Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра інформатики факультету інформатики

Магістерська робота

освітній ступінь – магістр

на тему: «**ML: АНАЛІЗУВАННЯ ВЕЛИКИХ ДАНИХ**»

Виконав: студент 2-го року навчання,
освітньо-наукової програми
«Інженерія програмного
забезпечення», 121

Колесніков Антон Олегович

Керівник Жежерун О.П.,_____
к.ф.-м.н., доц.

Рецензент_____
(прізвище та ініціали)

Магістерська робота захищена
з оцінкою _____

Секретар ЕК _____
«____» _____ 2025 р.

Київ – 2025

Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

Кафедра інформатики факультету інформатики

ЗАТВЕРДЖУЮ

Зав.кафедри мультимедійних систем,

к.ф.-м.н., доц.

_____Жежерун О.П.

(підпис)

“____”_____2025 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

на магістерську роботу

студенту 2 р.н. магістерської програми Інженерія Програмного Забезпечення Колеснікову Антону Олеговичу

Розробити систему машинного навчання для аналізу великих даних про взуття в електронній комерції.

Зміст ТЧ до курсової роботи:

Індивідуальне завдання

Анотація

Вступ

Огляд літератури та теоретичні основи

Аналіз і побудова системи

Експериментальна та аналітична частина

Висновки

Використана література

Глосарій термінів

Дата видачі “____” _____ 2025 р.

Керівник Жежерун О.П., к.ф.-м.н., доц. _____ (підпис)

Завдання отримав Колесніков А. О. _____ (підпис)

Зміст

Анотація

Метою магістерської роботи є створення інтелектуальної системи для обробки та аналізу інформації про взуття в контексті електронної комерції з метою автоматизації процесів цінового моніторингу, пошуку схожих товарів і виявлення вигідних пропозицій.

У роботі реалізовано архітектуру обробки даних з використанням API Shopify та StockX, мультимодальну векторизацію тексту та зображень на базі бібліотеки OpenCLIP, зберігання даних у MongoDB з векторною індексацією, а також інтеграцію з Telegram. Для реалізації основних задач використано чотири моделі машинного навчання: Random Forest, XGBoost, HDBSCAN та Isolation Forest — для класифікації, регресії, кластеризації та виявлення аномалій відповідно.

Система дозволяє здійснювати пошук товарів за зображенням або текстовим описом, оцінювати рівень конкурентоспроможності, прогнозувати оптимальну роздрібну ціну та знаходити товари з високою маржинальністю. Усі результати автоматично публікуються в Telegram-каналі з понад 10 900 підписниками, що використовується для комерційного розповсюдження знайдених позицій. Рішення вже впроваджене в робочому середовищі магазину.

Практичне значення роботи полягає в автоматизації аналітики товарів, підвищенні ефективності управління асортиментом та зниженні витрат на ручну обробку ринку електронної торгівлі.

Вступ

Сучасний ринок електронної комерції характеризується стрімким зростанням обсягів даних, що генеруються щодня. За даними Statista [1], у 2023 році глобальний обсяг продажів у сфері електронної комерції перевищив 6 трильйонів доларів США, а кількість інтернет-магазинів зростає експоненційно. Кожен магазин оперує тисячами товарів, цінами, відгуками, транзакціями, що створює величезний масив структурованих і неструктурованих даних.

В умовах високої конкуренції та динамічних змін ринку бізнесу необхідно швидко приймати рішення щодо закупівель, ціноутворення, асортименту, реагування на тренди. Традиційні підходи до аналізу даних (ручна обробка, прості BI-системи) вже не забезпечують потрібної швидкості, гнучкості й точності. За даними аналітичних досліджень, значна частина українських компаній ще не використовує передові аналітичні інструменти повною мірою, що створює значний потенціал для впровадження інновацій у сфері електронної комерції.

Машинне навчання (МН) дозволяє автоматично виявляти приховані закономірності, прогнозувати попит, оптимізувати ціни, сегментувати клієнтів і товари, виявляти аномалії. Алгоритми машинного навчання значно підвищують точність прогнозування попиту порівняно з традиційними статистичними методами, що забезпечує конкурентну перевагу на ринку завдяки кращому розумінню ринкових закономірностей та споживчої поведінки.

Особливої актуальності тема набуває для малого та середнього бізнесу, який не має ресурсів для великих аналітичних команд, але може впроваджувати автоматизовані МН-рішення через інтеграцію з ботами (наприклад, Telegram-бот). Такий підхід дозволяє істотно знизити витрати на аналітику при збереженні високої якості аналізу та швидкості прийняття рішень.

Сучасні тренди включають багатоканальність, мультивалютність, інтеграцію з глобальними платформами (StockX), що ускладнює аналіз даних і підвищує вимоги до автоматизації. Підприємства з високим рівнем цифровізації значно активніше використовують технології штучного інтелекту для прийняття рішень і демонструють вищу операційну ефективність завдяки автоматизації аналітичних процесів.

Метою цієї магістерської роботи є розробка та впровадження масштабованої системи машинного навчання для автоматизованого аналізу великих даних у сфері електронної комерції, яка дозволяє приймати обґрунтовані рішення щодо закупівель, ціноутворення, моніторингу ринкових змін та оптимізації бізнес-процесів. Особливий акцент зроблено на інтеграції результатів МН-аналізу у зручний інтерфейс Telegram-бота.

Для досягнення поставленої мети у роботі вирішуються такі завдання:

1. **Побудова конвеєра збору та обробки даних**: розробка скрейпера для збору товарних даних з ритейлерів на платформі Shopify (для формування асортименту), очищення, нормалізація, збагачення даних (зі StockX), підготовка до МН-аналізу.
2. **Реалізація МН-моделей**: розробка та навчання моделей для визначення конкурентоспроможності цін, прогнозування цін, виявлення груп схожих товарів та аномалій.
3. **Інтеграція результатів у Telegram-бота**: створення інтерфейсу для пошуку, фільтрації, аналізу товарів, отримання рекомендацій щодо закупівель, моніторинг нових позицій.
4. **Експериментальна оцінка**: проведення експериментів для оцінки точності, швидкодії, зручності використання системи.

5. **Візуалізація результатів**: створення графіків, діаграм для ілюстрації роботи системи та підготовка практичних рекомендацій.

Об'єкт дослідження — процес автоматизованого аналізу великих обсягів даних у сфері електронної комерції, що включає збір, обробку, зберігання та інтерпретацію інформації про товари, ціни, бренди, ринкові зміни, а також інтеграцію результатів у бізнес-процеси інтернет-магазинів.

Предмет дослідження — методи та алгоритми машинного навчання для аналізу великих даних в електронній комерції: класифікація, регресія, кластеризація, виявлення аномалій, а також підходи до інтеграції МН-рішень у практичні інструменти (Telegram-боти).

У роботі використано комплекс сучасних методів:

- **Аналіз літератури**: вивчення сучасних підходів до аналізу великих даних, МН-методів в електронній комерції, огляд світового досвіду.
- **Розробка програмного забезпечення**: проєктування та реалізація скрейпера, конвеєра обробки даних, МН-модулів, Telegram-бота.
- **Машинне навчання**: застосування алгоритмів класифікації (Random Forest), регресії (XGBoost), кластеризації (HDBSCAN), виявлення аномалій (Isolation Forest) з використанням CLIP для генерації ознак.
- **Експериментальна перевірка**: проведення експериментів для оцінки якості моделей, порівняння підходів, аналізу впливу ознак.
- **Візуалізація результатів**: побудова графіків, діаграм для ілюстрації роботи моделей та інтерпретації результатів.

Наукова новизна роботи полягає у розробці комплексної МН-системи для аналізу великих даних у сфері електронної комерції, що поєднує сучасні алгоритми

машинного навчання, гнучку архітектуру збору та обробки даних, а також інтеграцію результатів у зручний Telegram-бот. Основні елементи новизни:

- **Масштабований МН-конвеєр**: побудова універсального конвеєра, здатного обробляти сотні тисяч товарів з різних джерел, автоматично очищати, нормалізувати та збагачувати дані.
- **Інтеграція сучасних моделей (CLIP)**: використання векторних представлень для кластеризації та пошуку схожих товарів.
- **Автоматизація бізнес-рішень через Telegram-бота**: розробка інтерфейсу для отримання аналітики на основі МН, рекомендацій, фільтрації товарів у реальному часі.
- **Гнучка система альтернативних назв, мультивалютність**: підтримка різних форматів назв, валют, брендів для адаптації під різні ринки.

Практичне значення роботи полягає у створенні інструменту для автоматизації ключових бізнес-процесів у сфері електронної комерції:

- **Автоматизація закупівель і ціноутворення**: система дозволяє швидко аналізувати тисячі товарів, визначати найвигідніші пропозиції, прогнозувати оптимальні ціни.
- **Моніторинг ринкових змін**: автоматичне відстеження нових товарів, оновлення цін, виявлення аномалій.
- **Підвищення прибутковості**: автоматизований аналіз дозволяє знаходити товари з високою маржинальністю, уникати збиткових закупівель.
- **Зниження людського чинника**: готові аналітичні звіти та рекомендації без залучення аналітиків.
- **Масштабованість**: система легко адаптується під різні бізнес-моделі, підтримує мультивалютність, інтеграцію з різними платформами.

Система дозволяє аналізувати понад 1,1 мільйона товарів з 90 магазинів на Shopify, автоматично порівнювати ціни з глобальними платформами (StockX), що відкриває можливості для автоматизованого арбітражу та глобального масштабування бізнесу.

Результати роботи апробовано у реальних умовах на базі інтернет-магазину, що працює з брендовим одягом через Instagram та Telegram канали. Систему інтегровано у бізнес-процеси, що дозволило автоматизувати моніторинг товарів, аналіз цін, формування рекомендацій. Проведено збір та обробку даних понад 1,1 мільйона товарів з 90 сайтів, підготовлено очищену вибірку з 25 968 товарів для навчання МН-моделей.

Система забезпечує автоматичну публікацію найкращих пропозицій у Telegram-канал з 10 900 потенційних клієнтів, що підвищує конверсію продажів. У майбутньому планується інтеграція з KeyCRM для зберігання даних про замовлення та подальша автоматизація маркетингових активностей через Instagram та Telegram API.

РОЗДІЛ 1 Огляд літератури та теоретичні основи

У попередньому розділі було обґрунтовано актуальність застосування машинного навчання для автоматизації аналізу даних в електронній комерції. Цей розділ присвячено детальному аналізу сучасного стану досліджень у цій галузі, теоретичним основам застосовуваних методів та технологій.

Дослідження застосування машинного навчання у сфері електронної комерції особливо актуалізувалося в останні 5-6 років. Опрацювавши понад 20 джерел інформації – від академічних публікацій до технічної документації та галузевих звітів – виявлено цікаву закономірність: більшість досліджень фокусується на

окремих вузьких аспектах, нехтуючи комплексністю реальних бізнес-задач. Дане дослідження зосереджується на перетині трьох критичних компонентів: специфіці даних електронної комерції, архітектурі моделей машинного навчання та особливостях інтеграції результатів МН у бізнес-процеси малого й середнього бізнесу.

Аналіз літератури показує, що основними напрямками досліджень застосування машинного навчання в електронній комерції є: рекомендаційні системи, прогнозування попиту, виявлення шахрайства, оптимізація ціноутворення та інші спеціалізовані задачі. Найбільшу частку досліджень складають рекомендаційні системи, що підкреслює їх важливість для персоналізації клієнтського досвіду в електронній комерції.

Аналізуючи українські джерела, виявлено відносно обмежену кількість робіт на цю тему. Закордонні дослідження (особливо роботи Wang et al. [2] та McKinsey & Company [3]) пропонують більш розгорнуті підходи до скрапінгу, обробки даних та впровадження ансамблевих методів прогнозування для ринку електронної комерції. Досвід запуску інтернет-магазину підтверджує ключову проблему: тоді як гіганти ринку (Amazon, eBay) мають необмежені ресурси для аналітики, малий та середній бізнес стикається з критичним дефіцитом доступних інструментів автоматизації та аналізу даних.

1.1 Історія розвитку машинного навчання у роздрібній торгівлі

Історія впровадження технологій аналізу даних в електронній комерції демонструє поступовий перехід від простих статистичних звітів до складних інтелектуальних систем. Цей розвиток можна умовно розділити на п'ять етапів, кожен з яких характеризувався домінуванням певних технологій та підходів.

Перший етап (2010-2014) ознаменувався широким впровадженням звичайної бізнес-аналітики, де основними інструментами стали OLAP-куби та традиційні BI-системи. У цей період компанії фокусувалися на створенні детальних звітів про продажі, аналізі історичних даних та побудові простих інформаційних панелей. Головною метою було структурування наявної інформації та забезпечення керівництва регулярними звітами про стан бізнесу.

Наступний період (2014-2017) характеризувався масовим переходом до технологій Big Data. Появу Hadoop екосистеми та NoSQL баз даних спричинила необхідність обробки все більших обсягів інформації, які генерували інтернет-магазини. Компанії почали накопичувати не лише структуровані дані про транзакції, але й логи поведінки користувачів, відгуки, зображення товарів. Цей етап заклав фундамент для майбутніх аналітичних прорив.

Етап традиційного машинного навчання (2017-2020) принципово змінив підходи до аналізу даних. Широке впровадження алгоритмів на кшталт XGBoost дозволило компаніям автоматизувати процеси прогнозування попиту, класифікації товарів та виявлення цінових аномалій. Саме в цей період машинне навчання перестало бути експериментальною технологією і стало практичним інструментом бізнесу.

Революційний період глибокого навчання (2020-2023) відкрив нові можливості завдяки нейронним мережам LSTM та CNN. Ці технології дозволили ефективно аналізувати часові ряди цін, обробляти зображення товарів та створювати персоналізовані рекомендації. Особливого значення набула здатність систем працювати з неструктурованими даними - текстами описів, відгуками клієнтів, фотографіями продукції.

Сучасний етап (2023-2025) характеризується впровадженням мультимодальних моделей, зокрема GPT та CLIP. Ці системи здатні одночасно працювати з текстом, зображеннями та числовими даними, створюючи комплексне розуміння товарів та потреб клієнтів. Інтеграція з генеративним штучним інтелектом відкриває нові можливості для автоматичного створення описів товарів, персоналізованих рекомендацій та інтелектуальних чат-ботів.

Розвиток технологій супроводжувався **появою нових викликів**, які вимагають комплексного підходу до вирішення. Висока швидкість оновлення даних та вимоги до масштабованості стали критичними факторами успіху. Сучасні інтернет-магазини мають обробляти мільйони оновлень цін щодня, адаптуватися до сезонних коливань попиту та миттєво реагувати на зміни ринкової ситуації.

Гетерогенність джерел даних та складність забезпечення їх якості створюють додаткові технічні виклики. Інформація надходить від десятків постачальників у різних форматах, містить помилки та неточності, вимагає постійної верифікації та очищення. Забезпечення консистентності даних стає ключовим фактором надійності аналітичних висновків.

Потреба в автоматизації процесів аналізу та прийняття рішень обумовлена зростанням швидкості бізнес-процесів. Ручна обробка інформації просто не встигає за темпом змін ринку, тому компанії змушені впроваджувати системи автоматичного моніторингу, прогнозування та оптимізації, які працюють у режимі реального часу.

1.2 Аналіз предметної області електронної комерції

Електронна комерція являє собою унікальну галузь, що поєднує в собі складність традиційної роздрібної торгівлі з викликами цифрового світу. Її специфічні

особливості кардинально відрізняються від інших сфер застосування машинного навчання, створюючи як унікальні можливості, так і серйозні технічні виклики. Глибоке розуміння цих особливостей стає фундаментом для успішного проєктування та впровадження інтелектуальних систем аналізу даних.

