ДЕРЖАВНИЙ УНІВЕРСИТЕТ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА ІНФОРМАЦІЙНИХ СИСТЕМ ТА ТЕХНОЛОГІЙ

КВАЛІФІКАЦІЙНА РОБОТА

на тему:

«АНАЛІТИЧНА МОДЕЛЬ СПОЖИВЧОЇ ПОВЕДІНКИ НА РИНКУ E-COMMERCE»

на здобуття освітнього ступеня магістр
за спеціальності 124 Системний аналіз
(код,найменування спеціальності)
освітньо-професійної програми Інтелектуальні системи управління
(назва)

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

_____ Антон МАЙБОРОДА
(підпис) *(ім'я, ПРІЗВИЩЕ здобувача)*

Виконав:
здобувач вищої освіти
група САДМ-61 Антон МАЙБОРОДА
(ім'я, ПРІЗВИЩЕ)
Керівник Ровіл НАФЄЄВ
к.т.н. доцент *(ім'я, ПРІЗВИЩЕ)*

Рецензент: _____
(ім'я, ПРІЗВИЩЕ)

Київ 2026
ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Навчально-науковий інститут Інформаційних технологій

Кафедра Інформаційних систем та технологій
Ступінь вищої освіти магістр
Спеціальність 124 Системний аналіз
Освітньо-професійна програма Інтелектуальні системи управління

ЗАТВЕРДЖУЮ
Завідувач кафедрою ІСТ
Каміла СТОРЧАК

"____" _____2025 року

З А В Д А Н Н Я
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Майбороді Антону Олеговичу
(прізвище, ім'я, по батькові здобувача)

1.Тема кваліфікаційної роботи: Аналітична модель споживчої поведінки на ринку e-commerce
керівник кваліфікаційної роботи: к.ф.-м.н. Нафєєв Ровіл Касимович
(ім'я, ПРІЗВИЩЕ, науковий ступінь, вчене звання)
затверджені наказом Державного університету інформаційно-комунікаційних технологій від "___ " жовтня 2025 р. № ____

2. Строк подання кваліфікаційної роботи «26» Грудня 2025 р.

3. Вихідні дані кваліфікаційної роботи:
1. Дані поведінкові метрики користувачів.
2. База даних сеансів користувачів
4. Науково-технічна література.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити):
1. Дослідження та аналіз поведінки онлайн-користувачів.
2. Огляд методів машинного навчання для прогнозування намірів покупців.
3. Аналіз результатів моделей прогнозування поведінки користувачів.

5. Перелік ілюстраційного матеріалу: *презентація*
6. Дата видачі завдання «30» Жовтня 2025р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1.	Підбір технічної літератури		
2.	Обробка матеріалу		
3.	Аналіз поведінки споживачів в онлайн-середовищі		
4.	Дослідження методів машинного навчання		
5.	Інтеграція моделей прогнозування поведінки користувачів		
6.	Висновки за результатами аналізу		
7.	Розробка демонстраційних матеріалів.		
8.	Оформлення магістерської роботи		

Здобувач вищої освіти _____ Антон Майборода
(підпис) *(ім'я, ПРІЗВИЩЕ)*

Керівник кваліфікаційної роботи _____ Ровіл НАФЄЄВ
(підпис) *(ім'я, ПРІЗВИЩЕ)*

РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття ступня магістр: 78 стор., 28 рис., 5 табл., 38 джерел.

Мета роботи – визначення ключових чинників, що впливають на ухвалення рішень про онлайн-купівлю.

Об'єкт дослідження – процес створення прогностичних моделей аналізу поведінки.

Предмет дослідження – закономірності поведінки онлайн користувачів.

Короткий зміст роботи. У першому розділі магістерської роботи виконано аналіз особливостей поведінки споживачів в онлайн-середовищі. Проаналізовано функціонування онлайн-шопінгу та машинного навчання..

Виконано огляд досліджень у сфері онлайн даних та процесу їх збирання. Проаналізовано різні методи, які використовуються для досягнення основних результатів.

У третьому розділі описується методологія, спосіб, за допомогою якого модель виявляє закономірності поведінки користувачв у e-commerce. Презентували результати дослідження і роботи моделі.

КЛЮЧОВІ СЛОВА: ПОВЕДІНКА КОРИСТУВАЧІВ, АНАЛІЗ ДАНИХ, МАШИННЕ НАВЧАННЯ, ПРИЙНЯТТЯ РІШЕНЬ, КЛАСИФІКАЦІЯ, КЛАСТЕРИЗАЦІЯ, АЛГОРИТМИ НАВЧАННЯ, ДЕРЕВО РІШЕНЬ, АНАЛІТИКА КОРИСТУВАЧІВ, E-COMMERCE, ОНЛАЙН-ПОВЕДІНКА, ЗБІР ДАНИХ.

ABSTRACT

The text part of the qualification thesis for obtaining the Master's degree: 78 pages, 28 figures, 5 tables, 38 references.

Purpose of the research is to identify the key factors influencing decision-making in online purchasing.

Object of the research is the process of developing predictive models for behavior analysis.

Subject of the research is the patterns of online user behavior.

Brief summary of the thesis.

In the first chapter of the master's thesis, an analysis of consumer behavior characteristics in the online environment is conducted. The functioning of online shopping and machine learning is analyzed.

A review of studies in the field of online data and the process of their collection is carried out. Various methods used to achieve the main results are analyzed.

The third chapter describes the methodology and the approach by which the model identifies patterns of user behavior in e-commerce. The results of the study and the performance of the model are presented.

KEYWORDS: USER BEHAVIOR, DATA ANALYSIS, MACHINE LEARNING, DECISION-MAKING, CLASSIFICATION, CLUSTERING, LEARNING ALGORITHMS, DECISION TREE, USER ANALYTICS, E-COMMERCE, ONLINE BEHAVIOR, DATA COLLECTION.

ЗМІСТ

ВСТУП

Актуальність теми - Онлайн-шопінг та електронна комерція стали значною частиною світової економіки, а їхні програми стають основним засобом для людей, щоб знаходити, порівнювати та, зрештою, купувати товари. Розуміння тенденцій в онлайн-шопінгу та того, як кожна частина сайту електронної комерції, від відгуків клієнтів до посилань на соціальні мережі, може впливати на трафік та коефіцієнт конверсії, може допомогти онлайн-бізнесу краще обслуговувати своїх клієнтів та збільшувати свій дохід.

