібних магазинів були закриті, показали зниження загального обсягу продажів.

Зокрема, різке зниження продажів спостерігалось протягом серпня 2019 року, листопада 2020 р. та березня 2021 р.

3.8 Аналіз RFM

Аналіз RFM визначив 10 різних сегментів на основі купівельної поведінки клієнтів. Перші два сегменти, Champions і Loyal, представляють найцінніших клієнтів і зацікавлених клієнтів із загальною кількістю 13 060 (5,3%). Ці клієнти мають високу частоту покупок і давність, що вказує на те, що вони задоволені продуктами чи послугами, які пропонує бізнес. Оскільки косметику зазвичай купують на регулярній основі, наявність міцної бази лояльних клієнтів може забезпечити постійний потік доходу.

Тому важливо підтримувати їхнє задоволення та лояльність, надаючи відмінне обслуговування клієнтів та ексклюзивні пропозиції, винагороди для заохочення повторних покупок.

Сегмент «Потенційні лоялісти» становить 19% населення і складається з клієнтів, які нещодавно були активними, але не часто робили покупки. Цей сегмент дає можливість компанії підвищити лояльність клієнтів, допомогти повернутися, перетворивши їх на лояльних клієнтів, компанія має зосередитися на збільшенні їх частоти покупки та пропозиції програм лояльності, які відповідають їхнім уподобанням і збільшуються їх витрати.

Сегмент нових клієнтів становить значну частину (10,1%) бази клієнтів.

Це клієнти, які зробили свою першу покупку нещодавно і потенційно можуть стати лояльними та прибутковими клієнтами, якщо вони будуть задоволені своїм досвідом покупки. Роздрібна компанія може запропонувати код знижки або безкоштовний подарунок під час наступної покупки, щоб спонукати їх

повернутися. Крім того, компанія може запитати відгук від цих клієнтів, щоб краще зрозуміти їхні потреби та вподобання та використовувати цю інформацію для адаптації своїх пропозицій. Крім того, треба слідкувати за клієнтами після їх першої покупки, щоб показати, що бізнес цінує їхні дії. Це можна просто надіслати електронний лист із подякою або запропонувати опитування для збору відгуків і покращення досвіду клієнтів.

Загалом, надаючи чудове обслуговування клієнтів, пропонуючи персоналізовані акції, надання освітнього та інформаційного контенту про продукти/послуги та подальші дії після покупок фірма може створити лояльність клієнтів і зберегти цей специфічний сегмент.

Обидва сегменти «Перспективний» і «Потребують уваги» займають 11,1% від загальної кількості населення, що представляє можливість для роздрібною компанії повторно залучити цих клієнтів і перетворити їх на лояльних, пропонуючи персоналізовані пропозиції, щоб стимулювати покупки.

Сегмент «Знаходяться в режимі очікування» відображає 7,7% від загальної кількості клієнтів, які не робили покупок деякий час, але раніше були активними. Ці клієнти мають низький індекс частоти та нещодавності, які свідчать про те, що вони ризикують стати неактивними. Бізнес роздрібною торгівлі повинен діяти, щоб залучити цих клієнтів і запобігти їх втрати. Персоналізоване спілкування може бути ефективним, нагадуючи цим клієнтам про бренд і заохочуючи їх повернутися.

Аналіз RFM класифікував 9235 клієнтів у сегмент «Не можна їх втратити» (3,6% із загальної бази), яка включає клієнтів, які дуже зацікавлені в бренді, але нещодавно не робили покупки. Ці клієнти мають індекс високої частоти та індекс нещодавності свідчить про те, що вони лояльні, але ризикують стати неактивними. Роздрібний бізнес повинен зосередитися на утриманні цих клієнтів, пропонуючи персоналізовані рекламні акції та стимули, щоб заохотити їх зробити покупку. Крім того, повинна діяти програма лояльності та винагороди, яка для постійних клієнтів може бути потужним інструментом утримання клієнтів у цьому сегменті.