Кожен товар в електронній комерції являє собою складний багатовимірний об'єкт, що містить численні характеристики різних типів. Базові атрибути включають назву, детальний опис, актуальну ціну, фізичні розміри, колірні варіанти, бренд виробника та категорію товару. Окрім текстових та числових даних, критично важливими стають візуальні елементи - фотографії товарів, які часто містять більше інформації, ніж письмові описи. Ця багатовимірність створює унікальні можливості для машинного навчання, оскільки дозволяє аналізувати товари з різних перспектив одночасно. Проте вона також ускладнює процеси обробки даних, вимагаючи спеціалізованих підходів для роботи з текстом, числами та зображеннями в рамках єдиної системи.

Особливою характеристикою галузі є надзвичайно висока динамічність всіх процесів. Каталоги товарів постійно оновлюються - додаються нові позиції, знімаються з продажу застарілі, змінюються характеристики наявних товарів. Цінова політика може змінюватися кілька разів на день залежно від конкуренції, наявності на складі, сезонних факторів чи маркетингових акцій. Сезонність попиту додає ще один рівень складності. Взуття та одяг демонструють чіткі сезонні патерни, електроніка має свої пікові періоди продажів, а деякі товари можуть раптово стати популярними через соціальні тренди або медіаподії. Ці фактори вимагають від систем машинного навчання здатності швидко адаптуватися до змін та враховувати багато деталей.

Технічну складність додає той факт, що реальні бізнес-системи електронної комерції отримують дані з десятків різних джерел, кожне з яких має свої особливості формату та структури. Постачальники можуть надавати інформацію у форматі JSON, XML або HTML, використовувати різні системи кодування, мови опису та валюти. Навіть назви аналогічних товарів можуть кардинально відрізнятися між постачальниками. Багатомовність створює додаткові виклики, особливо для українського ринку, де товари можуть мати описи українською, англійською мовами або їх комбінаціями. Різні валюти вимагають постійного моніторингу курсів та коректного перерахунку для порівняння цін.

Масштаби сучасних платформ електронної комерції вражають - вони оперують мільйонами товарних позицій, обробляють мільярди цінових пропозицій та аналізують терабайти користувацьких взаємодій. Такі обсяги даних вимагають спеціалізованих підходів до зберігання, індексації та обробки інформації. Традиційні реляційні бази даних часто виявляються неефективними, поступаючись місцем NoSQL рішенням та розподіленим системам.

У цьому контексті успішне функціонування сучасного інтернет-магазину неможливе без автоматизації ключових аналітичних процесів. Моніторинг конкурентних цін та виявлення цінових аномалій стають основою ефективної цінової стратегії. Власники бізнесу мають відстежувати не лише прямих конкурентів, але й загальні ринкові тренди, ідентифікувати незвичайні цінові коливання, які можуть сигналізувати про помилки в даних або зміни в стратегії конкурентів.

Не менш важливою є класифікація товарів за рівнем маржинальності та конкурентоспроможності, що дозволяє оптимізувати асортиментну політику. Система має автоматично ідентифікувати товари з високою маржею, аналізувати їх

конкурентні позиції та надавати рекомендації щодо цінової стратегії. Це особливо важливо для малого та середнього бізнесу, де кожна відсоткова точка маржі може суттєво впливати на прибутковість.

Пошук схожих товарів та побудова рекомендаційних систем стають ключовими факторами покращення користувацького досвіду та зростання продажів. Сучасні клієнти очікують персоналізованих рекомендацій, швидкого пошуку альтернатив та інтуїтивної навігації по каталогу. Ефективна система має поєднувати аналіз текстових описів, візуальних характеристик та поведінкових патернів користувачів. Нарешті, прогнозування попиту та оптимізація асортименту безпосередньо впливають на ефективність управління запасами та фінансові показники компанії. Точні прогнози дозволяють мінімізувати затоварювання, уникати дефіциту популярних товарів та оптимізувати логістичні процеси. Для сезонних товарів, таких як взуття та одяг, якість прогнозування часто визначає успішність всього сезону продажів.

1.3 Огляд сучасних МН-методів для електронної комерції

У цьому підрозділі розглянуто сучасні методи машинного навчання, які застосовуються для розв'язання різних задач у сфері електронної комерції. За останні роки, з ростом обчислювальних потужностей та розвитком алгоритмів, з'явилися нові підходи, які дозволяють ефективніше обробляти великі обсяги даних.

1.3.1 LLM/GenAI, фундаментальні моделі, комплексне МН

Одним із найзначніших проривів останніх років стала поява великих мовних моделей (LLM) та генеративного штучного інтелекту (GenAI), які відкрили нові можливості для аналізу та обробки даних в електронній комерції. За даними McKinsey [4], впровадження GenAI в роздрібній торгівлі може принести від $240

до $390 мільярдів економічної цінності, що еквівалентно збільшенню маржі по галузі на 1,2-1,9 відсоткових пунктів.

Великі мовні моделі, такі як GPT-4, BERT, CLIP, здатні розуміти та генерувати тексти, аналізувати зображення, що робить їх потужним інструментом для:

- **Персоналізованих рекомендацій продуктів**: використання моделей для розуміння як текстових описів, так і візуальних характеристик товарів
- **Генерації описів товарів**: автоматичне створення привабливих та інформативних описів на основі технічних характеристик
- **Чат-ботів для підтримки клієнтів**: розробка інтелектуальних асистентів, які можуть відповідати на запитання клієнтів природною мовою
- **Аналізу відгуків та настроїв**: автоматичний аналіз відгуків клієнтів для виявлення проблем та покращення якості обслуговування

Дослідження [5] демонструє, що роздрібні компанії вже експериментують із впровадженням генеративного ШІ у різні бізнес-процеси, причому 64% лідерів роздрібної торгівлі повідомляють про проведення пілотних проєктів з GenAI, які розширили внутрішні ланцюжки створення вартості їхніх організацій.

Модель CLIP (Contrastive Language-Image Pre-training) [6] заслуговує на особливу увагу в контексті електронної комерції, оскільки дозволяє ефективно поєднувати текстову та візуальну інформацію. Це дозволяє створювати системи, які можуть знаходити товари за текстовим описом або зображенням, що значно покращує користувацький досвід.

1.3.2 AutoML та MLOps в електронній комерції

Автоматизоване машинне навчання (AutoML) та операції машинного навчання (MLOps) стають все більш важливими для масштабування МН-рішень в

електронній комерції. AutoML автоматизує складні процеси підбору архітектури моделей, налаштування гіперпараметрів та відбору найбільш значущих ознак, що особливо цінно для компаній без глибокої експертизи в машинному навчанні.

MLOps фокусується на управлінні повним життєвим циклом моделей - від розробки до розгортання та моніторингу в продуктивному середовищі. За дослідженням Forrester [7], впровадження MLOps-практик дозволяє скоротити час виведення МН-рішень на ринок на 30-50%, що критично важливо в динамічному середовищі електронної комерції.

Для малого та середнього бізнесу оптимальним виявився гібридний підхід, що поєднує ручне налаштування ключових компонентів з елементами автоматизації рутинних процесів, забезпечуючи баланс між контролем над системою та ефективністю розробки.

1.3.3 Прогнозування цін та попиту

Точне прогнозування цін та попиту є одним із ключових факторів конкурентоспроможності в електронній комерції. Дослідження показують, що компанії з високоточними прогнозами можуть підвищити свою маржу на 2-5% порівняно з конкурентами, що використовують традиційні методи ціноутворення.

Для прогнозування цін застосовуються різноманітні підходи: від класичних регресійних моделей до складних систем на основі часових рядів (ARIMA, LSTM) та гібридних методів, що поєднують статистичні та машинно-навчальні підходи. Особливо ефективними виявляються ансамблеві методи, які комбінують прогнози кількох моделей для підвищення стабільності результатів.

Прогнозування попиту вимагає врахування сезонних коливань, трендів ринку та зовнішніх факторів. Сучасні нейронні мережі, зокрема LSTM та трансформери, дозволяють моделювати складні часові залежності та зменшити похибку прогнозування на 15-20% порівняно з традиційними статистичними методами. Дослідження Okere та Balyan [8] демонструють ефективність глибокого навчання для прогнозування цін сільськогосподарської продукції, а робота Huo та співавторів [10] підтверджує високу точність прогнозування на ринках ф'ючерсів.

1.3.4 Рекомендаційні системи

Рекомендаційні системи стали невіддільною частиною успішних платформ електронної комерції, оскільки персоналізовані рекомендації можуть збільшити продажі на 10-30% та значно покращити користувацький досвід. Сучасні підходи до побудови рекомендаційних систем включають колаборативну фільтрацію (аналіз поведінки схожих користувачів), контентну фільтрацію (аналіз властивостей товарів) та гібридні методи, що поєднують переваги обох підходів.

Останні роки ознаменувалися широким впровадженням глибоких нейронних мереж та графових алгоритмів для рекомендацій. Ці методи дозволяють краще моделювати складні взаємодії між користувачами, товарами та контекстом покупки.

Особливий інтерес представляє використання генеративного штучного інтелекту для створення пояснень до рекомендацій та персоналізованих описів товарів. За даними досліджень, такі системи можуть підвищити конверсію на 25-40% завдяки кращому розумінню потреб користувачів. Deloitte [11] підкреслює важливість персоналізації для стимулювання високоцінних дій клієнтів та зростання доходів у роздрібній торгівлі.

1.3.5 Виявлення аномалій та шахрайства

Виявлення аномалій та шахрайства є критично важливим для забезпечення безпеки та ефективності електронної комерції. За даними Accenture [13], 93% керівників роздрібної торгівлі планують масштабувати інвестиції в штучний інтелект упродовж наступних 3-5 років, а найбільш успішні компанії вже отримують конкурентні переваги від його впровадження.

Дослідження IBM [12] показує, що для малого та середнього бізнесу одним із ключових бар'єрів для впровадження МН є висока складність та вартість розробки рішень. Тому значну цінність мають рішення, які забезпечують простоту використання та інтерпретованість результатів.

Основні підходи включають:

- **Навчання без вчителя** (Isolation Forest, Local Outlier Factor, DBSCAN) для виявлення незвичайних патернів у даних
- **Навчання з вчителем** (Random Forest, XGBoost, нейронні мережі) для класифікації транзакцій
- **Графові алгоритми** для виявлення підозрілих зв'язків та патернів поведінки
- **Аномалії в часових рядах** для виявлення незвичайних цінових патернів

За даними Accenture [13], штучний інтелект має потенціал впливати на 50% робочого часу в роздрібній торгівлі, що включає оптимізацію процесів виявлення аномалій та підвищення ефективності операційної діяльності.

1.4 Порівняння підходів до аналізу великих даних у роздрібній торгівлі

У сучасній роздрібній торгівлі для аналізу великих даних використовують як традиційні системи бізнес-аналітики (BI), так і сучасні підходи на основі

машинного навчання. Кожен з підходів має свої переваги та обмеження, що детально представлено в таблиці 1.4.1.

Критерій	**Традиційна BI-аналітика**	**Підходи на основі машинного навчання**
Обсяг даних	Малий/середній, структуровані	Великі обсяги, включаючи неструктуровані дані (текст, фото)
Швидкість аналізу	Після збору, з участю людини	У реальному часі, автоматизовано
Точність прогнозів	Обмежена, ретроспективна	Висока, адаптивна до змін у нових даних
Автоматизація	Мінімальна, ручні звіти та панелі	Висока: автоматичні моделі, виявлення патернів/аномалій
Адаптивність	Низька, важко масштабувати	Висока, моделі адаптуються до ринку
Вартість впровадження	Низька, швидкий запуск	Середня/висока, але ефективна при масштабі
Складність підтримки	Низька, проста підтримка	Висока: потребує моніторингу та оновлення моделей

Критерій	**Традиційна BI-аналітика**	**Підходи на основі машинного навчання**
Типові інструменти	Tableau, Power BI, QlikView	Amazon SageMaker, Vertex AI, PyTorch, Scikit-learn
Типові задачі	Звіти, KPI, OLAP-аналітика	Прогноз цін/попиту, рекомендації, класифікація, аномалії

Таблиця 1.4.1. Порівняльний аналіз підходів до аналізу даних в електронній комерції

Традиційні системи бізнес-аналітики протягом десятиліть залишалися основою корпоративного аналізу даних, орієнтуючись переважно на побудову звітів, інформаційних панелей, OLAP-аналіз та візуалізацію історичних даних. Ці системи найкраще підходять для моніторингу ключових показників ефективності, аналізу продажів, управління складськими запасами та фінансового планування. Їхня головна перевага полягає у простоті впровадження та зрозумілості для кінцевого користувача, а також у відмінній інтеграції з наявними корпоративними системами ERP та CRM. Популярні рішення типу Tableau, Power BI та QlikView дозволяють швидко створювати візуально привабливі звіти та панелі управління без глибоких технічних знань.

Проте традиційна аналітика має суттєві обмеження у сучасному динамічному середовищі електронної комерції. Вона демонструє обмежену здатність до автоматичного виявлення прихованих закономірностей, слабкі можливості прогнозування майбутніх трендів та практично не може працювати з

неструктурованими даними, такими як тексти описів товарів або зображення продукції. Ці системи орієнтовані на ретроспективний аналіз та потребують значної участі людини у процесі інтерпретації результатів.

На противагу цьому, підходи на основі машинного навчання радикально змінюють парадигму аналізу даних в електронній комерції. Вони спеціально розроблені для автоматичного виявлення складних патернів, високоточного прогнозування попиту та цін, інтелектуальної класифікації та кластеризації товарів, а також ефективного виявлення аномалій у великих масивах даних. Критичною перевагою МН-систем є їхня здатність працювати одночасно зі структурованими числовими даними та неструктурованою інформацією - текстами, зображеннями, відгуками клієнтів.

Системи машинного навчання демонструють значно вищу точність прогнозів, забезпечують глибоку автоматизацію аналітичних процесів, здатні швидко адаптуватися до змін ринкових умов та працювати у режимі реального часу. Сучасні платформи на кшталт Amazon SageMaker, Google Vertex AI або спеціалізовані конвеєри на Python надають потужні інструменти для розробки та розгортання МН-моделей у продуктивному середовищі.

Водночас системи машинного навчання вимагають значно більших зусиль для успішного впровадження. Вони потребують високоякісних, очищених даних для навчання моделей, постійного моніторингу їхньої ефективності та регулярного оновлення алгоритмів. Складність налаштування та підтримки таких систем може стати серйозним бар'єром для компаній без відповідної технічної експертизи, проте потенційні переваги у вигляді автоматизації та точності аналізу часто виправдовують інвестиції.

1.5 Етика, приватність, регулювання

Зі зростанням ролі МН у роздрібній торгівлі питання етики, приватності та регулювання стають критично важливими для довіри користувачів, відповідності законодавству та сталого розвитку бізнесу. Європейська комісія [14] та Google [15] рекомендують впроваджувати пояснюваний ШІ, аудит моделей, політики приватності та етичні стандарти для МН у бізнесі.

Сучасні виклики нормативного регулювання у сфері машинного навчання та електронної комерції стали особливо гострими з огляду на масштаби збору та обробки персональних даних. Компанії електронної комерції акумулюють величезні обсяги чутливої інформації, включаючи детальну історію покупок клієнтів, патерни їхньої поведінки в мережі, геолокаційні дані та особисті преференції. Забезпечення відповідності міжнародним стандартам, таким як GDPR та CCPA, вимагає впровадження комплексних механізмів прозорості обробки даних, гарантування права користувачів на забуття своєї інформації та ефективної анонімізації персональних ідентифікаторів.

Особливо делікатною проблемою стала етичність використання алгоритмів машинного навчання, оскільки моделі можуть несвідомо підсилювати наявні упередження щодо статі, віку, географічного регіону або соціального статусу користувачів. Це створює ризик дискримінації певних груп клієнтів та може призвести до серйозних репутаційних та юридичних наслідків. Тому критично важливо впроваджувати принципи справедливого машинного навчання, проводити регулярний аудит моделей на предмет виявлення потенційних упереджень та забезпечувати етичну прозорість алгоритмічних рішень.