Мета і завдання дослідження - Отримання глибокого розуміння поведінки споживачів у сфері онлайн-покупок, застосовуючи низку методів машинного навчання, робота спрямована на визначення ключових чинників, що впливають на ухвалення рішень про онлайн-купівлю. Загальна мета полягає у перетворенні аналітичних результатів на практичні тактики, які бізнес може застосовувати для зміцнення своєї цифрової присутності та ефективнішої взаємодії з клієнтами.

Об'єкт дослідження - розроблення прогностичних моделей, що надають компаніям уявлення про майбутні вподобання клієнтів за допомогою передових методів машинного навчання. Використовуючи такі прогностичні дані, компанії можуть точно налаштовувати стратегії взаємодії та конверсії, підвищуючи їхню ефективність у надзвичайно конкурентному цифровому середовищі. У сучасному світі, що стрімко розвивається, боротьба за цифрову видимість є гострою, а споживчі вподобання постійно змінюються - пошук ефективних маркетингових стратегій набуває вирішального значення. Для досягнення цієї мети було проведено ґрунтовні дослідження та аналіз, із використанням напрацювань [1-3]

Предмет дослідження - прогностична модель, що не лише з високою точністю прогнозує наміри щодо купівлі, але й залишається гнучкою та масштабованою, здатною адаптуватися до змін ринкової динаміки.

Практичне значення одержаних результатів - Це дослідження має ключове значення для вдосконалення цифрових маркетингових тактик і

поглиблення розуміння поведінки споживачів крізь призму передової аналітики.

Використовуючи потужність науки про дані та масштаби великих даних, воно забезпечує бізнес критично важливою інформацією щодо прогнозування та впливу на рішення клієнтів. Такі дані змінюють правила гри для цифрових маркетингових стратегій, даючи змогу компаніям формувати міцніший зв'язок зі своєю аудиторією та точніше адаптувати свої пропозиції. Крім того, отримані висновки збагачують наявний масив знань і створюють основу для подальших досягнень та дослідницьких ініціатив у сфері цифрової поведінки споживачів.

Наукова новизна розробки: У цьому дослідженні використовуються широкі можливості даних для вивчення поведінки онлайн-споживачів, однак водночас враховуються притаманні цьому інструменту обмеження, що визначають рамки отриманих результатів. Незважаючи на те, що набір даних, що використовується у дослідженні, є достатньо інформативним, він охоплює лише частину ширшого цифрового середовища, які хоч і цінні, проте не повністю всеохопні. Додатково, постійний розвиток технологій і зміна споживчих тенденцій ставлять під сумнів тривалу актуальність отриманих результатів. Попри ці обмеження, дослідження має на меті висвітлити типові моделі онлайн-покупок, закладаючи основу, що не лише допомагає бізнесам удосконалювати свої цифрові маркетингові стратегії, але й слугує відправною точкою для подальших наукових досліджень і пошуків.

1 АНАЛІЗ ОБЛАСТІ ДОСЛІДЖЕННЯ

1.1 Розуміння поведінки споживачів в онлайн-середовищі

Історія онлайн-шопінгу зазнала значної трансформації — від ранніх витоків у вигляді електронного обміну даними (EDI) у 1960-х роках до масового впровадження електронної комерції через Всесвітню мережу у 1990-х. Важливою віхою стало запровадження SSL-шифрування в середині 1990-х, що суттєво підвищило безпеку онлайн-транзакцій і рівень довіри споживачів. Масове поширення смартфонів і швидкісного інтернету ще більше змінило ринок, перетворивши мобільну комерцію на один із ключових чинників розвитку, який спростив і покращив досвід покупок. Така еволюція підкреслює визначальну роль технологій у формуванні ринкової динаміки та споживчих вподобань [4].

Зростання цифрової торгівлі, розглянуте у [5], позначає суттєвий зсув у роздрібній торгівлі та суспільній поведінці. Воно демонструє вплив соціальних і мобільних технологій, які спрямовують споживачів на онлайн-платформи, де оперативна ефективність і податкові переваги дозволяють онлайн-ретейлерам пропонувати конкурентні ціни та широкий вибір товарів. Така еволюція створює тиск на традиційні магазини, змушуючи їх упроваджувати інновації або ризикувати втратити актуальність, особливо в умовах домінування смартфонів у сфері онлайн-покупок. У результаті продавці змушені інтегрувати цифрові інструменти для забезпечення безшовного клієнтського шляху та адаптувати свої бізнес-моделі до змінних споживчих уподобань, аби зберегти темпи розвитку та конкурентоспроможність.

Аналіз поведінки онлайн-споживачів через психологічні моделі, зокрема зосереджуючись на стані потоку (flow), дає глибоке розуміння цифрової взаємодії та процесів ухвалення рішень. Стан потоку характеризується глибоким зануренням, у якому виклики та навички суттєво впливають на взаємодію користувача з онлайн-середовищем, а також на його поведінку під час перегляду

та здійснення покупок. Це підкреслює важливість узгодження цифрового середовища з можливостями користувача та підтримання залученості шляхом збалансованого поєднання викликів і навичок. Такий підхід свідчить, що оптимізація користувацького досвіду у цифрових просторах потребує стратегічного балансу, який стимулює залученість, зрештою підвищуючи глибину взаємодії та ймовірність покупки [6].

У сфері електронної комерції дизайн користувацького досвіду (UX) відіграє критичну роль у формуванні якісної взаємодії споживачів із цифровими платформами. Важливе дослідження [7] аналізує взаємозв'язок між дизайном користувацького інтерфейсу (UI) та UX, підкреслюючи їхній вплив на покращення зручності використання сайту й отримання задоволення від взаємодії, що впливає на рішення споживача від моменту першого візиту до здійснення покупки. У роботі наголошується на необхідності ставити в пріоритет UX для підвищення привабливості платформи, акцентуючи на важливості візуально привабливих і зручних дизайнів. Така стратегія є ключовою для підвищення утримання користувачів, їхньої лояльності та рівня продажів, підкреслюючи стратегічне значення UX-дизайну для вдосконалення e-commerce платформ відповідно до сучасних вимог та очікувань користувачів.