Клієнти, що перебувають у зоні ризику та сплячому режимі, представляють найменш цінні сегменти із загальною кількістю населення 114 445 (46,6%). Ці

клієнти мають низьку цінність Recency, тому важливо повторно залучати їх персоналізованими пропозиціями та рекламними акціями на основі минулої історії покупок клієнта, щоб запобігти їх відтоку. Бізнес також може вести опитування або відгук, щоб визначити причини їхньої бездіяльності, усунути їх і адаптувати відповідно до стратегії утримання. Проактивне обслуговування клієнтів також може допомогти визначити та вирішити потенційні проблеми, перш ніж вони призведуть до відтоку. Наприклад, першою дією може полягати в тому, щоб охопити клієнтів, чия кількість останніх покупок зменшилася, пропонуючи індивідуальні рішення, такі як різні тарифні плани або додаткові послуги. На рис. 3.23 представлено таблицю балів RFM.

```
> table(RFM_Scores$RF_Scores)

      About To Sleep      At Risk      Cant'lose them      Champions
      18993           63930           9235              3911
      Hibernating      Loyal Customers      Need Attention      New Customer
      50515           9149           8407              24723
      Potential Loyalists      Promising
      46497           10415

> |
```

Рисунок 3.23 – Таблиця балів RFM

Результати за сегментами RFM наведено в таблиці 3.2.

Таблиця 3.2 – Результати за сегментами RFM

Сегмент споживачів	Кількість споживачів
Чемпіони	3911 (1,6%)
Лояльні	9149 (3,7%)
Потенційні лоялісти	46497 (19%)
Нові споживачі	24723 (10,1%)
Перспективні	10415 (4,2%)
Потребують уваги	8407 (3,4%)
Знаходяться в режимі очікування	18993 (7,7%)
Не можна їх втратити	9235 (3,6%)
Під загрозою втрати	63930 (26%)
Бездіяльні споживачі	50515 (20,6%)

ВИСНОВКИ

Отримані результати дослідження підтвердили ефективність використання аналітики даних для прогнозування споживчого попиту на товари та послуги. Застосування методів аналізу даних дозволяє зробити більш точний та обґрунтований прогноз на майбутнє, що допомагає підприємствам оптимізувати свою діяльність та збільшувати ефективність бізнесу.

Завдяки аналізу даних можна виявити тенденції споживчого попиту, ідентифікувати ключові фактори, які впливають на нього, а також прогнозувати майбутні зміни. Це дозволяє підприємствам адаптувати свої стратегії та продукти до потреб споживачів, забезпечуючи конкурентні переваги на ринку.

Аналіз даних стосовно клієнтів революціонізував спосіб ведення бізнесу роздрібними торговцями, пропонуючи цінну інформацію про поведінку та вподобання клієнтів. Сегментуючи клієнтів за допомогою аналізу RFM, роздрібні продавці можуть визначити своїх найцінніших цільових клієнтів за допомогою персоналізованих маркетингових кампаній і покращити рівень утримання клієнтів.

Підхід сегментації забезпечує краще розуміння купівельної поведінки клієнтів, визначення моделей і тенденцій, що дозволяє роздрібним торговцям створювати більш персоналізовані та ефективні маркетингові стратегії.

Використання сегментації клієнтів і RFM-аналізу особливо важливо в секторі роздрібної торгівлі, де конкуренція є жорсткою, а лояльність клієнтів є ключем до успіху. Роздрібні торговці, які аналізують дані клієнтів і розробляють цільові маркетингові стратегії, мають перевагу щодо конкурентоспроможності на ринку, підвищення задоволеності та лояльності клієнтів, що призводить до підвищення доходів і прибутків. Крім того, розуміючи поведінку покупців і їх уподобання, роздрібні продавці можуть оптимізувати управління запасами, покращити ланцюжок поставок і підвищуючи ефективність, що призводить до економії коштів і підвищення прибутковості.

Загалом, інтеграція даних, аналітики клієнтів і сегментації клієнтів у роздрібні операції мають потенціал для революції в галузі. Аналізуючи дані

клієнта, роздрібні продавці можуть краще розуміти своїх клієнтів, покращувати взаємодію з клієнтами та підвищувати прибутковість. Таким чином, роздрібним торговцям необхідно інвестувати в інструменти і методи аналізу даних, такі як аналіз RFM та інші методи сегментації клієнтів, щоб залишатися конкурентоспроможними в галузі роздрібної торгівлі, що постійно змінюється.

Отже, використання аналітики даних для прогнозування споживчого попиту є важливим інструментом управління бізнесом, який дозволяє підприємствам збільшувати свою прибутковість та стабільність на ринку. Результати дослідження свідчать про необхідність впровадження аналітичних методів у практику управління підприємством для забезпечення його успішного розвитку.