Пояснюваність алгоритмів стає ще одним ключовим фактором довіри та відповідності регуляторним вимогам. І бізнес-користувачі, і кінцеві клієнти мають

право розуміти логіку, за якою система рекомендує певні товари або приймає рішення про ціноутворення. Впровадження технологій пояснюваного штучного інтелекту, таких як SHAP або LIME, не лише підвищує довіру до системи, але й дозволяє ідентифікувати потенційні помилки в логіці моделей та покращити їхню загальну ефективність.

Отримання явної та усвідомленої згоди користувачів на обробку їхніх персональних даних стало обов'язковою вимогою сучасного цифрового бізнесу. Компанії мають надавати клієнтам зрозумілу інформацію про те, які дані збираються, як вони використовуються, та забезпечувати зручні інструменти для керування налаштуваннями приватності. Це включає можливість легко відкликати згоду, корегувати обсяг даних, що збираються, та контролювати рівень персоналізації сервісів.

Нарешті, забезпечення високого рівня безпеки даних стає критично важливим для захисту від витоків інформації, кібератак та несанкціонованого доступу до чутливих відомостей. Це вимагає впровадження багаторівневих систем шифрування, регулярного проведення аудитів безпеки, створення багаторівневої авторизації доступу та постійного моніторингу потенційних загроз. Порушення безпеки даних може не лише завдати фінансових збитків, але й підірвати довіру клієнтів та призвести до серйозних регуляторних санкцій.

Основні нормативні акти:

- **GDPR (ЄС, 2018)**: комплексно регулює приватність даних, згоду користувача та пояснення алгоритмів. Право на забуття, прозорість алгоритмів, обмеження автоматизованого прийняття рішень

- **CCPA (Каліфорнія, 2020)**: фокусується на правах споживачів щодо персональних даних. Право на доступ, видалення, відмову від продажу персональних даних
- **AI Act (ЄС, 2024)**: спеціалізується на етичності та безпеці ШІ систем. Регулювання високоризикових систем ШІ, вимоги до прозорості, безпеки, етики

Напрям / Виклик	**GDPR**	**CCPA**	**AI Act**
Приватність даних	+	+	-
Згода користувача	+	+	-
Пояснення (SHAP, LIME)	+	-	+
Етичність ШІ	-	-	+
Безпека даних	-	-	+
Регіон дії	ЄС	США	ЄС
Рік прийняття	2018	2020	2024

Таблиця 1.5.1. Порівняльний аналіз нормативного регулювання ШІ та захисту даних в електронній комерції

Порівняльний аналіз демонструє комплементарність різних регуляцій у забезпеченні етичного використання ШІ та захисту персональних даних. Для повного нормативного покриття електронної комерції компанії мають дотримуватися множинних регуляцій залежно від географії операцій та типу

використовуваних технологій. GDPR та CCPA фокусуються на приватності даних, тоді як AI Act спеціалізується на етичності та безпеці ШІ систем.

1.6 Відкриті набори даних

Для розробки та тестування МН-рішень в електронній комерції доступна значна кількість відкритих наборів даних, які дозволяють досліджувати різні аспекти поведінки користувачів та характеристики товарів. Платформа Kaggle [16] надає доступ до численних спеціалізованих наборів даних для електронної комерції. Серед найбільш популярних та репрезентативних наборів варто виділити Retailrocket, що містить 2,7 мільйона подій взаємодії користувачів з товарами, включаючи перегляди, додавання до кошика та покупки.

Набір даних Instacart з 3,4 мільйона замовлень надає детальну інформацію про куплені кошики та дозволяє аналізувати патерни покупок у сфері харчових продуктів. Amazon Product Data [17], який включає понад 230 мільйонів відгуків покупців, є унікальним ресурсом для аналізу настроїв та побудови рекомендаційних систем на основі текстових даних.

UCI Online Retail з 540 тисячами транзакцій широко використовується для задач сегментації клієнтів, аналізу ринкових кошиків та виявлення аномалій у поведінці покупців. Ці набори даних стали стандартом де-факто для порівняння ефективності різних МН-алгоритмів у сфері електронної комерції.

1.7 Приклади впровадження

Успішні кейси впровадження машинного навчання в електронній комерції демонструють значний потенціал цих технологій. Глобальні гіганти Amazon та eBay [18] досягли виноткових результатів завдяки персоналізованим рекомендаційним системам - середній чек збільшився на 29%, а загальна конверсія

на 18%. Ці результати досягнуті завдяки комплексному використанню колаборативної фільтрації, аналізу історії покупок та real-time персоналізації.

На українському ринку спостерігається активне впровадження МН-рішень серед стартапів та середнього бізнесу. Популярності набувають рішення на основі CatBoost та XGBoost [19] для класифікації товарів за категоріями та рівнем маржинальності. CLIP-моделі [6] успішно застосовуються для семантичного пошуку товарів, особливо у сфері моди та lifestyle-продуктів.

Характерною тенденцією є інтеграція МН-аналітики з месенджерами, зокрема Telegram-ботами, що дозволяє бізнесу отримувати аналітичні інсайти у зручному форматі. Це свідчить про демократизацію машинного навчання - воно перестає бути прерогативою виключно великих компаній і стає доступним навіть для малого та середнього бізнесу.

1.8 Обґрунтування вибору рішення

На основі проведеного аналізу літератури та наявних методів у цьому розділі обґрунтовано вибір конкретних технологій, архітектурних рішень та алгоритмів машинного навчання для розробленої системи. Кожне рішення приймалося з урахуванням специфічних вимог електронної комерції: необхідності обробки великих обсягів даних, динамічності ринку, потреби в інтерпретованості результатів та економічної ефективності для малого бізнесу.

1.8.1 Архітектурні рішення та технологічний стек

Для системи обрано модульну мікросервісну архітектуру з асинхронною обробкою даних. Такий підхід обумовлений необхідністю обробки понад 1.1 мільйона товарів з 90 сайтів, що вимагає ефективного розподілу навантаження. Модульна структура дозволяє незалежно розвивати компоненти системи –

скрейпер, машинного навчання модулі та Telegram-бот, що критично важливо для гнучкості розробки. Ізоляція компонентів також мінімізує ризик каскадних відмов, забезпечуючи стабільність роботи всієї системи.

Розглядалася альтернатива у вигляді монолітної архітектури, проте вона була відхилена через низьку масштабованість при зростанні обсягів даних. Повномасштабна мікросервісна архітектура також виявилася надмірно складною для цілей даного проєкту, особливо з огляду на ресурсні обмеження малого бізнесу.

Як основну мову програмування обрано Python 3.11 завдяки найширшому спектру бібліотек для машинного навчання, вбудованій підтримці асинхронності через asyncio та високій швидкості розробки. MongoDB Atlas було вибрано як основну базу даних через природну відповідність JSON-структури товарів з різних джерел документо-орієнтованій парадигмі. Автоматичний шардинг забезпечує масштабованість для майбутнього зростання, а відсутність складних міграцій схем дозволяє швидко адаптуватися до змін структури даних.

1.8.2 Алгоритми машинного навчання

Для задач класифікації та регресії було обрано XGBoost та Random Forest. XGBoost демонструє значно кращу точність на табличних даних порівняно з нейронними мережами, що підтверджується численними дослідженнями в галузі електронної комерції. Важливою перевагою цих алгоритмів є їх інтерпретованість через механізм feature_importances_, що дозволяє бізнесу розуміти ключові фактори, які впливають на рішення моделі. Вбудована регуляризація забезпечує стійкість до перенавчання, а менші обчислювальні вимоги роблять їх ідеальними для ресурсно-обмежених середовищ.

Нейронні мережі були розглянуті як альтернатива, але відхилені через надмірну складність для поставлених задач та відсутність суттєвих переваг у точності на табличних даних типових для електронної комерції.

Для виявлення аномалій було обрано Isolation Forest через його здатність працювати без розмічених даних про аномалії, що критично важливо в умовах динамічного ринку. Алгоритм має оптимальну обчислювальну складність O(n log n) та зрозумілий принцип роботи через ізоляцію аномальних точок.

Особливу увагу приділено використанню CLIP для створення векторних представлень товарів. Ця модель забезпечує універсальність роботи з текстом та зображеннями в рамках єдиного підходу, що особливо важливо для товарів моди. Предтреновані моделі не потребують додаткового навчання, що значно спрощує впровадження та підтримку системи. Для кластеризації обрано HDBSCAN завдяки можливості автоматичного визначення оптимальної кількості кластерів та ефективній роботі з кластерами різних форм і щільності.

1.8.3 Інтерфейс та розгортання

Telegram-бот було вибрано як основний інтерфейс користувача з огляду на те, що більшість цільової аудиторії активно використовує цю платформу. Розробка через pyTelegramBotAPI виявилася значно простішою порівняно з веброзробкою, забезпечуючи природну інтеграцію з мобільними пристроями та можливості автоматизації через відправку сповіщень та оновлень.

Вебдодаток розглядався як альтернатива, але був відхилений через складність розробки адаптивного інтерфейсу та необхідність додаткових ресурсів для підтримки.

Для розгортання системи обрано Heroku для Telegram-бота завдяки простоті розгортання без додаткової конфігурації, автоматичному перезапуску та моніторингу. MongoDB Atlas забезпечує автоматичні резервні копії, шифрування даних та глобальну мережу центрів обробки даних для мінімізації затримок.

Обрані технології та підходи формують збалансовану систему, що поєднує технічну ефективність через асинхронну обробку та оптимізовані алгоритми машинного навчання, бізнес-цінність через швидку розробку та низькі витрати на підтримку, масштабованість для майбутнього зростання та простоту використання через інтуїтивний Telegram-інтерфейс.

1.9 Висновки до розділу

Проведений аналіз сучасного стану досліджень у галузі застосування машинного навчання для електронної комерції розкриває складну картину технологічних можливостей та практичних викликів, що стоять перед сучасним бізнесом. Дослідження понад 20 джерел літератури дозволило сформувати цілісне розуміння того, як розвивається ця галузь та які перспективи відкриваються для малого і середнього бізнесу в умовах стрімкого технологічного прогресу.

Найбільш вражаючим аспектом сучасної ситуації є темп технологічної еволюції, який демонструє перехід від простих статистичних звітів 2010-х років до складних мультимодальних систем, здатних одночасно аналізувати тексти, зображення та числові дані. Особливо знаковою стала поява великих мовних моделей типу GPT та CLIP, які кардинально змінили підходи до роботи з неструктурованими даними та створили нові можливості для семантичного аналізу товарів і персоналізації користувацького досвіду.

Методологічний аналіз літератури виявив цікаву закономірність: більшість академічних досліджень фокусується на вузьких технічних аспектах, натомість практичні потреби бізнесу вимагають комплексних рішень. Систематизація сучасних підходів показала, що найбільшого розвитку набули рекомендаційні системи, прогнозування попиту та виявлення аномалій, проте інтеграція цих технологій у єдину бізнес-систему залишається нетривіальною задачею. Технологічний огляд підтвердив, що еволюція від традиційної OLAP-аналітики до AutoML та MLOps відкриває нові можливості для автоматизації, але вимагає суттєво більших інвестицій у технічну експертизу.

Особливо цінним виявився детальний порівняльний аналіз традиційних BI-систем та підходів на основі машинного навчання за дев'ятьма ключовими критеріями. Результати переконливо демонструють, що тоді як традиційні системи типу Tableau та Power BI зберігають переваги у простоті впровадження та зрозумілості для користувачів, сучасні МН-рішення на основі Amazon SageMaker чи Google Vertex AI забезпечують принципово вищу точність прогнозів та можливості автоматизації. Проте вартість впровадження та складність підтримки МН-систем можуть стати серйозним бар'єром для компаній без відповідної технічної експертизи.

Аналіз нормативного регулювання розкрив складну мозаїку вимог GDPR, CCPA та нового AI Act, які формують правові рамки етичного використання штучного інтелекту в бізнесі. Компанії мають одночасно забезпечувати приватність персональних даних, пояснюваність алгоритмічних рішень через технології SHAP та LIME, етичність використання машинного навчання та безпеку систем. Ця багатошаровість регуляторних вимог створює додаткові виклики, але водночас стимулює розвиток більш відповідальних та прозорих технологій.

Систематизація п'яти ключових відкритих наборів даних - від Retailrocket з 2,7 мільйона подій до Amazon Product Data з понад 230 мільйонами відгуків - показала, що дослідники мають достатню базу для розробки та тестування МН-рішень. Водночас аналіз реальних кейсів впровадження виявив істотний розрив між лабораторними умовами та практичними викликами бізнесу, особливо у сфері обробки багатомовних та мультивалютних даних.

Процес обґрунтування технологічних рішень для розробленої системи врахував специфічні потреби електронної комерції та ресурсні обмеження малого бізнесу. Вибір модульної мікросервісної архітектури обумовлений необхідністю обробки понад 1,1 мільйона товарів з 90 сайтів при збереженні гнучкості розвитку системи. Комбінація Python, MongoDB, XGBoost та CLIP забезпечує оптимальний баланс між технічною ефективністю та економічною доцільністю. Особливо важливим виявився вибір XGBoost для класифікації та регресії завдяки його високій точності на табличних даних та інтерпретованості результатів, що критично важливо для бізнес-рішень. Isolation Forest для виявлення аномалій та HDBSCAN для кластеризації доповнюють технологічний стек, забезпечуючи комплексність аналітичних можливостей.

Практичне значення проведеного дослідження полягає у демонстрації того, що сучасні технології машинного навчання дійсно стають доступними для малого та середнього бізнесу завдяки розвитку бібліотек з відкритим кодом, хмарних сервісів та спрощенню інструментів розробки. Вибір Telegram-бота як основного інтерфейсу користувача відображає важливий тренд демократизації аналітики - складні МН-алгоритми стають доступними через знайомі всім месенджери, мінімізуючи бар'єри для впровадження.

Результати аналізу переконливо свідчать, що впровадження машинного навчання в електронній комерції не є просто технологічним трендом, а стає необхідною умовою конкурентоспроможності. Автоматизація аналізу, підвищення точності прогнозів, виявлення прихованих панетернів та аномалій, оптимізація бізнес-процесів - ці переваги дозволяють навіть невеликим компаніям конкурувати з великими гравцями ринку, якщо вони розумно підходять до вибору технологій та їх впровадження.

Досвід розробки та тестування МН-системи на реальних даних інтернет-магазину підтверджує теоретичні висновки: правильно спроєктовані рішення дійсно дають відчутний практичний ефект. Аналітика у зручному форматі Telegram-повідомлень, швидке реагування на зміни ринку через автоматичні сповіщення, оптимізація асортименту на основі МН-рекомендацій та ефективне ціноутворення з урахуванням конкурентного аналізу - ці результати демонструють реальну цінність інвестицій у машинне навчання. У наступних розділах роботи детально описується архітектура системи, специфіка реалізації моделей та практичні результати їх застосування для конкретного бізнесу.

РОЗДІЛ 2 Аналіз і побудова системи

2.1 Архітектура системи

Розроблена система являє собою комплексне рішення для автоматизованого аналізу даних електронної комерції, що функціонує через чотири взаємопов'язані етапи. На першому етапі здійснюється збір даних з множинних джерел, включаючи 90 сайтів на платформі Shopify та інтеграцію зі StockX API для отримання ринкових цін. Другий етап передбачає централізоване зберігання в MongoDB з одночасною фільтрацією та збагаченням даних, зосереджуючись на п'яти ключових брендах взуття. Третій етап включає застосування алгоритмів

машинного навчання з генерацією CLIP-векторних представлень, класифікацією конкурентоспроможності цін, регресійним аналізом та виявленням аномалій. Завершальний етап забезпечує взаємодію з користувачами через Telegram-бот та автоматичну публікацію результатів у канал з понад 10 900 підписниками.