1.2 Машинне навчання в електронній комерції

Машинне навчання (ML) та глибинне навчання — ключові складові штучного інтелекту (AI), які докорінно змінюють електронну комерцію, забезпечуючи розширене розпізнавання закономірностей і ухвалення рішень на основі даних. Ці технології, використовуючи такі алгоритми, як нейронні мережі, Random Forest, дерева рішень, Naive Bayes і SVM, істотно підвищують ефективність виявлення шахрайства, а метод SMOTE допомагає долати дисбаланс даних і покращує продуктивність алгоритмів. Наприклад, SVM оптимізує задачі класифікації та регресії, що є важливими для персоналізації взаємодії з клієнтами [8].

Використовуючи нейронні мережі, такі як CNN, глибинне навчання аналізує складні структури даних, підвищує точність розпізнавання зображень і мовлення, а також сприяє глибшому розумінню споживчих уподобань за допомогою передової аналітики . Комплексне застосування ML в e-commerce — від рекомендацій товарів до аналізу клієнтського шляху — демонструє зрушення в бік більш адаптивних, безпечних та орієнтованих на клієнта платформ [9][10].

Персоналізація є центральним напрямом впливу ML на e-commerce: алгоритми аналізують великі масиви користувацьких даних, щоб формувати індивідуальні рекомендації товарів. За допомогою логістичної регресії та Naïve Bayes проводиться аналіз настроїв споживачів на основі відгуків, що покращує користувацький досвід і лояльність. Такий імовірнісний підхід до аналізу настроїв дає змогу впроваджувати динамічні стратегії персоналізації, які змінюються відповідно до споживчих вподобань [11].

Крім того, інтеграція моделей глибинного навчання, таких як RNN, LSTM і GRU, забезпечує виокремлення складних закономірностей у послідовних даних, що дає глибше розуміння поведінки користувачів, підвищуючи їхню задоволеність і стимулюючи розвиток бізнесу [10][12].

Поєднання ML з технологією блокчейн для забезпечення безпеки транзакцій і застосування прогностичної аналітики для прогнозування ринкових трендів є подальшим розширенням ролі ML у сфері e-commerce, відкриваючи перспективи інноваційного, безпечного та високо персоналізованого онлайн-шопінгу [13]. Дослідження ролі машинного навчання в електронній комерції підкреслює глибокий вплив AI-технологій на трансформацію онлайн-покупок. Використовуючи ML і глибинне навчання, платформи e-commerce посилюють персоналізацію, підвищують рівень безпеки та встановлюють нові стандарти операційної ефективності.

Переходячи від базових застосувань до прогностичного моделювання, аналітики та подальших рішень, стає очевидним, що майбутнє досліджень і практик у сфері e-commerce нерозривно пов’язане з постійним розвитком і інтеграцією інтелектуальних систем. У наступних розділах буде детальніше

розглянуто прогностичні моделі, виклики та майбутні напрями, що ще раз підкреслить ключову роль AI у стимулюванні інновацій в електронній комерції.

У сфері електронної комерції прогностичне моделювання докорінно змінює підходи бізнесу до прогнозування трендів, поведінки клієнтів та потреб в інвентарі. Використовуючи складні алгоритми, такі як лінійна регресія та нейронні мережі, ретейлери ефективно орієнтуються в динамічному цифровому середовищі та зберігають конкурентну перевагу. Дослідження, наведене у [14], демонструє ефективність методів машинного навчання, таких як LSTM та XGBClassifier, у прогнозуванні рівня задоволеності клієнтів. Отримані результати дають змогу формувати персоналізовані пропозиції та підвищувати операційну ефективність, зміцнюючи позиції бізнесу на передовій e-commerce інновацій.

Крім того, дослідження, наведені у [15], підкреслюють ключову роль прогностичного моделювання у вдосконаленні систем виявлення шахрайства та підвищенні безпеки транзакцій в електронній комерції. Інтеграція прогностичного моделювання в бізнес-стратегії e-commerce підвищує адаптивність та оперативність реагування на ринкові зміни. Компанії, що активно використовують прогностичну аналітику, отримують конкурентну перевагу та можуть забезпечити стале зростання й успіх у цифровому середовищі.

1.3 Виклики онлайн-шопінгу та машинного навчання

У сфері електронної комерції зростання застосування технологій збору даних підкреслює необхідність захисту приватності [16]. Важливо досягти балансу між використанням даних клієнтів для персоналізованого досвіду та захистом конфіденційної інформації. Новітні технології, такі як блокчейн і шифрування, відкривають можливості для посилення безпеки даних [17]. Однак збереження рівноваги між корисністю даних і приватністю користувачів залишається складним завданням, що потребує подальших досліджень [18].

Алгоритмічні упередження в системах машинного навчання створюють перешкоди для забезпечення справедливості в онлайн-шопінгу. Упередження в навчальних даних можуть спотворювати результати, впливаючи на клієнтський досвід і рівень довіри. Серед заходів для мінімізації таких упереджень — використання алгоритмів, орієнтованих на справедливість, та формування різноманітних навчальних вибірок. Проте забезпечення алгоритмічної справедливості вимагає постійної уваги та адаптації. Необхідно розробляти додаткові стратегії для виявлення та усунення упереджень у моделях машинного навчання

Інтеграція віртуальної реальності (VR) у сферу електронної комерції має потенціал докорінно змінити досвід онлайн-покупок, тоді як доповнена реальність (AR) відкриває унікальні можливості для покращення візуалізації товарів та залучення клієнтів. Ці імерсійні технології створюють середовище, подібне до звичайних торговельних просторів, як підкреслено в [19]. Тривають наукові дослідження, спрямовані на подолання технологічних бар’єрів та оптимізацію користувацького досвіду, що забезпечить безшовні й інтуїтивно зрозумілі взаємодії під час онлайн-шопінгу. Використання AR і VR в електронній комерції є важливим напрямом майбутніх інновацій та розвитку індустрії, як зазначено в [20].

У цьому розділі буде проводиться всебічний огляд і аналіз сучасних досліджень та наукових робіт. Це охоплює детальну оцінку застосованих методологій, отриманих результатів і висновків, зроблених дослідниками у відповідній галузі.