ПЕРЕЛІК ПОСИЛАНЬ

1. Abbasi Ahmed, Sarker Suprateek, Chiang Roger HL. Big data research in information systems: Toward an inclusive research agenda. *J Assoc Inf Syst* 2016;17(2):
2. Chen Hsinchun, Chiang Roger HL, Storey Veda C. Business intelligence and analytics: From big data to big impact. *MIS Q* 2012;1165–88.
3. Müller Roland M, Thoring Katja. Design thinking vs. lean startup: A comparison of two user-driven innovation strategies. In: *Leading through design*, Vol. 151. 2012, p. 91–106.
4. Kagermann Henning. Change through digitization—Value creation in the age of Industry 4.0. In: *Management of permanent change*. Springer; 2015, p. 23–45.
5. Edmondson Amy C, Nembhard Ingrid M. Product development and learning in project teams: The challenges are the benefits. *J Prod Innov Manage* 2009;26(2):123–38.
6. Nijssen Edwin J, Frambach Ruud T. Determinants of the adoption of new product development tools by industrial firms. *Ind Mark Manag* 2000;29(2):121–31.
7. Hanelt Andre, Piccinini Everlin, Gregory Robert W, Hildebrandt Björn, Kolbe Lutz M. Digital transformation of primarily physical industries-exploring the impact of digital trends on business models of automobile manufacturers. In: *Wirtschaftsinformatik*. 2015, p. 1313–27.
8. Johanson Mathias, Belenki Stanislav, Jalminger Jonas, Fant Magnus, Gjertz Mats. Big automotive data: Leveraging large volumes of data for knowledge-driven product development. In: *2014 IEEE international conference on big data (Big data)*. IEEE; 2014, p. 736–41.
9. Jones Matthew. What we talk about when we talk about (big) data. *J Strateg Inf Syst* 2019;28(1):3–16.
10. Reinartz Werner, Wiegand Nico, Imschloss Monika. The impact of digital transformation on the retailing value chain. *Int J Res Mark* 2019;36(3):350–66.
11. Micus Christian, Smeets Jasper, Böhm Markus, Krcmar Helmut. Customer inte?gration in product development using big data: An example of using fleet data to determine parking behavior. In: *PACIS*. 2022, p. 160.

12. Davenport TH, Harris JG. Competing on analytics: The new science of winning. Language, 15. Jg., Nr. 217p, S. 24cm, Harvard Business School Press; 2007.

13. Park YoungKi, El Sawy Omar A, Fiss Peer. The role of business intelligence and communication technologies in organizational agility: a configurational approach. J Assoc Inf Syst 2017;18(9):1.

14. Blackhurst Jennifer, Wu Tong, O'Grady Peter. PCDM: a decision support model?ing methodology for supply chain, product and process design decisions. J Oper Manage 2005;23(3–4):325–43.

15. Eppinger Steven, Ulrich Karl. Product design and development. McGraw-Hill Higher Education; 2015. [

16. Dean Jr James W, Sharfman Mark P. Does decision process matter? A study of strategic decision-making effectiveness. Acad Manag J 2017;39(2).

17. Kelleher John D, Mac Namee Brian, D'arcy Aoife. Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies. MIT Press; 2020.

18. Chandler Daniel, Munday Rod. A dictionary of social media. Oxford University Press; 2016.

19. Christopher Andrew N. Interpreting and using statistics in psychological research. SAGE Publications; 2016.

20. Bahar Yelken 2005. ERP System Implementation: A C Beko A.Ė. Ėstanbul Kltr University Faculty Of Engineering & Architecture, Department Of Industrial Engineering.

21. Jiang Yingjie (2005), "Critical Success Factors in ERP Implementation in Finland", M.Sc. Thesis in Accounting, the Swedish School of Economics and Business Administration.

22. Kvavik, R. B., Katz, R. N., Beecher, K., Caruso, J., King, P., Voludakis, J., & Williams, L. A. (2002). The promise and performance of enterprise systems for higher education (ERS0204).

23. Leo Zornada, "Implementing ERP system in higher Education Institutions", 27th Int. Conf. Information Technology Interface ITI 2005, June 20- 23,2005.

ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО - КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ТЕЛЕКОМУНІКАЦІЙ

КВАЛІФІКАЦІЙНА РОБОТА

**ВИКОРИСТАННЯ АНАЛІТИКИ ДАНИХ ДЛЯ
ПРОГНОЗУВАННЯ СПОЖИВЧОГО ПОПИТУ
НА ТОВАРИ ТА ПОСЛУГИ**

виконав студент: ФІЛІПОВ Артем
керівник: КОЗАЧЕНКО С.Я., к.е.н., доцент

1

Мета кваліфікаційної роботи:

покращення ефективності управління споживчим попитом підприємства шляхом використання сучасних методів аналітики даних

Об'єкт дослідження:

методи аналітики даних для прогнозування споживчого попиту на товари та послуги

Предмет дослідження:

споживчий попит на товари та послуги

2

Актуальність:

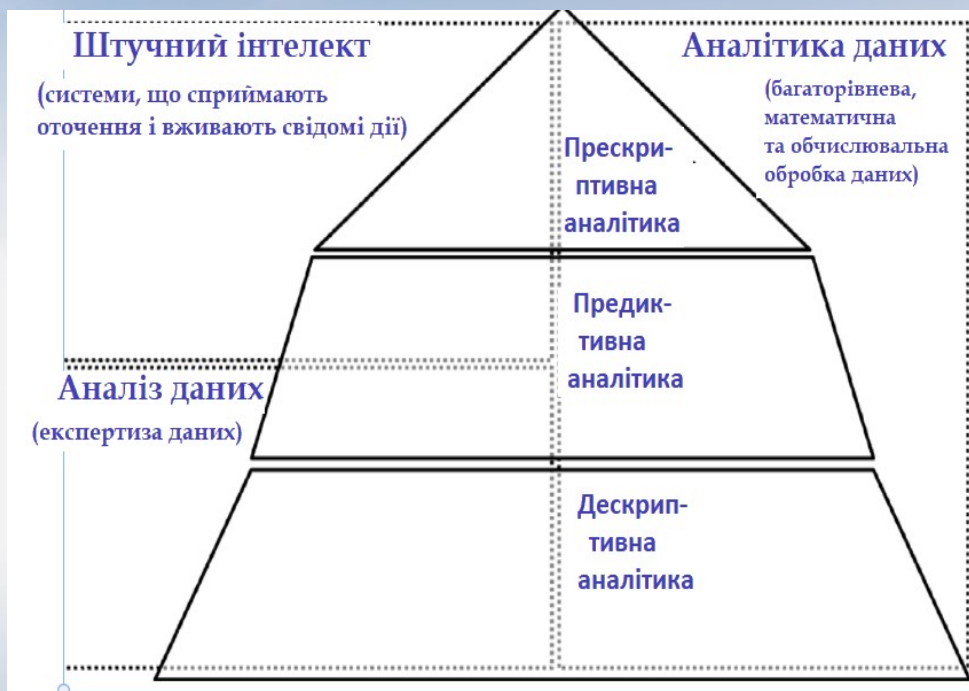
Аналітика даних стає все більш важливою для підприємств у сучасному світі. Прогнозування споживчого попиту на товари та послуги дозволяє підприємствам оптимально планувати виробництво, запаси, маркетингові кампанії та інші стратегічні рішення, що в умовах постійних змін і конкуренції на ринку є дуже актуальним.

За допомогою аналітики даних можна виявити та аналізувати ключові тенденції споживчого ринку, визначити фактори, що впливають на попит, побудувати прогнозні моделі та розробляти стратегії залучення нових клієнтів. При цьому підприємствам потрібно швидко реагувати на зміни в споживчому попиті, щоб забезпечити стабільний розвиток бізнесу.

Таким чином, кваліфікаційна робота на тему "Використання аналітики даних для прогнозування споживчого попиту на товари та послуги" є важливою для практичних застосувань у галузі маркетингу, логістики та стратегічного планування підприємства.