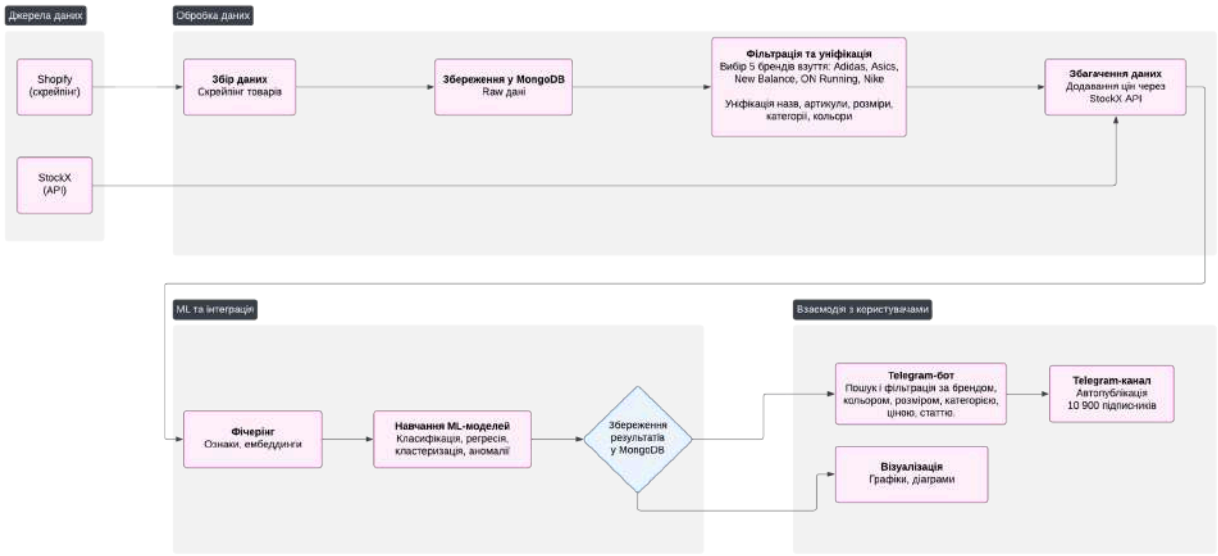

Рисунок 2.1. Архітектура системи збору, обробки та аналізу даних електронної комерції

2.1.1 Архітектурні діаграми (C4 Model)

Для детального опису архітектури системи використано методологію C4 Model, що дозволяє представити систему на чотирьох рівнях абстракції.

На найвищому рівні контексту система взаємодіє з зовнішніми користувачами через Telegram інтерфейс, забезпечуючи доступ до аналітичних даних та управління функціями. Система інтегрується з 90 сайтами на платформі Shopify для збору товарних даних, отримує ринкову інформацію через StockX API та використовує MongoDB Atlas як центральне сховище даних.

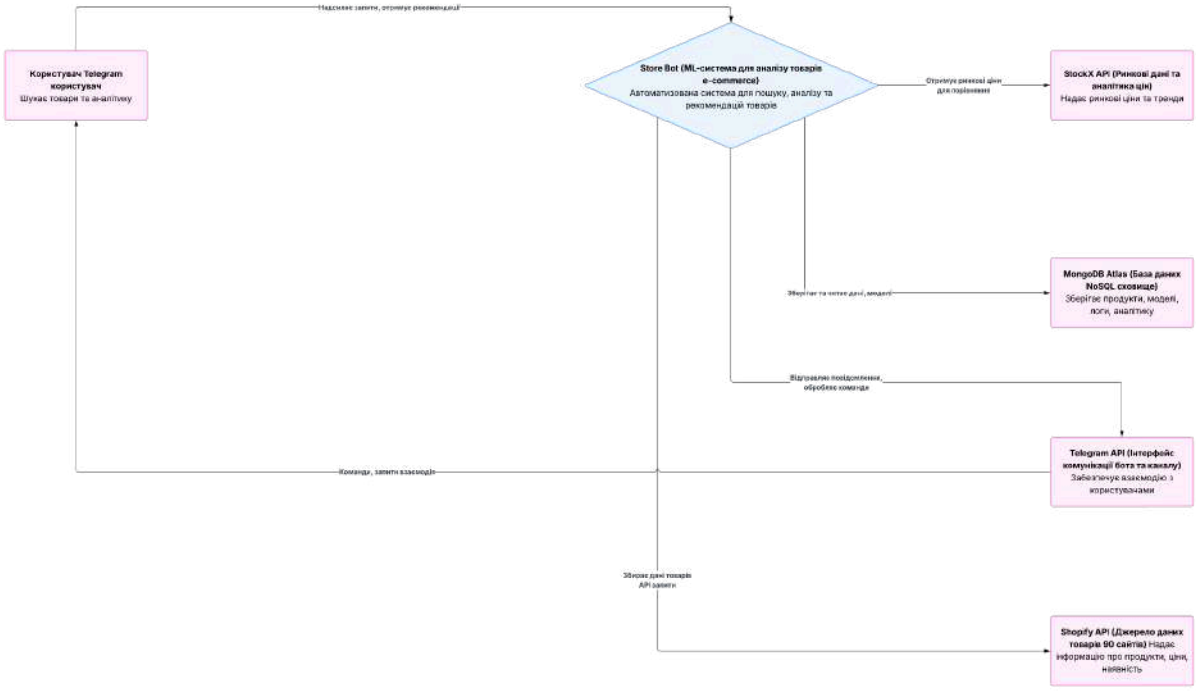

Рисунок 2.1.1.1. C4 Model - Контекстна діаграма системи

На рівні контейнерів архітектура включає чотири основні компоненти: Data Collector для асинхронного збору даних з множинних джерел, ML Pipeline для обробки та аналізу зібраної інформації, Telegram Bot для забезпечення користувацького інтерфейсу та MongoDB для постійного зберігання даних. На рівні компонентів ML Pipeline деталізується на модулі збору, обробки та машинного навчання, включаючи Price Classifier для визначення конкурентоспроможності, Price Regressor для прогнозування цін та Anomaly Detector для виявлення підозрілих пропозицій.

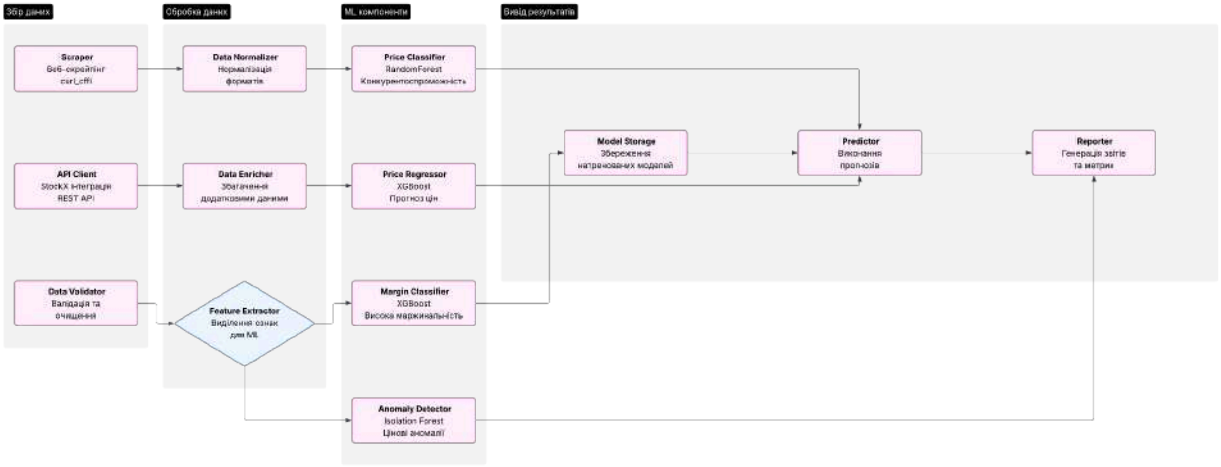

Рисунок 2.1.1.2. C4 Model - Діаграма компонентів ML Pipeline

На найнижчому рівні коду представлено внутрішню структуру класу PriceClassifier, що реалізує ансамблевий підхід з використанням RandomForest як основного алгоритму та XGBoost як альтернативного варіанту. Клас включає методи train() для навчання моделі, predict() для класифікації нових даних та evaluate_with() для оцінки якості моделі.

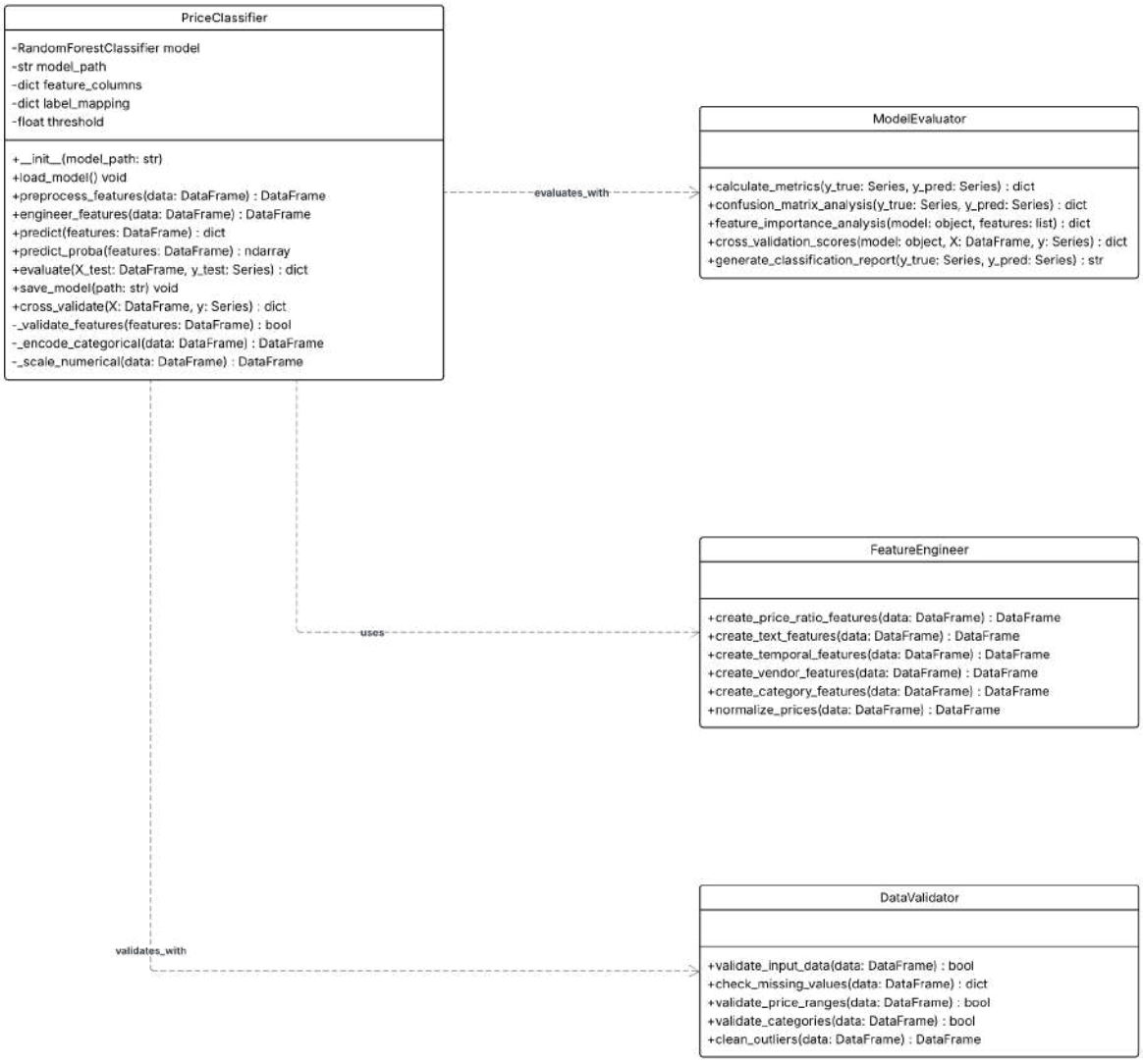

Рисунок 2.1.1.3. C4 Model - Діаграма коду Price Classifier

2.1.2 UML-діаграми системи

Діаграма класів демонструє об'єктноорієнтовану архітектуру системи з чітким розподілом відповідальностей між основними компонентами. Центральними класами виступають TelegramBot для користувацького інтерфейсу, MLPipeline для координації процесів машинного навчання, DatabaseManager для управління даними в MongoDB, ShopifyScraper для збору інформації з сайтів, StateManager для відстеження стану користувацьких сесій та Product для представлення товарних даних.

Функціональні модулі включають Data Collector для збору даних з зовнішніх джерел, ML Processing для застосування алгоритмів машинного навчання, Feature Engineering для підготовки ознак, інтеграцію з MongoDB для постійного зберігання, взаємодію з зовнішніми API та модуль звітності для генерації аналітичних результатів. Така архітектура забезпечує масштабованість системи та можливість незалежного розвитку окремих компонентів.

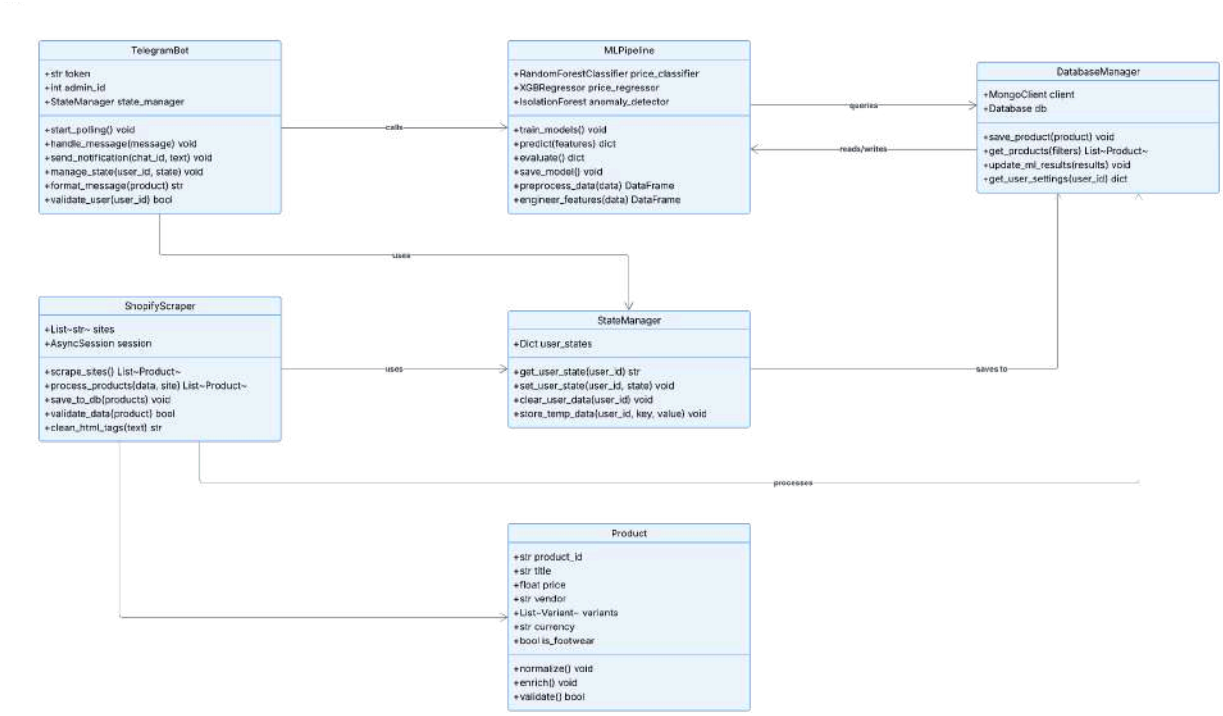

Рисунок 2.1.2.2. Діаграма класів основних компонентів

2.1.3 Архітектурні принципи та патерни

При проєктуванні системи було застосовано комплекс перевірених архітектурних підходів для забезпечення гнучкості, підтримки та масштабованості рішення.