У дослідженні [21] вивчають веб-майнінг поведінки користувачів з метою персоналізації вебсередовища. Вони пропонують метод для формування агрегованих профілів використання шляхом кластеризації користувацьких транзакцій та переглядів сторінок. Ключовим аспектом є розрахунок ваги значущості переглядів сторінок у межах профілю використання, що допомагає вирішувати проблеми розрідженості даних і масштабованості. Формула для цього розрахунку виглядає так:

$$weight(p_{prc}) = \frac{1}{|C|} \sum_{t \in c} w(p_t) \quad (1)$$

Ця формула нормалізує важливість переглядів сторінок, враховуючи частоту їх появи у транзакціях всередині кластера. Завдяки цьому результат стає точнішим і дозволяє створювати більш персоналізований веб-досвід. У дослідженні продемонстровано, що такий підхід покращує якість рекомендацій, що, у свою чергу, позитивно впливає на залученість та задоволення користувачів.

У роботі [22] подано докладний аналіз поведінки онлайн-споживачів, де розрізняються сценарії купівлі, пошуку та звичайного перегляду сторінок за допомогою clickstream-даних. Авторка вводить показники на рівні сесії, такі як “PAGES” та “PGTIME”, які описують відвідування інтернет-магазину. Також використовуються метрики “DIFFCAT” і “DIFFPROD” для фіксації різноманітності дій користувача під час сесії. Кластерний аналіз дозволяє згрупувати різні типи сесій і виявити поведінкові шаблони, що відповідають певним стратегіям користувачів. Такий підхід підкреслює важливість розуміння деталей поведінки онлайн-покупців для створення більш точних маркетингових стратегій.

У дослідженні [23] аналізують класифікацію сесій електронних покупців із використанням методу опорних векторів (SVM), зосереджуючись на відмінностях між сценаріями перегляду та фактичної купівлі. Їхній підхід спрямований на максимізацію відстані між класами при одночасному штрафуванні неправильних класифікацій, що відповідає базовій оптимізаційній задачі SVM. Нижче наведено формулу, яка описує це оптимізаційне завдання.

$$L(\alpha) = \frac{1}{2}\alpha' Q \alpha - e'\alpha \quad (2)$$

Ця формула підкреслює здатність SVM ефективно класифікувати дані великої вимірності, що демонструє його практичну користь у сфері аналітики електронної комерції, зокрема для точного передбачення сесій, які закінчиться

покупкою.

У дослідженні [24] розглядаються методи машинного навчання для підвищення показника завершення електронних транзакцій. Робота зосереджена на моделюванні бізнес-процесів та аналізі даних за допомогою Google Analytics. У дослідженні застосовано класифікатор Байєса та багатошарові нейронні мережі прямого поширення для прогнозування поведінки користувачів і вдосконалення бізнес-стратегій в інтернет-середовищі. Такий підхід демонструє потенціал машинного навчання для оптимізації веб-компонентів відповідно до бізнес-цілей та пропонує практичну модель підвищення частки успішно завершених онлайн-транзакцій.

У дослідженні [25] запропонували структуру прогнозування поведінки онлайн-покупців, зосереджену на визначенні намірів придбання в реальному часі та виявленні ризику покидання сайту. Дослідники сформували набір даних із 12 330 сесій, кожна з яких була позначена за фактом здійснення покупки. Цей репрезентативний датасет став основою для двох моделей прогнозування.

Перша модель поєднувала дані про перегляди сторінок із інформацією про поведінку користувача в сесії та використовувала три алгоритми машинного навчання: Random Forest, SVM і Multilayer Perceptron (MLP). Найвищу точність — 87,24% — продемонстрував саме MLP.

Друга модель була побудована на рекурентній нейронній мережі з архітектурою LSTM, яка працювала з послідовними даними, і забезпечила ще вищу точність — 87,94%.

У подальшому дослідники розширили модель, додавши ансамбль штучних нейронних мереж (ANN), що дало змогу точніше відобразити складні взаємозв’язки між ознаками датасета та результатами прогнозування.

Окремо варто відзначити дослідження Сакар [25] і Кабіра [26], які зробили значний внесок у розвиток методів аналізу поведінки онлайн-покупців та підвищення точності прогнозних моделей. У цих роботах оцінюється ефективність різних алгоритмів машинного навчання, зокрема ансамблевих методів, таких як Random Forest та Gradient Boosting, які демонструють

покращення результатів під час роботи з базовими наборами даних. У їхній роботі досліджуються класифікатори Байєса та алгоритм рішучих дерев C4.5, що також сприяє кращому розумінню побудови ефективних прогнозних моделей у складних умовах. Random Forest підтверджує свою високу результативність, демонструючи переваги ансамблевого навчання у задачах прогнозування [27].

У цьому розділі наведено також кілька базових алгоритмів машинного навчання, зокрема Random Forest (RF), Multilayer Perceptron (MLP) та Long Short-Term Memory (LSTM), з посиланням на їх ключові наукові праці [27, 28, 29].

1.4 Висновки першого розділу

Дослідження методів динамічного прогнозування онлайн-покупок із використанням моделей машинного навчання демонструє вагомий вплив штучного інтелекту на розвиток електронної комерції. Простеження історичної еволюції онлайн-торгівлі, поєднане з аналізом сучасних моделей прогнозування та аналітики, підкреслює трансформаційну роль інтелектуальних систем у формуванні цифрового ринку.

Ключові роботи таких дослідників, як Сакар та ін. [25] і Кабір та ін. [26], створили міцне підґрунтя для розуміння поведінки онлайн-покупців та підвищення точності прогнозування. Їхні інноваційні підходи, що ґрунтуються на ансамблевих методах і сучасних алгоритмах машинного навчання, наочно демонструють потенціал аналітики даних для стимулювання розвитку бізнесу та підвищення задоволеності клієнтів.

У цьому дослідженні закладено намір розвинути окреслені наукові основи, поглиблюючи аналіз поведінкових звичок та вдосконалюючи методи прогнозування в онлайн-торгівлі. Використовуючи датасет, представлений у роботі [25], планується більш детально проаналізувати особливості поведінки споживачів і створити розширені моделі прогнозування, що забезпечуватимуть практичні аналітичні висновки для компаній електронної комерції.

Загалом, метою є внесення внеску у розвиток досліджень та практик у

сфері електронної комерції шляхом інтеграції передових методів машинного навчання та комплексного аналізу даних. Використання технологій штучного інтелекту відкриває нові можливості для інновацій та сприяє покращенню якості взаємодії користувачів з онлайн-платформами.