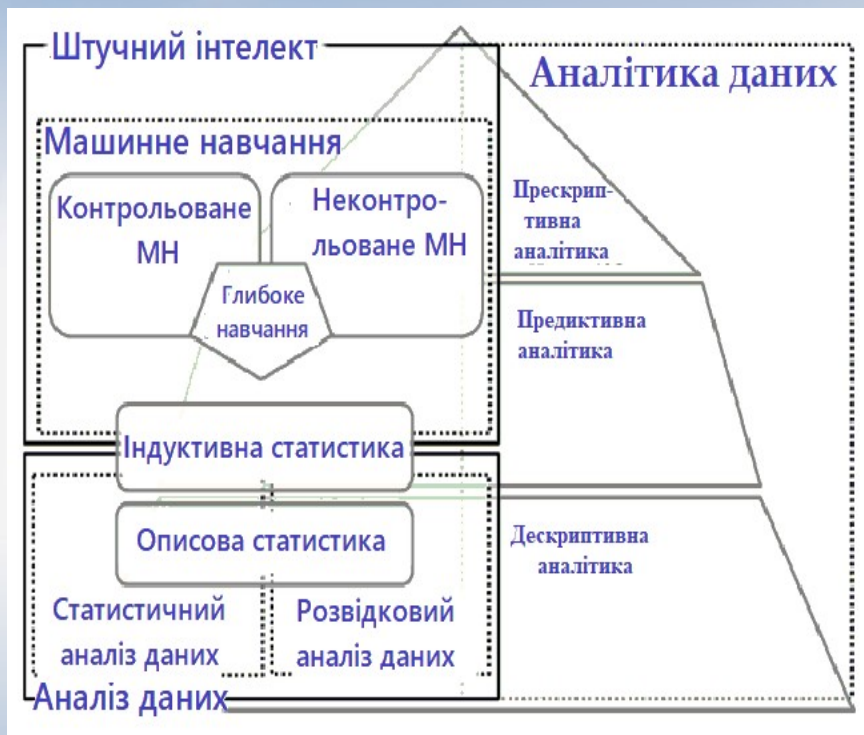
3

Таксономія аналізу даних з трьома рівнями



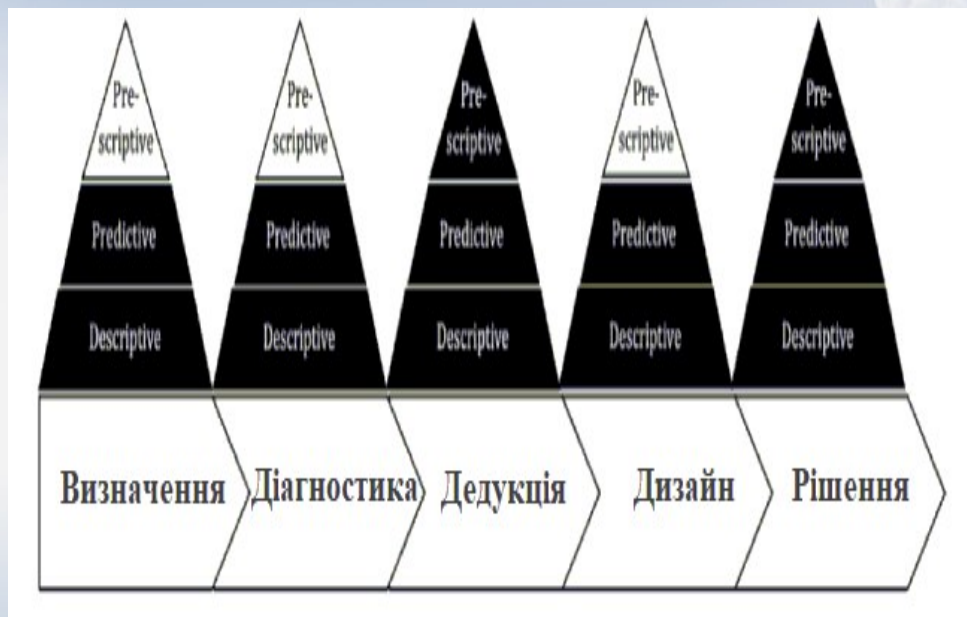
4

Таксономія машинного навчання і аналітики даних



5

Аналітична основа разом із загальним процесом розробки продукту



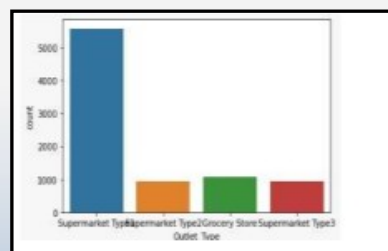
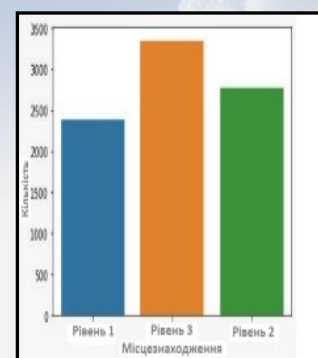
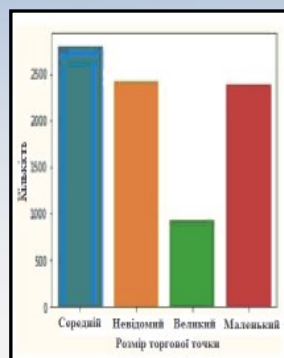
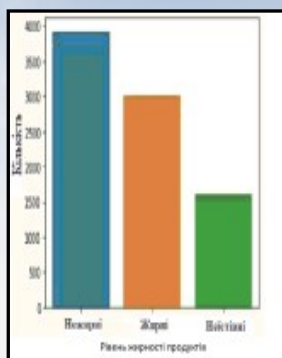
6