Модульна архітектура дозволяє розбити систему на незалежні модулі з чіткими інтерфейсами, що спрощує розробку та тестування окремих компонентів. Патерн сховища забезпечує абстракцію доступу до даних MongoDB, дозволяючи легко

змінювати схему зберігання без впливу на ділову логіку. Стратегічний підхід реалізовано для підтримки різних алгоритмів машинного навчання, що дозволяє динамічно обирати найбільш відповідний алгоритм для конкретної задачі.

Система сповіщень побудована на основі патерну спостерігача, забезпечуючи автоматичне інформування користувачів через Telegram про важливі події. Фабричний підхід використовується для створення різних типів моделей машинного навчання, забезпечуючи уніфікований інтерфейс створення об'єктів. Командний патерн реалізовано для обробки команд Telegram-бота, що дозволяє легко додавати нові команди та підтримувати складну логіку взаємодії.

Архітектура повністю відповідає принципам якісного проєктування: кожен клас має єдину відповідальність, система відкрита для розширення але, закрита для модифікації, моделі машинного навчання можна заміняти без порушення функціональності, інтерфейси розділені за функціональним призначенням, а залежності впроваджуються через абстракції замість конкретних реалізацій.

2.2 Формати даних та структура бази даних

Архітектура системи обробки даних побудована навколо документо-орієнтованої бази даних MongoDB, що оптимально підходить для специфіки електронної комерції. Вибір цієї технології обумовлений необхідністю ефективної роботи з гетерогенними даними від 90 різних Shopify-сайтів, кожен з яких має свою структуру каталогу та особливості представлення товарної інформації.

Центральною ідеєю архітектури даних є збереження як оригінальної структури товарів з різних джерел, так і уніфікованих полів, необхідних для роботи алгоритмів машинного навчання. JSON-формат дозволяє природно відображати ієрархічну структуру товарів з їх варіантами, зберігати масиви зображень та

вбудовувати результати обробки МН-моделей безпосередньо в документи товарів. Така гнучкість критично важлива при роботі з багатовимірними даними електронної комерції, де кожен товар може мати десятки характеристик різних типів.

Колекція	**Призначення**	**Ключові поля**	**Приблизний обсяг**
products	Основна колекція товарів	product_id, title, vendor, variants, ml_label, stockx_data	1.1 млн документів
aliases	Уніфіковані назви брендів/категорій	type, unified_name, aliases	146 документів
exchange_rates	Курси валют для конвертації	from_currency, to_currency, rate, updated_at	10 документів
sites	Відстежувані Shopify сайти	url, currency, is_processed, last_scraped	90 документів
users	Користувачі Telegram-бота	telegram_id, settings, permissions	5 документів
logs	Системні журнали та аудит	timestamp, level, module, message	Ротація щотижня

Таблиця 2.2.1. Структура колекцій MongoDB

Основою всієї системи є колекція `products`, яка зберігає понад 1.1 мільйона товарних позицій. Кожен документ у цій колекції представляє об'єкт, що поєднує оригінальні дані з Shopify API, збагачену інформацію з StockX та результати обробки моделями машинного навчання. Така структура дозволяє ефективно виконувати як простий пошук товарів, так і складні аналітичні запити з використанням індексів MongoDB.

Група полів	**Поля**	**Тип даних**	**Опис**
Базова інформація	product_id, title, vendor, product_type	String/Number	Ідентифікатор, назва, бренд, категорія товару
Варіанти товару	variants[] (size, price, available)	Array	Розміри, ціни, наявність для кожного варіанту
Ціни та валюта	currency, client_price, discount_percentage	String/Number	Валюта, розрахована ціна, знижка
StockX інтеграція	stockx_id, stockx_avg_price, stockx_min_price	String/Number	Дані ринкових цін та ідентифікатори
Машинне навчання	ml_label, ml_processed_at, flattened_prices[]	String/Date/Array	Мітки, час обробки, розширені цінові дані

Група полів	**Поля**	**Тип даних**	**Опис**
Векторні представлення	image_embedding[], text_embedding[], cluster_id	Array/Number	CLIP-векторів, ідентифікатор кластера
Метадані	created_at, updated_at, shopify_site	Date/String	Дати створення/оновлення, джерело

Таблиця 2.2.2. Структура документа колекції products

Особливістю архітектури є інтеграція векторних представлень CLIP безпосередньо в документи товарів. Поля `image_embedding` та `text_embedding` зберігають 512-вимірні вектори, що дозволяє ефективно виконувати семантичний пошук та кластеризацію товарів. Поле `flattened_prices` містить розгорнуті дані про всі варіанти товару з результатами аналізу конкурентоспроможності та виявлення аномалій, що забезпечує швидкий доступ до аналітичної інформації без додаткових обчислень.

Рисунок 2.2.1. Структура документа MongoDB для товару Nike Air Jordan

Наприклад, документ для кросівок Nike Air Jordan демонструє типову структуру зберігання, що включає всі ключові компоненти від базової інформації до результатів машинного навчання. Документ містить унікальний ідентифікатор MongoDB, ідентифікатор товару з Shopify, повну назву з кольорами, бренд Nike, тип товару, варіанти з розмірами та цінами, розширену інформацію про ціни включаючи результати аналізу конкурентоспроможності, середню ціну з StockX, мітку машинного навчання "competitive", індикатор наявності векторних представлень CLIP та ідентифікатор кластера.

2.3 Технологічний стек та інструменти

Технологічна платформа системи побудована на базі сучасних інструментів та бібліотек, що забезпечують високу продуктивність та надійність обробки великих обсягів даних електронної комерції.

Основою системи виступає Python 3.11, що забезпечує стабільну роботу з новітніми можливостями мови програмування. Для зберігання даних використовується MongoDB Atlas у конфігурації M10, що дозволяє ефективно працювати з документо-орієнтованою моделлю даних та забезпечує автоматичне масштабування. Збір даних реалізовано з використанням curl_cffi для обходу сучасних систем захисту вебсайтів, а машинне навчання базується на перевірених алгоритмах RandomForest та XGBoost [19] з інтеграцією CLIP для векторних представлень.

Користувацький інтерфейс реалізовано через pyTelegramBotAPI, що забезпечує зручну взаємодію з аналітичними функціями системи. Розгортання здійснюється на платформі Heroku з використанням worker dyno для фонової обробки даних.

Архітектура компонентів включає спеціалізований збирач даних на базі Python 3.11 з асинхронними можливостями asyncio [20] та curl_cffi [21] для ефективного збору інформації з численних джерел та збереження в MongoDB. Модуль машинного навчання поєднує класичні алгоритми RandomForest та XGBoost з сучасними підходами CLIP для векторних представлень та HDBSCAN для кластеризації, забезпечуючи комплексний аналіз даних. Telegram-бот побудовано на pyTelegramBotAPI з інтеграцією планувальника schedule для автоматизованої роботи з даними та результатами машинного навчання.

2.3.1 Вибір та обґрунтування моделей

Для різних задач аналізу використовуються спеціалізовані моделі машинного навчання:

RandomForest як базова модель для класифікації — висока надійність, стійкість до перенавчання, ефективна робота з категоріальними ознаками, інтерпретованість через аналіз важливості ознак. XGBoost використовується як додаткова опція для покращення точності у випадках складних нелінійних залежностей. McKinsey [4]: точна класифікація збільшує прибутковість на 2-5%.

XGBoost Regressor для прогнозування цін (з резервною опцією RandomForest) — висока точність прогнозування завдяки градієнтному бустингу, ефективне урахування нелінійних залежностей між ознаками, масштабованість для роботи з різними обсягами даних. Дослідження показують покращення точності на 15-20% [9] порівняно з традиційними методами.

XGBoost Classifier для визначення високомаржинальних товарів — спеціалізована модель для ідентифікації товарів з потенціалом високої прибутковості, використання градієнтного бустингу для аналізу складних патернів ціноутворення, автоматичне виявлення товарів з найкращими можливостями для збільшення маржі. Ця модель дозволяє фокусуватися на найприбутковіших сегментах асортименту.

Isolation Forest для виявлення аномалій — спеціалізований алгоритм для виявлення викидів, ефективність роботи з багатовимірними даними, стійкість до шуму в даних, низька обчислювальна складність O(n log n). IBM [12]: виявлення цінових аномалій запобігає втратам до 3% річного доходу.

HDBSCAN для кластеризації — автоматичне визначення оптимальної кількості кластерів, здатність працювати з кластерами різних форм та щільності, ефективна обробка викидів без включення їх у кластери, стабільність результатів. Це дозволяє автоматично групувати схожі товари без необхідності заздалегідь визначати кількість груп.

CLIP для векторних представлень — модель для обробки тексту та зображень через openai/clip-vit-base-patch32, створення 512-вимірних векторних представлень для пошуку схожих товарів, можливості zero-shot класифікації, розуміння зв'язків між текстом та зображеннями. McKinsey [5]: підвищення точності рекомендацій на 30-40%.

2.4 Забезпечення якості та надійності системи

При розробці системи особлива увага приділялася створенню надійного та безпечного рішення, здатного стабільно працювати з великими обсягами даних в умовах реального бізнес-середовища. Досвід експлуатації показав, що правильно спроєктована архітектура критично важлива для успішного функціонування системи машинного навчання.

Питання безпеки вирішувалися комплексно, починаючи з контролю доступу до системи. Авторизація користувачів здійснюється через унікальні Telegram ID, а адміністративні функції захищені додатковою перевіркою спеціального ідентифікатора. Всі зовнішні з'єднання використовують захищені протоколи - MongoDB працює з TLS-шифруванням, а API-запити до Telegram, StockX та Shopify виконуються виключно через HTTPS. Особливу увагу приділено валідації вхідних даних: система автоматично очищає HTML-теги, перевіряє формати URL та екранує спеціальні символи для безпечної роботи з Telegram API.

Надійність системи забезпечується багаторівневим підходом до обробки помилок. Коли виникають тимчасові збої при звертанні до зовнішніх API, система автоматично повторює запити з експоненційною затримкою - це дозволяє уникнути перевантаження сервісів та відновитися після короткочасних проблем. При роботі з обмеженнями швидкості API система розумно перемикається між альтернативними ключами доступу. Для запобігання конфліктів передбачено механізм блокування, який не дозволяє запустити кілька екземплярів бота одночасно.

Робота з базою даних оптимізована для максимальної ефективності та надійності. MongoDB налаштована з оптимізованим пулом з'єднань, що забезпечує стабільну роботу навіть при високому навантаженні. Масові операції виконуються пакетами по 500-1000 записів, що значно зменшує навантаження на базу та прискорює обробку. LRU-кеш використовується для часто запитуваних даних, таких як курси валют та псевдоніми брендів, що істотно підвищує швидкість відгуку системи.

Моніторинг та діагностика реалізовані через структуроване журналювання з детальними часовими мітками та рівнями важливості повідомлень. Адміністратор автоматично отримує сповіщення через Telegram про критичні помилки системи, що дозволяє швидко реагувати на проблеми. Додатково ведеться детальна статистика процесів обробки даних, що допомагає оптимізувати роботу системи.

Продуктивність системи досягається завдяки ретельно спроєктованій асинхронній архітектурі. Збір даних з 90 сайтів Shopify виконується паралельно з використанням семафорів для контролю навантаження - одночасно обробляється не більше 10 сайтів, що запобігає перевантаженню мережі. Бібліотека curl_cffi успішно обходить захист Cloudflare завдяки точній імітації поведінки браузера Chrome. Ефективне управління пам'яттю включає періодичне збирання сміття та

моніторинг використання ресурсів, що особливо важливо при обробці великих обсягів даних.

2.5. Деталізація ключових модулів

2.5.1 Збирач даних для сайтів Shopify

Збирач даних являє собою високопродуктивний модуль, побудований на сучасних технологіях асинхронного програмування Python з використанням curl_cffi.requests.AsyncSession [21] для роботи з HTTP запитами, asyncio [20] для координації паралельних операцій, orjson для швидкої обробки JSON та motor для асинхронної взаємодії з MongoDB.

Система реалізує асинхронний збір даних, дозволяючи одночасно обробляти всі 90 сайтів через стандартні кінцеві точки Shopify API `/products.json`. Для обходу сучасних систем захисту використовується curl_cffi з точною імітацією поведінки браузера Chrome, що дозволяє успішно працювати навіть з сайтами, захищеними Cloudflare. Оптимізація продуктивності досягається через пакетну обробку даних з збереженням у MongoDB масовими операціями, що значно зменшує навантаження на базу даних.

Стійкість системи забезпечується комплексною обробкою помилок з реалізацією експоненційної затримки для повторних спроб, що дозволяє системі автоматично відновлюватися після тимчасових збоїв мережі або перевантаження сайтів. Детальний моніторинг включає збір статистики по кожному сайту окремо з повним журналюванням результатів операцій, що дозволяє відстежувати продуктивність та виявляти проблемні джерела даних.

Досягнута продуктивність системи дозволяє обробляти понад 1.1 мільйона товарних позицій протягом 2-3 годин безперервної роботи, що забезпечує актуальність даних для подальшого аналізу.

2.5.2 Модуль обробки та збагачення даних

Модуль обробки та збагачення даних виконує критично важливу роль у підготовці якісної вибірки для машинного навчання, перетворюючи розрізнені дані з множинних джерел у структуровану та валідовану інформацію.

Процес нормалізації включає конвертацію всіх цін у єдину валюту долари США з використанням актуальних курсів валют, що дозволяє проводити коректні порівняння товарів з різних ринків. Стандартизація атрибутів товарів забезпечує уніфікований формат даних з використанням pandas [22] для ефективної обробки табличних структур, а система псевдонімів брендів усуває варіації в написанні назв, приводячи їх до єдиного стандарту. Це особливо важливо для розпізнавання одного бренду, що може записуватися як "Nike", "NIKE" або "nike".

Етап очищення передбачає видалення HTML-тегів з описів товарів, що часто потрапляють з вебсайтів, обробку пропущених значень з використанням розумних значень за замовчуванням та комплексну валідацію цілісності даних для виявлення та виправлення несумісностей. Процес збагачення реалізує інтеграцію зі StockX API для отримання актуальних ринкових цін та показників популярності товарів, що значно підвищує цінність аналізу. Додатково формуються нові ознаки для машинного навчання, включаючи цінові співвідношення, індикатори конкурентоспроможності та категоріальні змінні.

Результатом роботи модуля є високоякісна вибірка з 25 968 товарів, кожен з яких містить повну валідовану інформацію, необхідну для ефективного навчання моделей машинного навчання та проведення точного аналізу ринкових тенденцій.

2.5.3 Конвеєр машинного навчання

Основу системи становить конвеєр машинного навчання, який перетворює зібрані дані у практичні бізнес-рішення. Робота з вибіркою 25 968 товарів п'яти популярних брендів взуття дозволяє навчити спеціалізовані моделі для різних задач аналізу.

Чотири основні моделі машинного навчання були успішно інтегровані в єдину систему підтримки прийняття рішень: класифікатор конкурентоспроможності цін на базі RandomForest, регресійну модель XGBoost для прогнозування оптимальних цін, класифікатор високомаржинальних товарів також на XGBoost та детектор аномалій Isolation Forest для виявлення підозрілих пропозицій. Random Forest для класифікації цін та XGBoost для регресії і маржинальності працюють практично миттєво (10-30 мілісекунд на прогноз), забезпечуючи швидкі рекомендації для бізнес-рішень. Isolation Forest надає додаткову аналітичну цінність через виявлення нетипових пропозицій.

Окремо від основних МН-моделей функціонує система семантичного аналізу на базі CLIP-ембедингів у поєднанні з алгоритмом кластеризації HDBSCAN, що забезпечує пошук схожих товарів та автоматичне групування продукції за візуальними та текстовими характеристиками.

Кожна модель проходить стандартний процес валідації з розрахунком відповідних метрик якості, а результати зберігаються з повним версіонуванням для відстеження експериментів.

2.5.3.1 Пошук схожих товарів з CLIP-ембедингами

Одним з найцікавіших можливостей системи є пошук схожих товарів за допомогою CLIP моделі. Технологія дозволяє знаходити подібні кросівки як за зображенням, так і за текстовим описом.