РОЗДІЛ 2 ОПИС НАБОРУ ДАНИХ

2.1 Джерело даних та процес їх збирання

Набір даних являє собою структуровану таблицю, що містить широкий спектр ознак, які дають можливість проаналізувати поведінку онлайн-покупців та їхню готовність здійснювати покупки. До складу цих ознак входять як числові, так і категоріальні змінні, що забезпечує всебічне уявлення про взаємодію користувачів з платформою електронної комерції.

Числові параметри охоплюють такі метрики, як кількість відвіданих сторінок, тривалість перебування на різних категоріях сторінок, а також показники Google Analytics, зокрема bounce rate, exit rate та page value.
Категоріальні ознаки включають інформацію про операційну систему користувача, браузер, регіон, тип трафіку, тип відвідувача, індикатор вихідних днів, місяць активності та факт отримання доходу.

У цьому розділі здійснюється детальний аналіз структури та змісту набора даних, розглядається розподіл ознак і їхні типи. Також проводиться попередній розвідувальний аналіз даних, який дає змогу зрозуміти, як різні характеристики впливають на прийняття користувачами рішення про покупку та які взаємозв’язки існують між ключовими змінними.

Набір даних, використаний у цьому дослідженні, походить із Online Shoppers Purchasing Intention Dataset, оприлюдненого Sakar та співавторами у межах їх роботи, присвяченої прогнозуванню намірів онлайн-покупців у реальному часі [25]. Цей набір даних є у відкритому доступі, який широко використовується для розміщення навчальних та дослідницьких датасетів.

Online Shoppers Purchasing Intention Dataset містить анонімізовані дані про сесії користувачів, зібрані на одному сайті електронної комерції протягом визначеного періоду. Процес збору включав відстеження взаємодій користувачів із вебресурсом та фіксацію таких характеристик, як кількість відвіданих сторінок, тривалість перегляду різних категорій сторінок, а також метрики

Google Analytics — bounce rate, exit rate та page value. Для підвищення достовірності даних було вжито заходів для мінімізації впливу потенційних чинників, таких як маркетингові кампанії, сезонні події, специфічні профілі користувачів або окремі часові періоди. Кожен запис у датасеті відповідає окремій користувацькій сесії.

У рамках цього дослідження використовується вибірка з **12 330** сесій, сформована на основі даних Google Analytics, що дозволяє проаналізувати поведінку онлайн-покупців. Хоча за обсягом цей датасет не можна однозначно віднести до категорії “великих даних”, характер аналітики електронної комерції — великий обсяг, швидкість надходження даних та різноманітність інформації — потенційно наближує його до цієї сфери. Запропоновані аналітичні методи є масштабованими та можуть бути адаптовані для значно більших масивів даних, що підсилює актуальність даного підходу в контексті big data.

Числові ознаки включають кількість переглянутих сторінок, час перебування у кожній категорії та показники Google Analytics. Категоріальні ознаки охоплюють інформацію про операційну систему, браузер, регіон, тип трафіку, тип користувача, ознаку вихідного дня, місяць та наявність покупки.

Завдяки ретельно організованому процесу збору та обробки даних Online Shoppers Purchasing Intention Dataset є цінним джерелом для дослідження поведінки споживачів у сфері електронної комерції.

Дослідники, які бажають отримати цей набір даних, можуть завантажити його за посиланням, зазначеним у джерелі [30].

2.2 Структура та вміст DataFrame

Набір даних містить 12 330 записів, розподілених по 18 стовпцях, що окреслюють різні аспекти поведінки користувачів під час сеансів електронної комерції. Ці атрибути включають як числові, так і категоріальні характеристики, що забезпечує повне уявлення про взаємодію користувачів на платформі.

Таблиця 1: Опис числових атрибутів

Числові атрибути	
АТРИБУТ	**ОПИС**
Administrative	Рахує кількість адміністративних сторінок, відвіданих під час сесії
Administrative Duration	Відображає час, проведений на адміністративних сторінках
Informational	Рахує кількість інформаційних сторінок, відвіданих користувачем
Informational Duration	Сукупний час, проведений на інформаційних сторінках
ProductRelated	Підсумовує перегляди сторінок, пов’язаних із продуктами
ProductRelated Duration	Загальний час, проведений на сторінках продуктів
BounceRates	Відсоток сесій, у яких було переглянуто лише одну сторінку
ExitRates	Частота виходів користувачів зі сторінки
PageValues	Середня цінність сторінок, відвіданих перед транзакцією
SpecialDay	Показує наближеність до особливих подій

Таблиця 2: Опис категоріальних атрибутів

Категоріальні атрибути	
АТРИБУТ	**ОПИС**
Month	Місяць сесії, критично важливий для аналізу сезонних тенденцій
OperatingSystems	Ідентифікує операційну систему, яку використовує користувач

Browser	Браузер, що використовувався для сесії, впливає на рендеринг сайту
Region	Розташування користувача для аналізу сегментації ринку
TrafficType	Джерело трафіку, що веде на вебсайт
VisitorType	Розрізняє нових і повернених відвідувачів
Weekend	Булевий сигнал, якщо візит відбувся у вихідні
Revenue	Булевий індикатор того, чи призвела сесія до транзакції

Проведений розвідувальний аналіз даних дав змогу виявити низку важливих спостережень, кожне з яких доповнює загальну картину поведінки онлайн-покупців. Отримані результати охоплюють як тонкі закономірності у зміні показників *BounceRates* та *ExitRates*, що частково відображають психологію користувача під час взаємодії з платформою, так і циклічну динаміку транзакцій, зафіксовану атрибутом *Month*. Кожен із цих інсайтів слугує орієнтиром для формування ефективних стратегій у сфері електронної комерції.

Датасет передає не просто набори чисел — він відображає історії залученості користувачів, сезонних коливань у витратах та своєрідного «цифрового діалогу» між покупцем і платформою.

Таблиця нижче подає узагальнені статистичні показники числових змінних датасету. Вона дозволяє оцінити центральні тенденції, варіативність та особливості розподілу ключових характеристик, що формують поведінку онлайн-покупців.