Модулі системи аналітики даних



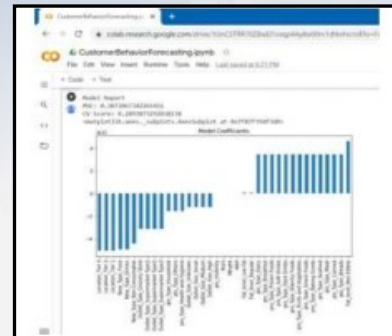
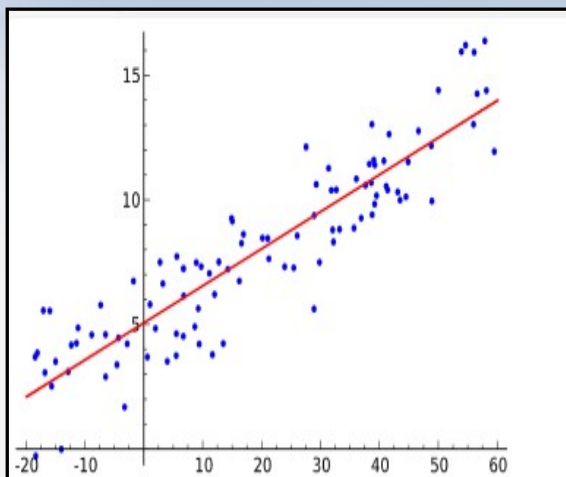
7

Графіки розподілу продуктів за критеріями



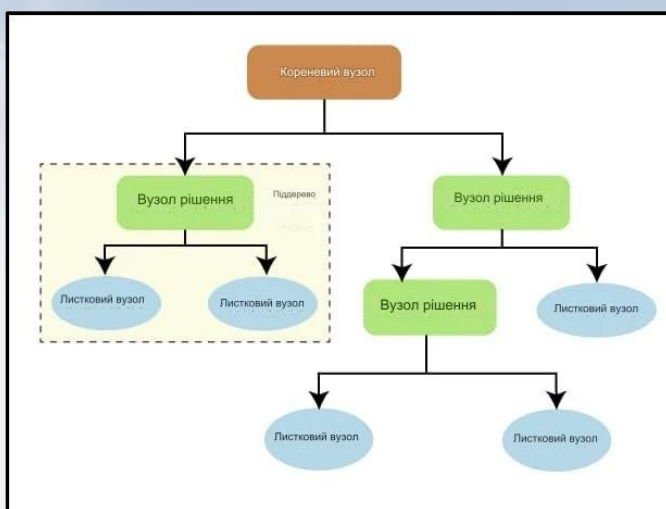
8

Застосування алгоритму лінійної регресії



9

Застосування алгоритму випадкового лісу



10

Сегментація споживачів на основі значень R і F



11

ВИСНОВКИ

- Отримані результати дослідження підтвердили ефективність використання аналітики даних для прогнозування споживчого попиту на товари та послуги. Застосування методів аналізу даних дозволяє зробити більш точний та обґрунтований прогноз на майбутнє, що допомагає підприємствам оптимізувати свою діяльність та збільшувати ефективність бізнесу.
- Завдяки аналізу даних можна виявити тенденції споживчого попиту, ідентифікувати ключові фактори, які впливають на нього, а також прогнозувати майбутні зміни. Це дозволяє підприємствам адаптувати свої стратегії та продукти до потреб споживачів, забезпечуючи конкурентні переваги на ринку.
- Таким чином, використання аналітики даних для прогнозування споживчого попиту є важливим інструментом управління бізнесом, який дозволяє підприємствам збільшувати свою прибутковість та стабільність на ринку. Результати дослідження свідчать про необхідність впровадження аналітичних методів у практику управління підприємством для забезпечення його успішного розвитку.

12