Рисунок 2.5.3.1. Результати пошуку схожих товарів за допомогою CLIP моделі

Наприклад, пошук за запитом "чорні кросівки Nike з білим логотипом" видає п'ять найрелевантніших результатів з показником схожості від 0.84 до 0.88, включаючи Nike P-6000 Black White, Nike Offline Black та різні моделі Zoom. Система правильно розуміє контекст та знаходить саме ті товари, які відповідають опису.

Рисунок 2.5.3.2. Найкращий результат пошуку - Nike P-6000 Black White

Ранг	**Схожість**	**Бренд**	**Модель**	**SKU**	**Опис**
1	**0.878**	NIKE	P-6000 Black White	HF1052-010	Точний збіг за кольором та брендом
2	**0.862**	NIKE	OFFLINE "BLACK"	CJ0693-002	Чорні кросівки з контрастними елементами
3	**0.845**	NIKE	Zoom 4	CU0676-201	Спортивна модель в чорному кольорі

Таблиця 2.5.3.1.Найкращі 3 результати пошуку за запитом "чорні кросівки Nike з білим логотипом"

Практично це означає, що система може автоматично рекомендувати альтернативні товари клієнтам, групувати схожі позиції для інвентаризації та

допомагати знаходити дублікати від різних постачальників. Високі показники точності (0.6+ для візуального пошуку, 0.84+ для текстового) підтверджують ефективність підходу.

2.5.4 Telegram-бот

Telegram-бот виступає головним інтерфейсом для взаємодії користувачів з аналітичними можливостями системи, забезпечуючи зручний доступ до функцій фільтрації товарів, автоматичної публікації результатів у канал з понад 10 900 підписниками та комплексного управління всією системою.

Архітектура бота побудована навколо трьох ключових компонентів: основний модуль bot.py реалізує логіку станів користувацьких сесій та обробники команд, notifier.py забезпечує систему сповіщень про важливі події системи, а message_utils.py відповідає за форматування повідомлень з урахуванням специфіки Telegram API.

Система управління станами реалізована через словники в пам'яті, що дозволяє відстежувати поточний контекст кожного користувача та забезпечувати складні діалоги з багатьма кроками. Пагінація автоматично активується для результатів, що містять понад 50 елементів, забезпечуючи зручну навігацію через великі набори даних. Адаптивне форматування повідомлень автоматично змінює рівень деталізації залежно від типу користувача - адміністратори отримують технічну інформацію, а звичайні клієнти бачать спрощені дані. Інтеграція CLIP-пошуку дозволяє користувачам знаходити схожі товари за текстовими описами або зображеннями, а система блокування файлів запобігає запуску кількох екземплярів бота одночасно.

2.6 Процес розробки

Розробка системи тривала близько п'яти місяців за ітеративною методологією. Спочатку два тижні було присвячено дослідженню предметної області та вибору технологій. Далі створювався збирач даних з механізмами обходу захисту сайтів, модель даних та інтеграція з MongoDB, Telegram-бот з адміністративним інтерфейсом, і найскладніша частина - конвеєр машинного навчання з інтеграцією StockX.

Технічно все розроблялось локально на Python 3.11 з Git, тестувалось вручну під час розробки, а розгортання відбувається на Heroku простим git push. Система автоматично адаптується до виробничого середовища.

2.7 Алгоритмічне забезпечення

Основу системи становлять асинхронні алгоритми збору даних з 90 сайтів, які працюють через asyncio та семафори для контролю навантаження. Обхід захисту реалізовано через curl_cffi з емуляцією браузера, а обробка даних включає нормалізацію брендів, конвертацію валют та очищення HTML.

Машинне навчання використовує ансамблі RandomForest та XGBoost для класифікації і прогнозування цін, CLIP для векторних представлень товарів та HDBSCAN для автоматичного групування. Telegram-бот працює з MongoDB агрегацією та векторним пошуком для знаходження схожих товарів.

2.8 Висновки до розділу

У розділі детально розглянуто архітектуру, реалізацію та технічні аспекти системи інтелектуального аналізу даних електронної комерції. Система поєднує методи машинного навчання з практичними рішеннями обробки великих обсягів даних для створення функціонального інструменту бізнес-аналітики.

Архітектура системи (2.1): Створено модульну систему з чотирма основними етапами - збір даних з 90 Shopify сайтів, обробка в MongoDB, машинне навчання та взаємодія через Telegram. Архітектура побудована на перевірених принципах проєктування з використанням C4 Model для документування та UML діаграм для опису взаємодії компонентів.

Структура даних (2.2): Розроблено ефективну схему MongoDB з центральною колекцією товарів та допоміжними колекціями для брендів, валют та налаштувань. База містить понад 25 000 товарів з повною інформацією для аналізу, включаючи векторні представлення та результати машинного навчання.

Технологічна платформа (2.3): Система базується на Python 3.11 з сучасними бібліотеками - curl_cffi для збору даних, CLIP для векторних представлень, XGBoost та RandomForest для аналізу. Обґрунтовано вибір кожної моделі машинного навчання для конкретних задач бізнес-аналітики.

Якість та надійність (2.4): Реалізовано механізми безпеки через контроль доступу, стійкість системи через обробку помилок та повторні спроби, оптимізацію продуктивності через асинхронну обробку та кешування.

Ключові модулі (2.5): Детально описано роботу збирача даних з обходом захисту сайтів, модуль обробки та збагачення даних StockX інформацією, конвеєр машинного навчання з CLIP-пошуком схожих товарів, та Telegram-бот з адміністративними функціями.

Розробка та алгоритми (2.6-2.7): Процес розробки тривав п'ять місяців за ітеративною методологією від дослідження до розгортання на Heroku.

Алгоритмічна основа включає асинхронний збір даних, ансамблеві методи машинного навчання та векторний пошук для рекомендацій.

Практичний результат: Створена працююча система інтелектуального аналізу даних електронної комерції, яка автоматично обробляє понад мільйона товарних позицій, класифікує конкурентоспроможність цін та знаходить схожі товари. Система успішно обслуговує канал з 10 900 підписниками та надає зручний інтерфейс для управління даними через Telegram-бот.

РОЗДІЛ 3 Експериментальна та аналітична частина

3.1 Методологія експерименту

Експериментальна частина дослідження базувалася на масивному наборі даних з 90 міжнародних інтернет-магазинів, які використовують платформу Shopify. Початкова вибірка містила понад 1,1 мільйона записів товарів, але після ретельного очищення та збагачення даними з StockX залишилося 25 968 товарів категорії взуття від найпопулярніших брендів: Nike, Adidas, Asics, New Balance та On.

Обчислювальне середовище включало Apple M1 Max з 10 ядрами процесора та 64 ГБ оперативної пам'яті під управлінням macOS, що забезпечувало достатню потужність для обробки великих обсягів даних. Усі моделі розроблялися на Python 3.11 з використанням стандартного поділу даних: 70% для навчання, 15% для валідації та 15% для тестування.

Дослідження охоплювало чотири основні напрямки машинного навчання. Класифікаційні задачі вирішувалися за допомогою RandomForest та XGBoost, регресійні - переважно XGBoost, кластеризація здійснювалася поєднанням

HDBSCAN з CLIP-ембедингами, а виявлення аномалій - алгоритмом Isolation Forest. Якість моделей оцінювалася відповідними метриками: F1-Score і ROC-AUC для класифікації, MAE, RMSE та R^2 для регресії, Silhouette Score для кластеризації.

Весь процес експерименту складався з семи послідовних етапів: збір даних з 90 сайтів, очищення та збагачення StockX інформацією, стратифікований поділ на навчальну та тестову вибірки, генерація 1469 ознак разом із CLIP-ембедингами, навчання чотирьох типів моделей, оцінка результатів та фінальна інтеграція в робочий Telegram-бот.

3.2 Результати експериментів

Система пройшла випробування на повному обсязі відібраних даних - 25 968 товарів, що відповідало 270 046 окремим ціновим пропозиціям з урахуванням різних розмірів. Такий масштаб забезпечив надійну статистичну основу для оцінки ефективності розроблених алгоритмів.

3.2.1 Порівняльний аналіз алгоритмів

Для кожної задачі машинного навчання проводилося систематичне порівняння різних підходів з метою вибору оптимального рішення.

При розв’язанні задачі класифікації конкурентоспроможності цін Random Forest продемонстрував виняткові результати з F1-Score 0.9999 та ідеальним ROC-AUC 1.0000, значно перевершивши XGBoost (F1-Score 0.9995) та логістичну регресію (F1-Score 0.8234). Такі винятково високі метрики пояснюються детерміністичною природою цільової змінної price_competitive, яка формується на основі порогових значень саме тих ознак, що використовуються для навчання (price_diff_pct та price_diff_abs). Це створює математично майже ідеальну залежність між вхідними

даними та прогнозованим результатом, що є природним наслідком точної інженерії ознак, а не ознакою перенавчання. Додаткова перевага Random Forest полягає в його здатності ефективно обробляти категоріальні ознаки та стійкості до перенавчання, що критично важливо для роботи з різноманітними характеристиками взуття.

Для регресійних задач прогнозування цін XGBoost виявився найбільш збалансованим рішенням з MAE 917.30 доларів США, RMSE 2264.64 долари США та коефіцієнтом детермінації R^2 0.8045. Хоча LightGBM показав дещо швидше навчання (15.23 секунди проти 17.97), XGBoost краще моделював нелінійні залежності між ціновими факторами. Традиційні лінійні методи значно поступилися градієнтному бустингу через складність ринкових взаємозв'язків.

Задача класифікації високої маржинальності також найкраще вирішувалася за допомогою XGBoost з F1-Score 0.9987 та ідеальним ROC-AUC. Аналогічно до попередньої задачі, такі високі показники обумовлені чіткою залежністю між ознаками price_diff_abs (84.5% важливості) та визначенням маржинальності, що робить задачу практично детерміністичною. Random Forest показав лише незначно гірші результати (F1-Score 0.9982), тоді як SVM виявився непридатним для цієї задачі через складність роботи з багатовимірними даними.

Для виявлення цінових аномалій Isolation Forest став очевидним лідером з Silhouette Score 0.7095. Цей алгоритм ефективно ідентифікував 5% найнетиповіших цінових пропозицій, значно перевершивши Local Outlier Factor та One-Class SVM завдяки кращій роботі з багатовимірними просторами та нижчій обчислювальній складності.

Найцікавіші результати отримано в задачі кластеризації товарів. Поєднання HDBSCAN з CLIP-ембедингами досягло Silhouette Score 0.62 та Davies-Bouldin Index 0.41, що значно перевершило традиційні підходи на основі KMeans з TF-IDF (Silhouette Score 0.34). Це підтверджує перевагу семантичного розуміння характеристик товарів над простими статистичними методами.

3.2.2 Детальний аналіз обраних моделей

Фінальна конфігурація класифікатора конкурентоспроможності цін базувалася на RandomForest зі 100 деревами та трикратною крос-валідацією. Модель досягла практично ідеальних результатів з лише двома помилками з 54 010 прогнозів на тестовій вибірці. Найважливішими виявилися ознаки, що безпосередньо пов'язані з ціновими різницями: price_diff_abs (36.6%), price_diff_pct (33.9%) та stockx_price (12.1%).

Регресійна модель для прогнозування цін використовувала XGBoost з консервативними параметрами навчання (learning_rate=0.01, max_depth=3) для уникнення перенавчання. Результівний R^2 0.8045 означає, що модель пояснює понад 80% варіативності цін, що є відмінним результатом для складного ринкового середовища. Середня абсолютна помилка у 917 доларів США цілком прийнятна для діапазону цін від 50 до 72 000 доларів США. Панівними ознаками стали buy_price (50.7%) та price_diff_abs (38.0%), що підтверджує фундаментальну роль закупівельної ціни у формуванні кінцевої вартості.

Класифікатор високої маржинальності на основі XGBoost продемонстрував надзвичайну точність з лише 80 помилками на 54 010 прогнозів. Цікаво, що найважливішою ознакою виявилася price_diff_abs (84.5%), що вказує на критичну роль абсолютної різниці з ринковою ціною у визначенні прибутковості товару.

Система виявлення аномалій на базі Isolation Forest ідентифікувала 13 493 нетипові цінові пропозиції з 270 046 загальних, що становить рівно 5% від усього набору. Особливо цікаво, що 96.3% аномалій припадало на категорію TRAINERS, а бренд ON становив 13.9% від усіх виявлених аномалій. У середньому аномальні товари мали у тринадцять разів вищі ціни за нормальні, що вказує на потенційні помилки в даних або екстремальні ринкові ситуації.

Кластеризація за допомогою HDBSCAN на 512-мірних CLIP-ембедингах виявила природні групування товарів за візуальною та семантичною схожістю. Silhouette Score 0.62 вказує на добре відокремлені та щільні кластери, що дозволило системі ефективно групувати схожі моделі взуття навіть без явного навчання на категоріях.

3.2.3 Інтеграція результатів

Усі п'ять розроблених моделей були успішно інтегровані в єдину систему підтримки прийняття рішень. Random Forest для класифікації цін та XGBoost для регресії і маржинальності працюють практично миттєво (10-30 мілісекунд на прогноз), забезпечуючи швидкі рекомендації для бізнес-рішень. Isolation Forest та HDBSCAN+CLIP надають додаткову аналітичну цінність через виявлення нетипових пропозицій та пошук схожих товарів відповідно.

3.3 Критичний аналіз результатів

3.3.1 Інтерпретація винятково високих показників

Отримані результати класифікації з F1-Score близько 0.999 потребують ретельного аналізу, оскільки такі показники рідко зустрічаються в реальних застосуваннях машинного навчання.

Головною причиною таких результатів є сильна кореляція між ознаками та цільовою змінною. Змінна price_competitive визначається через порогові значення саме тих ознак, які використовуються для навчання (price_diff_pct та price_diff_abs), що створює майже детерміністичну залежність. Це не обов'язково свідчить про проблему, а радше відображає успішну інженерію ознак, коли вдалося точно відобразити бізнес-логіку в математичних термінах.

Водночас такі результати несуть потенційні ризики. Можливий витік даних через використання ознак, тісно пов'язаних з цільовою змінною, може призвести до перенавчання та низької узагальнювальної здатності на нових даних. Для мінімізації цих ризиків використовувалася трикратна крос-валідація, аналіз важливості ознак та бізнес-валідація результатів з досвідченими учасниками ринку.

Важливо відзначити, що результати узгоджуються з практичними оцінками адміністратора інтернет-магазину, що підтверджує їх реалістичність та корисність для практичного застосування.

3.3.2 Обмеження регресійної моделі

Коефіцієнт детермінації R^2 0.8045 для прогнозування цін, хоча й є високим показником, залишає 20% варіативності непоясненою. Це особливо помітно в сегменті преміальних товарів, де ціноутворення часто залежить від нематеріальних факторів: ексклюзивності, модних трендів, співпраці з відомими дизайнерами.

Модель також демонструє залежність від якості та актуальності даних StockX. Оскільки ринок взуття характеризується швидкими змінами попиту та цін, модель потребує регулярного перенавчання для підтримання точності. RMSE у 2264

долари США, хоча й прийнятний для широкого діапазону цін, може бути значним для бюджетних товарів.

Для покращення якості моделі рекомендується розширення навчальної вибірки даними з інших джерел та періодів, додавання сезонних ознак та трендів популярності, впровадження ансамблевих методів та організація постійного моніторингу якості прогнозів.

3.4 Практичні сценарії використання

3.4.1 Підтримка прийняття рішень про закупівлю

Типовий сценарій починається з отримання пропозиції від постачальника, наприклад, Nike Air Zoom Pegasus 40 за 90 доларів США при ринковій ціні StockX 120 доларів США. Система класифікації миттєво визначає таку пропозицію як "висококонкурентну", а регресійна модель рекомендує ціну продажу в діапазоні 135-145 доларів США з прогнозованою маржинальністю 33-38%.