Таблиця 3: Зведена статистика для числових ознак

Feature	Середній	Стандарт	Мин.	Медіана	Макс.	75%
Administrative	2.315	3.322	0.000	1.000	27.000	4.000
Administrative Duration	80.819	176.779	0.000	7.500	3398.750	93.256

Informational	0.504	1.270	0.000	0.000	24.000	0.000
Informational _ Duration	34.472	140.749	0.000	0.000	2549.375	0.000
ProductRelated	31.731	44.476	0.000	18.000	705.000	38.000
ProductRelated _ Duration	1194.746	1913.669	0.000	598.937	63973.522	1464.157
BounceRates	0.022	0.048	0.000	0.003	0.200	0.017
ExitRates	0.043	0.049	0.000	0.025	0.200	0.050
PageValues	5.889	18.568	0.000	0.000	361.764	0.000
SpecialDay	0.061	0.199	0.000	0.000	1.000	0.000
OperatingSystems	2.124	0.911	1.000	2.000	8.000	3.000
Browser	2.357	1.717	1.000	2.000	13.000	2.000
Region	3.147	2.402	1.000	3.000	9.000	4.000
TrafficType	4.070	4.025	1.000	2.000	20.000	4.000

2.3 Візуалізація

Візуалізація даних за допомогою графічних зображень є ключем до виявлення основних закономірностей та ідей, які може знадобитися повною мірою передати зведена статистика. Для ілюстрації розподілу числових атрибутів та дослідження міжзмінних зв’язків використовуються різні графічні методи, включаючи гістограми, діаграми розсіювання та коробкові діаграми.

2.3.1 Показники відмов та виходів:

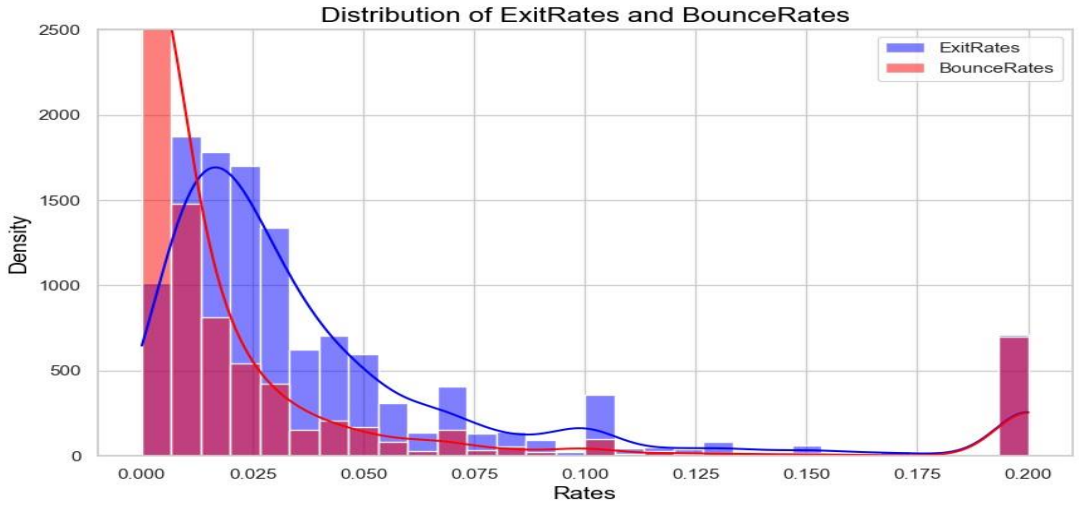


Рисунок 1- Гістограма залежності показників відмов від показників виходів

На рисунку 1 візуально показано розподіл показників «BounceRates» та «ExitRates» за допомогою гістограм та оцінок щільності ядра. Ця візуалізація

надає цінну інформацію про частоту та закономірності цих показників протягом сеансів веб-сайту, пропонуючи глибше розуміння поведінки користувачів. На рисунку підкреслюється різниця між показником відмов та показником виходу – двома критичними показниками залученості користувачів. Показник відмов відображає частку відвідувань однієї сторінки, що вказує на випадки, коли користувачі залишають веб-сайт після перегляду лише однієї сторінки. На противагу цьому, показник виходу означає швидкість, з якою користувачі залишають сайт, незалежно від кількості сторінок, відвіданих протягом сеансу. Рисунок показує ширший розподіл показників виходу порівняно з показниками відмов, що свідчить про різні рівні залученості та взаємодії користувачів у різних розділах веб-сайту. Це спостереження підкреслює важливість комплексного аналізу обох показників для виявлення областей для покращення та оптимізації загального користувацького досвіду.

2.3.2 Показники відмов та виходів за доходом:

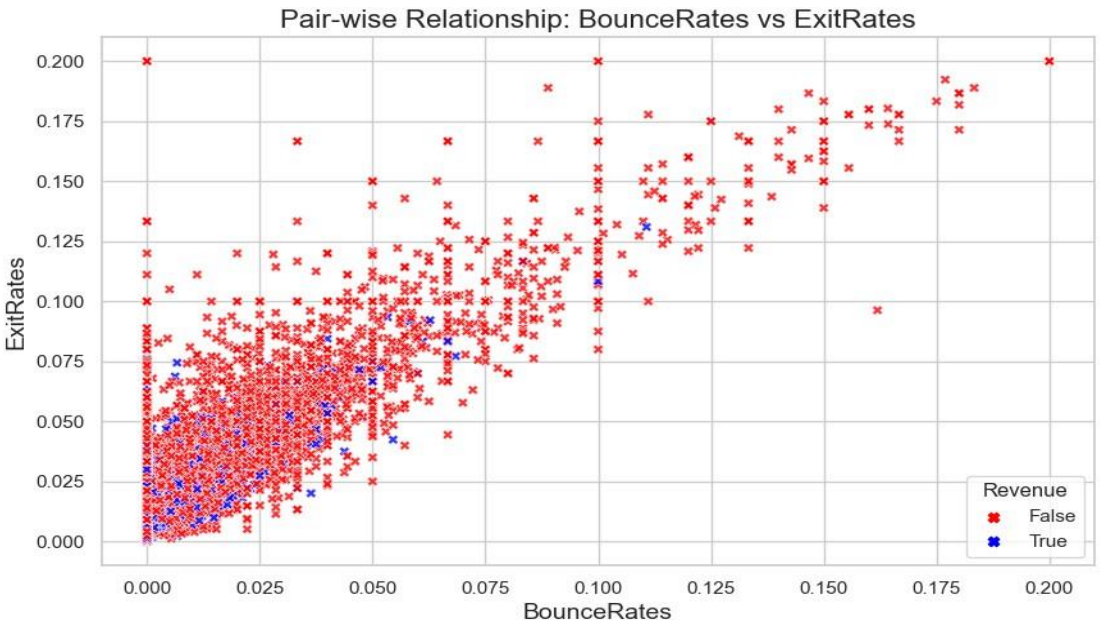


Рисунок 2- Показники відмов та виходів за доходом

На рисунку 2 зображено кореляцію між показниками відмов та виходів залежно від статусу доходу, що проливає світло на суттєвий дисбаланс класів (не

купують проти купують). Цей дисбаланс створює серйозну проблему для традиційних методології машинного навчання. Однак завдяки застосуванню передових методів, таких як надмірна вибірка (SMOTE), ансамблеві методи (пакування, стекування, підвищення) та альтернативні метрики оцінки (криві точності та повного відтворення), стає можливим зменшити ці обмеження та підвищити точність і стабільність моделей.

2.3.3 Розподіл класів доходу

На рисунку 3 наведено візуальне представлення розподілу класів доходу в наборі даних, класифікуючи випадки на основі статусу завершення транзакцій. Серед загальної кількості вибірок 84,5% (10 422 випадки) відповідають випадкам, коли транзакції не були завершені, тоді як решта 15,5% (1908 випадків) вказують на випадки, коли транзакції були успішно завершені. Ця візуалізація пропонує цінну інформацію про поширеність як позитивних, так і негативних вибірок класів, що є ключовим для розробки надійних моделей машинного навчання, спрямованих на точне прогнозування результатів транзакцій.

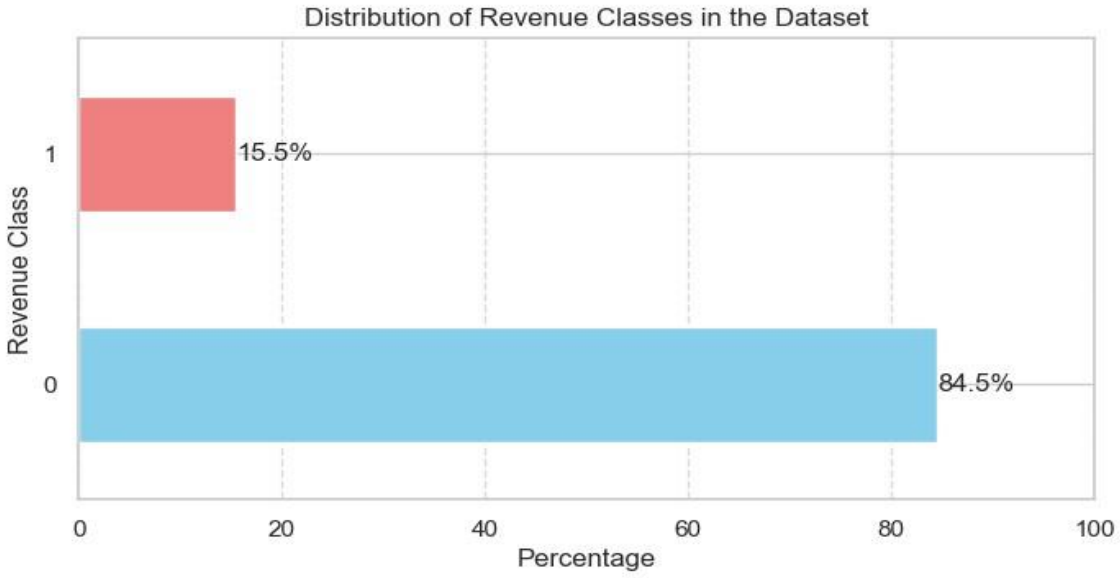


Рисунок 3- Розподіл доходів

2.3.4 Розподіл типів відвідувачів

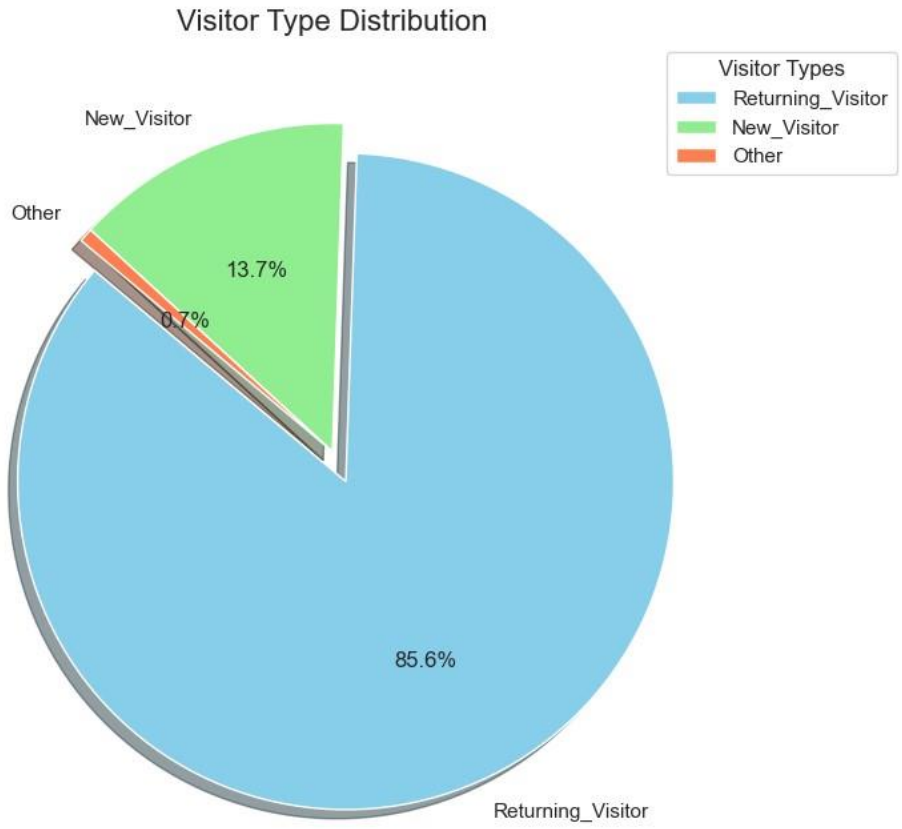


Рисунок 4 - Розподіл відвідувачів

На рисунку 4 користувачів веб-сайту класифіковано на три категорії: нові, постійні та гості. Постійні клієнти – це користувачі, які вже мають створений обліковий запис на веб-сайті чи порталі та регулярно повертаються для отримання послуг або інформації. Нові клієнти – це відвідувачі, які вперше проходять процедуру реєстрації, щоб отримати доступ до функціоналу ресурсу. Гості – це користувачі, які переглядають веб-сайт без створення облікового запису або авторизації; зазвичай вони ознайомлюються з основним контентом, порівнюють інформацію чи оцінюють зручність ресурсу перед потенційною реєстрацією. Такий поділ допомагає точніше аналізувати поведінку аудиторії, оцінювати ефективність залучення нових користувачів і визначати рівень лояльності існуючих.