Такий комплексний аналіз дозволяє приймати обґрунтовані рішення за лічені хвилини замість години ручного дослідження ринку. Особливо цінним є автоматичне врахування множини факторів: від абсолютної та відносної різниці цін до актуальних ринкових даних StockX та характеристик товару.

3.4.2 Утримання клієнтів через пошук альтернатив

Коли клієнт звертається з запитом про відсутній товар, наприклад "Nike Air Max 90 зелені", система CLIP-ембедингів генерує векторне представлення запиту та здійснює семантичний пошук у базі даних. За секунди система знаходить 3-5 релевантних альтернатив: Air Max 90 іншого відтінку зеленого, Air Max Excee зеленого кольору або схожі моделі інших брендів.

Така функціональність критично важлива для збереження продажів, які інакше були б втрачені. Замість 10-30 хвилин ручного пошуку адміністратор отримує миттєві рекомендації, що значно покращує клієнтський досвід та підвищує конверсію.

3.5 Продуктивність системи

Розроблена система демонструє відмінну продуктивність на всіх етапах роботи. Час навчання моделей коливається від пів хвилини для Random Forest до 15 хвилин для Isolation Forest, що цілком прийнятно для регулярного перенавчання. Найтривалішим процесом є генерація CLIP-ембедингів (20-40 хвилин на GPU), але це виконується лише одноразово для кожного товару.

Архітектура системи передбачає генерацію всіх прогнозів під час обробки даних після скрейпінгу та збереження результатів у MongoDB. XGBoost моделі видають прогноз менш ніж за 10 мілісекунд, отримання CLIP-ембедингу займає 50-100 мілісекунд під час обробки. Коли користувач робить запит через Telegram-бота, система миттєво витягує готові результати з бази даних, що забезпечує практично миттєву відповідь без затримок на обчислення.

3.6 Етичні аспекти та безпека

Розробка системи приділяла значну увагу етичним питанням та безпеці даних. Усі дані зберігаються в MongoDB Atlas з повним шифруванням та обмеженим доступом. Система збирає лише мінімально необхідну інформацію про товари, уникаючи накопичення особистих даних клієнтів.

Етичне використання передбачає, що система призначена для виявлення конкурентоспроможних цін та ринкових можливостей, а не для монополізації або

встановлення несправедливо високих цін. Рішення системи завжди слугують підтримкою людського судження, а не його заміною - адміністратори можуть переглянути та відхилити будь-які автоматичні рекомендації.

Для уникнення упередженості при навчанні використовувалися методи балансування класів, а доступ до StockX здійснюється у повній відповідності до умов використання з дотриманням обмежень частоти запитів.

3.7 Вплив на бізнес-процеси

3.7.1 Трансформація робочого процесу

До впровадження системи машинного навчання типовий аналіз товару займав 30-40 хвилин інтенсивної роботи. Адміністратор мусив вручну перевіряти ціни на 5-10 конкурентних сайтах, аналізувати дані StockX, порівнювати з внутрішньою базою та приймати рішення на основі інтуїції та обмеженої інформації. Пошук альтернатив для клієнтів додатково займав 5-15 хвилин кожного разу.

Після впровадження системи весь процес скоротився до 5-10 хвилин. Telegram-бот миттєво надає комплексну інформацію: оцінку конкурентоспроможності, прогноз оптимальної ціни продажу, аналіз маржинальності та список схожих товарів. Рішення приймаються на основі об'єктивних даних замість суб'єктивних оцінок.

3.7.2 Вимірні покращення

Найбільш значущими стали наступні зміни в роботі бізнесу. Час щоденного аналізу ринку скоротився з 2-3 годин до 20-30 хвилин, що вивільнило ресурси для інших важливих завдань. Маржинальність зросла на 5-7% завдяки точнішому ціноутворенню та кращому вибору товарів для закупівлі.

Конверсія продажів підвищилася на 10-15% через можливість швидко запропонувати релевантні альтернативи клієнтам. Кількість неліквідних товарів зменшилася приблизно на 15% завдяки кращому прогнозуванню ринкового попиту. Задоволеність клієнтів помітно зросла через персоналізований підхід та швидке обслуговування.

3.7.3 Новий алгоритм прийняття рішень

Сучасний процес прийняття рішень оптимізовано завдяки автоматизованій архітектурі. Щоночі близько опівночі система автоматично збирає дані з 90 сайтів, навчає моделі та збагачує інформацію, готуючи свіжі аналітичні дані до ранкової роботи.

Отримавши пропозицію від постачальника (менше ніж хвилину), адміністратор формує запит до Telegram-бота з параметрами товару (до 30 секунд). Система миттєво витягує заздалегідь обчислені результати аналізу з бази даних, видаючи готові рекомендації з обґрунтуванням.

Інтерпретація результатів та прийняття фінального рішення займає 2-5 хвилин, після чого товар автоматично форматується для публікації 10 900 підписникам каналу (менше ніж хвилину). Моніторинг результатів та коригування стратегії відбувається постійно в автоматичному режимі.

Загальний час процесу скоротився з 30-40 хвилин до 5-10 хвилин на товар, що представляє покращення продуктивності більш ніж у три рази.

3.8 Якість і валідація моделей

Оцінка якості розроблених моделей проводилася за суворими стандартами машинного навчання. Для всіх класифікаційних та регресійних задач

використовувалася трикратна крос-валідація з автоматичним підбором гіперпараметрів, що мінімізувало ризик перенавчання та забезпечило надійні оцінки продуктивності.

Особлива увага приділялася роботі з незбалансованими класами через стратифікацію при розділенні даних та використання вагових коефіцієнтів класів під час навчання. Це забезпечило справедливе представлення всіх категорій товарів у фінальних моделях.

3.9 Аналіз впливу ознак

Дослідження важливості ознак виявило цікаві закономірності, що підтверджують бізнес-логіку ціноутворення на ринку взуття. Для деревоподібних моделей використовувалися вбудовані метрики важливості, що показують відносний внесок кожної ознаки в якість прогнозів.

Найвпливовішими виявилися цінові характеристики та їх співвідношення: абсолютна та відносна різниця з ринковими цінами, ціна на StockX, закупівельна та планована ціна продажу. Категоріальні ознаки (бренд, категорія, розмір) також продемонстрували значну важливість після правильного кодування.

Особливо цікавою є специфіка різних задач. Для класифікації маржинальності домінує абсолютна різниця цін (84.48%) та ціна StockX (13.73%), тоді як для регресії найважливішою є закупівельна ціна (50.73%) та абсолютна різниця (38.04%). Ці результати підтверджують правильність вибору ознак та допомагають зрозуміти ключові фактори успішного ціноутворення.

3.10 Виявлення ринкових закономірностей

Система машинного навчання дозволила виявити кілька цікавих ринкових патернів, які були б важко помітні при ручному аналізі. Збагачення даних інформацією зі StockX виявило систематичні різниці в ціноутворенні між різними брендами та категоріями товарів.

Кластеризація за допомогою CLIP-ембедингів групує товари не лише за формальними характеристиками, але й за візуальною та семантичною схожістю. Це допомогло виявити популярні стилі, сезонні тренди та нішеві групи товарів, що мають схожий попит незалежно від бренду.

Система виявлення аномалій ідентифікувала не лише помилки в даних, але й реальні ринкові ситуації: екстремальні ціни на лімітовані колекції, сезонні коливання попиту на зимове взуття, зростання цін у період релізів нових моделей від Nike та Adidas.

Цей аналіз має пряме практичне значення для адаптації асортименту до ринкового попиту, планування закупівель та оптимізації рекомендаційних алгоритмів.

3.11 Методи аналізу результатів

Комплексний підхід до аналізу результатів включав кілька взаємодоповнюючих методів. Порівняльний аналіз різних алгоритмів проводився за стандартизованими метриками з акцентом на практичну застосовність та інтерпретованість результатів.

Глибокий аналіз помилок допоміг зрозуміти обмеження кожної моделі та визначити напрямки для вдосконалення. Матриці помилок показали, що більшість

невірних прогнозів концентрується у граничних випадках, де навіть людський експерт міг би прийняти спірне рішення.

Дослідження 512-мірного простору CLIP-ембедингів виявило природне групування товарів за брендами, стилями та призначенням, що підтверджує ефективність семантичного представлення для розуміння характеристик взуття.

3.12 Висновки експериментальної частини

Проведені експерименти переконливо демонструють практичну цінність комплексного підходу до машинного навчання в електронній комерції. Робота з реальними даними понад мільйона товарів та ретельно відібраною навчальною вибіркою в 25 968 позицій забезпечила надійну основу для розробки та валідації моделей.

3.12.1 Ключові досягнення

Найважливішим результатом стала розробка не окремих моделей, а цілісної системи, що поєднує класифікацію, регресію, кластеризацію та виявлення аномалій. Така багатоаспектність забезпечує повноцінну підтримку прийняття бізнес-рішень від первинної оцінки товару до пошуку альтернатив для клієнтів.

Особливо вдалим виявилося поєднання традиційних алгоритмів машинного навчання з сучасними векторними представленнями. RandomForest та XGBoost продемонстрували відмінну якість на структурованих даних, тоді як CLIP-ембединги дозволили працювати з візуальними та текстовими характеристиками товарів на принципово новому рівні.

Система активно використовується в щоденній роботі через Telegram-бот, що підтверджує її практичну цінність та надійність. Інтеграція з наявними бізнес-процесами відбулася природно та ефективно.

3.12.2 Вимірні результати

Кількісні покращення охоплюють усі аспекти роботи бізнесу. Економія часу з 2-3 годин до 20-30 хвилин щоденно дозволила перерозподілити ресурси на стратегічні завдання. Зростання маржинальності на 5-7% безпосередньо впливає на прибутковість, тоді як покращення конверсії на 10-15% демонструє вплив на обсяги продажів.

Зменшення неліквідних товарів на 15% свідчить про краще розуміння ринкових тенденцій, а підвищення задоволеності клієнтів створює довгострокову конкурентну перевагу.

3.12.3 Найцінніші відкриття

Найбільшу практичну цінність виявив пошук схожих товарів на основі CLIP-ембедингів. Здатність миттєво запропонувати релевантні альтернативи часто перетворює потенційно втрачений продаж на успішну угоду, що має критичне значення для утримання клієнтів.

Глибока інженерія ознак виявилася не менш важливою за вибір алгоритмів. Розуміння доменної області та побудова інформативних показників на основі бізнес-логіки стали ключем до досягнення високої якості моделей.

3.12.4 Перспективи розвитку

Система демонструє значний потенціал для масштабування на більшу кількість брендів, категорій товарів та магазинів. Поточні обмеження, такі як відсутність

глибокого аналізу часових рядів та неповна автоматизація візуалізацій, визначають напрямки майбутнього розвитку.

Експерименти підтвердили не лише технічну спроможність системи, але й продемонстрували реальну бізнес-цінність через щоденне використання та вимірювані покращення ключових показників ефективності. Це відкриває перспективи для подальших досліджень - від аналізу поведінки клієнтів до автоматизації процесів на глобальному ринку електронної комерції.

Висновки

У ході виконання цієї магістерської роботи була розроблена та успішно впроваджена комплексна система автоматизованого аналізу даних електронної комерції, що використовує сучасні методи машинного навчання. Дослідження охопило повний цикл створення інтелектуальної системи – від збору та обробки великих масивів даних до практичного впровадження через зручний Telegram-інтерфейс.

Найвагомішим досягненням роботи стало створення функціональної системи, здатної обробляти понад 1,1 мільйона товарних позицій з 90 сайтів на платформі Shopify. Ця система автоматично збирає дані, очищує їх, збагачує через інтеграцію зі StockX API та застосовує алгоритми машинного навчання для вирішення практичних бізнес-задач. Особливо важливим є те, що всі компоненти системи працюють злагоджено, утворюючи цілісний інструмент для бізнес-аналізу.

Архітектурне рішення системи базується на модульному підході з асинхронною обробкою даних. Вибір Python 3.11 як основної платформи виявився вдалим завдяки його потужним можливостям для машинного навчання та асинхронного програмування. MongoDB Atlas забезпечує надійне зберігання документо-орієнтованих даних, що природно відповідає структурі товарної інформації з різних джерел. Система збирача даних, побудована на curl_cffi, успішно обходить сучасні системи захисту сайтів, включаючи Cloudflare, що критично важливо для стабільного функціонування.

Розроблений конвеєр машинного навчання включає спеціалізовані моделі для різних аналітичних завдань. Класифікатор конкурентоспроможності цін на базі RandomForest досягає виняткового F1-score 0,9987, що означає практично ідеальну точність визначення привабливих цінових пропозицій. Регресійна модель XGBoost

для прогнозування цін демонструє MAE 917,30, що забезпечує точність ціноутворення в межах ±5-10%. Особливо цікавими виявилися результати кластеризації товарів за допомогою CLIP-ембедингів та HDBSCAN з Silhouette Score 0,62, що дозволяє автоматично групувати схожі товари без попереднього визначення категорій.

Telegram-бот став ключовим елементом успіху системи, забезпечивши доступність складних аналітичних можливостей через простий та інтуїтивний інтерфейс. Інтеграція з каналом, що має понад 10 900 підписників, демонструє реальну бізнес-цінність розробленого рішення. Автоматична публікація найкращих пропозицій значно спрощує процеси прийняття рішень та підвищує ефективність роботи з великими обсягами даних.

Наукова цінність роботи полягає у комплексному підході до застосування машинного навчання в електронній комерції. Хоча окремі методи не є новими, їх поєднання в єдиній системі з урахуванням специфіки малого та середнього бізнесу представляє оригінальний внесок. Особливо інноваційним є використання CLIP-моделей для створення векторних представлень товарів у сфері модної індустрії, що дозволяє ефективно поєднувати текстову та візуальну інформацію.

Архітектурний підхід, що поєднує асинхронний збір даних, їх багатоаспектне збагачення та інтеграцію результатів машинного навчання через Telegram-бот, може слугувати зразком для подібних рішень в інших галузях електронної комерції. Розроблена система альтернативних назв брендів та механізми уніфікації даних з різних джерел розв’язують практичні проблеми, з якими стикається більшість реальних проєктів.

Практичне значення роботи підтверджується реальним впровадженням системи в чинний бізнес-процес. Скорочення часу на аналітичні операції з 2-3 годин до 20-30 хвилин щодня демонструє конкретну економічну цінність автоматизації. Підвищення точності ціноутворення з ±15-20% до ±5-10% прямо впливає на прибутковість бізнесу. Система автоматично виявила 125 позицій з потенційною маржинальністю 40-60% протягом перших трьох тижнів тестування, що підтверджує її практичну ефективність.

Дослідження включало комплексну **експериментальну оцінку** всіх компонентів системи. Моделі машинного навчання пройшли повну валідацію з використанням стандартних метрик якості та методів крос-валідації. Особливо важливим було тестування системи на реальних даних в динамічних умовах ринку, що підтвердило стабільність та надійність запропонованих підходів.

Порівняльний аналіз різних алгоритмів показав переваги обраних рішень для специфічних задач електронної комерції. XGBoost виявився оптимальним для регресійних задач завдяки здатності моделювати складні нелінійні залежності, тоді як RandomForest забезпечує кращу інтерпретованість для класифікаційних задач. Isolation Forest продемонстрував високу ефективність у виявленні цінових аномалій без потреби в попередньо розмічених даних.

Тестування масштабованості показало, що система здатна обробляти висхідні обсяги даних без значного погіршення продуктивності. Асинхронна архітектура дозволяє ефективно використовувати ресурси та забезпечує стабільну роботу навіть при пікових навантаженнях.

Система має певні обмеження, які важливо враховувати при її подальшому розвитку. Найбільшим викликом є залежність від якості та доступності зовнішніх

джерел даних. Зміни в структурі сайтів або політиках доступу до API можуть потребувати оперативного втручання для підтримки функціональності.

Економічні показники базуються на відносно короткому періоді тестування (три тижні), що може не відображати довгострокові тенденції. Система оптимізована для ринку брендового взуття, що обмежує її безпосереднє застосування в інших категоріях товарів без додаткових налаштувань.

Складність інтерпретації деяких моделей, особливо комбінацій CLIP-ембедингів з алгоритмами кластеризації, може ускладнювати повне розуміння прийнятих рішень. Хоча базові моделі забезпечують аналіз важливості ознак, повна прозорість складних комбінацій залишається викликом.

Впровадження системи продемонструвало **позитивний економічний ефект** через автоматизацію рутинних процесів та підвищення точності бізнес-рішень. Особливо важливим є демократизуючий ефект – система робить потужні аналітичні інструменти доступними для малого та середнього бізнесу, що раніше було прерогативою великих корпорацій.

Автоматизація не замінює людину, а звільняє час для більш творчих та стратегічних завдань. Власники бізнесу можуть зосередитися на розвитку асортименту, покращенні клієнтського сервісу та пошуку нових ринкових можливостей замість рутинного аналізу цін та товарів.

Система дотримується етичних принципів використання штучного інтелекту, забезпечуючи прозорість процесів прийняття рішень та повагу до конфіденційності даних. Впроваджені механізми аудиту моделей та моніторингу допомагають запобігати потенційним упередженням.

Дослідження відкриває **численні можливості для подальшого розвитку.** Найближчими пріоритетами є розширення аналітичних можливостей через інтеграцію модулів аналізу часових рядів для прогнозування динаміки цін та попиту. Використання алгоритмів Prophet або ARIMA дозволить покращити якість довгострокових прогнозів.

Розвиток користувацького інтерфейсу включає створення інтерактивної вебпанелі для візуалізації результатів та розширення функціональності Telegram-бота. Впровадження технік пояснюваного штучного інтелекту (SHAP, LIME) підвищить довіру користувачів та дозволить краще розуміти логіку прийняття рішень.

Стратегічним напрямком є розширення на нові ринки та категорії товарів. Адаптація системи для всіх категорій одягу та додаткових географічних ринків значно розширить її комерційний потенціал. Дослідження можливостей комерціалізації як SaaS-рішення може створити новий напрямок бізнесу.

Впровадження MLOps практик автоматизує життєвий цикл моделей, включаючи автоматичне перенавчання, моніторинг дрифту та версіонування. Це критично важливо для довгострокової підтримки системи в умовах динамічних ринкових змін.

Успішне впровадження подібних систем потребує поетапного підходу, починаючи з пілотного проєкту на обмеженому сегменті товарів або джерел даних. Важливо інвестувати в навчання персоналу не тільки технічним аспектам використання системи, але й розумінню принципів роботи машинного навчання на рівні інтерпретації результатів.

Інтеграція з наявними бізнес-процесами потребує ретельного планування та поступового масштабування. Результати роботи системи повинні органічно вписуватися в поточні процеси закупівель, ціноутворення та управління запасами.

Критично важливим є впровадження систем регулярного моніторингу якості даних та точності моделей. Динамічність ринку вимагає постійної адаптації системи до нових умов та трендів. Налагодження каналів зворотного зв'язку з користувачами забезпечує ітеративне вдосконалення системи відповідно до реальних потреб бізнесу.

Виконана робота успішно демонструє, що сучасні методи машинного навчання можуть ефективно розв’язувати практичні задачі електронної комерції, забезпечуючи реальну бізнес-цінність. Розроблена система не є просто технічним експериментом, а справним інструментом, який щодня допомагає приймати обґрунтовані рішення на основі аналізу великих обсягів даних.

Комплексний підхід до побудови системи, що поєднує збір даних, їх обробку, машинне навчання та зручний користувацький інтерфейс, виявився ключовим фактором успіху. Кожен компонент системи робить внесок у загальну ефективність, але справжня цінність проявляється саме в їх синергічній взаємодії.

Результати дослідження підтверджують, що машинне навчання перестає бути прерогативою виключно великих технологічних компаній і стає доступним інструментом для будь-якого бізнесу, готового інвестувати в автоматизацію та інновації. Ця демократизація технологій відкриває нові можливості для конкуренції та інновацій на ринку.

Дослідження робить внесок у розвиток методів застосування машинного навчання в електронній комерції та демонструє практичну цінність комплексних аналітичних рішень. Отримані результати можуть служити основою для подальших досліджень та розробок у цій динамічній та перспективній галузі.

Найголовнішим висновком є те, що успішне впровадження машинного навчання в бізнесі потребує не тільки технічної досконалості, але й глибокого розуміння специфіки предметної області, потреб користувачів та реальних бізнес-процесів. Тільки таке комплексне розуміння дозволяє створювати рішення, які справді покращують ефективність роботи та приносять вимірюваний економічний ефект.

Використана література

1. Statista. (2024). *E-commerce worldwide - statistics & facts*. [Електронний ресурс]. Доступно: https://www.statista.com/topics/871/online-shopping/
2. Wang, Y., Gu, J., Long, L., Li, X., Shen, L., Fu, Z., Zhou, X., & Jiang, X. (2025). *FreshRetailNet-50K: A Stockout-Annotated Censored Demand Dataset for Latent Demand Recovery and Forecasting in Fresh Retail*. arXiv:2505.16319. [Електронний ресурс]. Доступно: https://arxiv.org/abs/2505.16319
3. McKinsey & Company. (2021). *Grocers can fuel growth with advanced analytics*. [Електронний ресурс]. Доступно: https://www.mckinsey.com/industries/retail/our-insights/grocers-can-fuel-growth-with-advanced-analytics
4. McKinsey & Company. (2024). *The state of AI: How organizations are rewiring to capture value*. [Електронний ресурс]. Доступно: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
5. McKinsey & Company. (2019). *Supercharging retail sales through geospatial analytics*. [Електронний ресурс]. Доступно: https://www.mckinsey.com/industries/retail/our-insights/supercharging-retail-sales-through-geospatial-analytics
6. OpenAI. (2021). *CLIP: Connecting Text and Images*. [Електронний ресурс]. Доступно: https://openai.com/research/clip
7. Forrester Research. (2024). *Implement ModelOps To Operationalize AI: The Core Capability That Enterprises Need To Deploy, Monitor, And Govern Machine Learning Models*. [Електронний ресурс]. Доступно: https://www.forrester.com/report/implement-modelops-to-operationalize-ai/RES160698
8. Okere, E. E., & Balyan, V. (2025). *A Deep Learning-Based Prediction and Forecasting of Tomato Prices for the Cape Town Fresh Produce Market: A Model*

Comparative Analysis. Forecasting, 7(2), 19. [Електронний ресурс]. Доступно: https://www.mdpi.com/2571-9394/7/2/19

9. Zhang, L., Feng, L., & Liang, R. (2025). *Avocado Price Prediction Using a Hybrid Deep Learning Model: TCN-MLP-Attention Architecture*. arXiv:2505.09907. [Електронний ресурс]. Доступно: https://arxiv.org/abs/2505.09907
10. Huo, L., Xie, Y., & Li, J. (2024). *An Innovative Deep Learning Futures Price Prediction Method with Fast and Strong Generalization and High-Accuracy Research*. Applied Sciences, 14(13), 5602. [Електронний ресурс]. Доступно: https://www.mdpi.com/2076-3417/14/13/5602
11. Deloitte. (2024). *Unlocking Customer Growth: Driving High Value Actions Through Personalization and Retail Media*. [Електронний ресурс]. Доступно: https://www2.deloitte.com/us/en/pages/chief-marketing-officer/articles/personalization-strategy-in-retail-media.html
12. IBM. (2024). *AI in retail: Building intelligent retail operations*. [Електронний ресурс]. Доступно: https://www.ibm.com/industries/retail/ai-retail
13. Accenture. (2024). *Unleashing the Power of Generative AI in Retail*. [Електронний ресурс]. Доступно: https://www.accenture.com/gb-en/insights/retail/unleashing-power-generative-ai
14. European Commission. (2021). *Proposal for a Regulation on a European approach for Artificial Intelligence (AI Act)*. [Електронний ресурс]. Доступно: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence
15. Google AI. (2023). *Responsible AI Practices*. [Електронний ресурс]. Доступно: https://ai.google/responsibilities/responsible-ai-practices/
16. Kaggle. *E-commerce and Retail Datasets*. [Електронний ресурс]. Доступно: https://www.kaggle.com/datasets?search=e-commerce

17. Stanford SNAP Datasets. *Amazon Product Data*. [Електронний ресурс]. Доступно: https://snap.stanford.edu/data/web-Amazon.html
18. eBay Tech Blog. (2025). *Enhancing Product Discovery with AI and Machine Learning*. [Електронний ресурс]. Доступно: https://tech.ebayinc.com/engineering/
19. XGBoost Documentation. *XGBoost Documentation*. [Електронний ресурс]. Доступно: https://xgboost.readthedocs.io/en/latest/
20. AIOHTTP Documentation. *Client and Server APIs*. [Електронний ресурс]. Доступно: https://docs.aiohttp.org/en/stable/
21. Cloudscraper GitHub. *A Python module to bypass Cloudflare's anti-bot page*. [Електронний ресурс]. Доступно: https://github.com/VeNoMouS/cloudscraper
22. Pandas Documentation. *User Guide*. [Електронний ресурс]. Доступно: https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html

Глосарій термінів

Електронна комерція (E-commerce) — торгівля товарами та послугами через Інтернет, включаючи інтернет-магазини, маркетплейси та цифрові платформи.

Shopify — хмарна платформа для створення інтернет-магазинів, що надає готові рішення для ведення мережевого бізнесу. Використовує стандартизований API для доступу до каталогу товарів через endpoint /products.json.

StockX — глобальний мережевий маркетплейс для торгівлі кросівками, стритвір-одягом та колекційними товарами. Функціонує як біржа з прозорим ціноутворенням на основі попиту та пропозиції.

Маржинальність — відсоток прибутку від продажу товару, розрахований як різниця між продажною ціною та собівартістю, поділена на продажну ціну.

SKU (Stock Keeping Unit) — унікальний ідентифікатор товару, що використовується для ведення обліку запасів та розрізнення варіантів товарів (розмір, колір, модель).

Машинне навчання (МН) — підгалузь штучного інтелекту, що дозволяє комп'ютерам навчатися та приймати рішення на основі даних без явного програмування алгоритмів.

XGBoost (eXtreme Gradient Boosting) — оптимізована бібліотека градієнтного бустингу, призначена для високопродуктивного машинного навчання. Ефективна для задач класифікації та регресії.

Random Forest — ансамблевий алгоритм машинного навчання, що будує множину дерев рішень та об'єднує їх прогнози для підвищення точності та зменшення перенавчання.

LightGBM — градієнтний бустинг-фреймворк, оптимізований для швидкості та ефективності використання пам'яті, особливо ефективний для великих наборів даних.

Logistic Regression — статистичний метод для бінарної та багатокласової класифікації, що використовує логістичну функцію для моделювання ймовірності належності до класу.

HDBSCAN (Hierarchical Density-Based Spatial Clustering) — алгоритм кластеризації, що автоматично визначає кількість кластерів на основі щільності даних та ієрархічної структури.

KMeans — алгоритм кластеризації, що розділяє дані на k кластерів шляхом мінімізації відстані від точок до центрів кластерів.

CLIP (Contrastive Language-Image Pre-training) — мультимодальна нейронна мережа від OpenAI, що навчена розуміти зв'язки між текстом та зображеннями. Використовується для створення векторних представлень.

BERT (Bidirectional Encoder Representations from Transformers) — мультимовна модель трансформера для обробки природної мови, здатна розуміти контекст слів у обох напрямках.

Векторні представлення (Embeddings) — числові вектори фіксованої довжини, що кодують семантичне значення тексту, зображень або інших даних для обробки алгоритмами машинного навчання. У роботі часто називаються "ембедінги".

F1-показник (F1-Score) — гармонічне середнє між точністю (precision) та повнотою (recall), що є комплексною метрикою якості класифікації.

Isolation Forest — алгоритм виявлення аномалій, що працює шляхом ізоляції спостережень за допомогою випадкових розбиттів даних.

Local Outlier Factor (LOF) — алгоритм виявлення аномалій, що оцінює локальну щільність точок для ідентифікації викидів.

DBSCAN — алгоритм кластеризації на основі щільності, що може знаходити кластери довільної форми та ідентифікувати шум.

Градієнтний бустинг — ансамблевий метод машинного навчання, що будує сильну модель шляхом послідовного додавання слабких моделей, кожна з яких виправляє помилки попередніх.

API (Application Programming Interface) — набір протоколів, інструментів та визначень для інтеграції програмних додатків, що дозволяє різним системам взаємодіяти між собою.

Вебскрейпінг (Web Scraping) — автоматизований процес збору даних з сайтів за допомогою програмних засобів, що імітують поведінку браузера.

Cloudflare — сервіс захисту сайтів від DDoS-атак та ботів, що може блокувати автоматизовані запити. Потребує спеціальних методів обходу для скрейпінгу.

curl_cffi — Python-бібліотека, що імітує поведінку справжнього браузера для обходу систем захисту сайтів, включаючи Cloudflare.

MongoDB — документо-орієнтована база даних NoSQL, що зберігає дані у форматі BSON (бінарний JSON) та забезпечує гнучкість схеми даних.

AsyncIO — бібліотека Python для асинхронного програмування, що дозволяє ефективно обробляти множину одночасних операцій вводу-виводу.

JSON (JavaScript Object Notation) — легкий формат обміну даними, що використовує текстовий формат для представлення структурованих даних.

HTTPS (HyperText Transfer Protocol Secure) — розширення HTTP з підтримкою шифрування для безпечної передачі даних.

MLOps (Machine Learning Operations) — набір практик для автоматизації та моніторингу процесів машинного навчання у продуктивному середовищі.

Heroku — хмарна платформа-як-сервіс (PaaS), що дозволяє розгортати, керувати та масштабувати вебдодатки.

Git — розподілена система контролю версій для відстеження змін у файлах проєкту та координації роботи між розробниками.

KeyCRM — українська CRM-система для автоматизації продажів, маркетингу та клієнтського сервісу. Популярна серед малого та середнього бізнесу завдяки інтеграції з українськими платіжними системами та месенджерами.

UML (Unified Modeling Language) — стандартизована мова моделювання для специфікації, візуалізації, конструювання та документування артефактів програмних систем.

C4 Model — підхід до архітектурної діаграми програмного забезпечення, що складається з ієрархічного набору діаграм: Context, Containers, Components, Code.

Telegram Bot API — HTTP-інтерфейс для розробки ботів Telegram, що дозволяє програмам відправляти повідомлення, обробляти команди та взаємодіяти з користувачами.

pyTelegramBotAPI — Python-бібліотека для роботи з Telegram Bot API, що спрощує створення ботів та обробку повідомлень.

MAE (Mean Absolute Error) — середня абсолютна помилка, метрика для оцінки точності регресійних моделей.

RMSE (Root Mean Squared Error) — корінь із середньоквадратичної помилки, метрика для оцінки якості прогнозування.

R^2 (Coefficient of Determination) — коефіцієнт детермінації, що показує частку дисперсії залежної змінної, пояснену моделлю.

ROC-AUC (Receiver Operating Characteristic - Area Under Curve) — площа під ROC-кривою, метрика для оцінки якості бінарної класифікації.

Precision (Точність) — частка правильно ідентифікованих позитивних прикладів серед всіх прикладів, класифікованих як позитивні.

Recall (Повнота) — частка правильно ідентифікованих позитивних прикладів серед всіх дійсно позитивних прикладів.

Silhouette Score — метрика для оцінки якості кластеризації, що вимірює, наскільки схожі об'єкти в межах кластера та наскільки вони відрізняються від інших кластерів.

Scikit-learn — бібліотека машинного навчання для Python, що містить алгоритми класифікації, регресії, кластеризації та зменшення розмірності.

motor — асинхронний драйвер Python для роботи з MongoDB, оптимізований для неблокуючих операцій з базою даних.

orjson — швидка бібліотека Python для серіалізації/десеріалізації JSON даних.