2.3.5 Аналіз транзакцій у спеціальні дні:

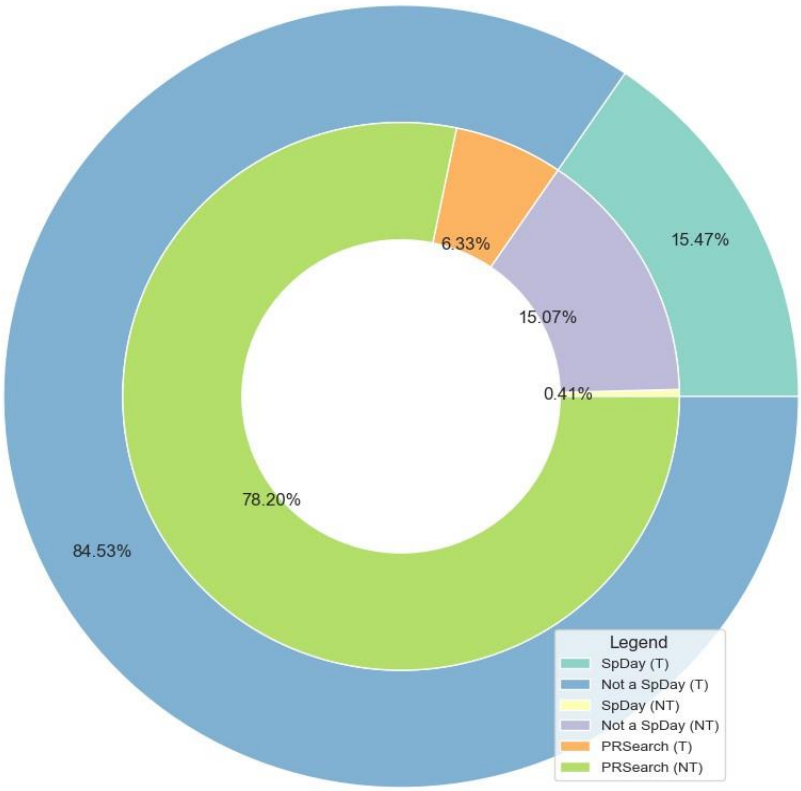
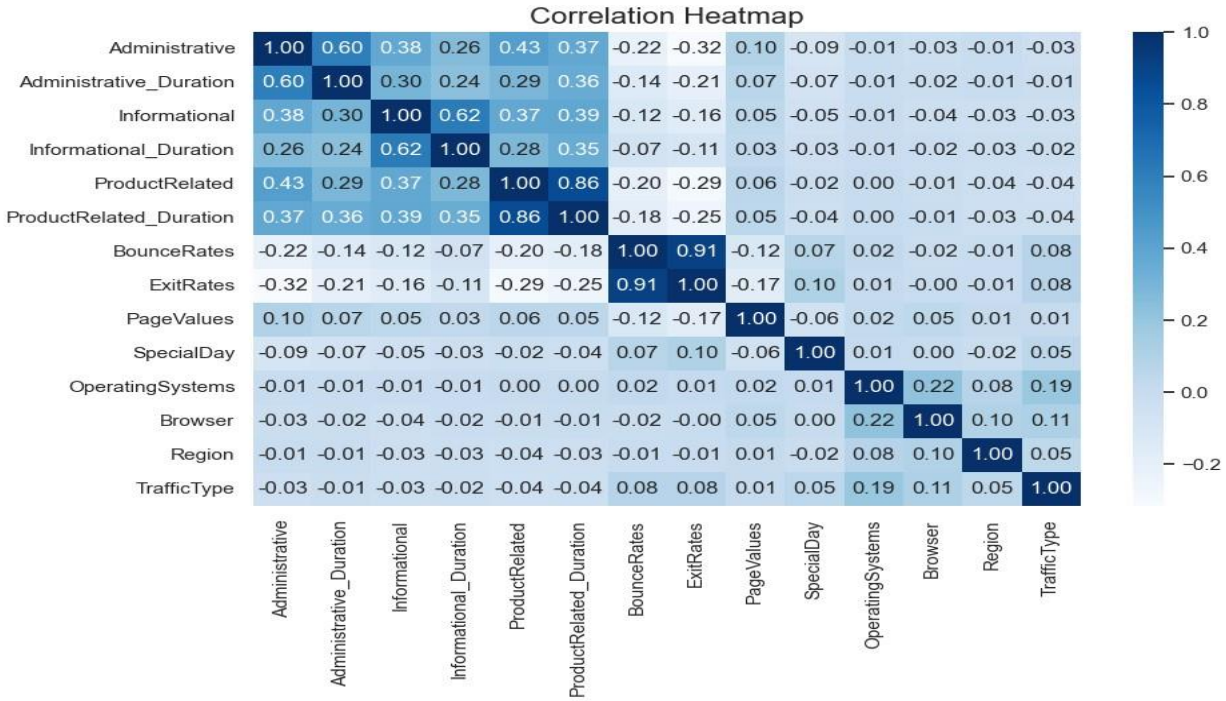


Рисунок 5 - Інфографіка транзакцій у особливі дні

На цьому рисунку зображено розподіл пошукових запитів, пов’язаних з продуктом, що призвели до транзакції, з акцентом на вплив особливих днів. Для ілюстрації рисунка використовується така термінологія:

- PRSearch – Пошук, пов’язаний з продуктом
- SpDay – Особливий день (наприклад, Різдво, Великдень)
- T – Транзакцію завершено

- NT – Транзакцію не завершено Кореляційний аналіз



Коефіцієнт кореляції між двома векторами ознак P та Q задається як

Рисунок 6 - Теплова карта кореляції

$$\rho_{P,Q}=\frac{COV(P,Q)}{\rho_P \cdot \rho_Q} \quad (3)$$

де COV(P,Q) позначає варіацію між P та Q, а ρі позначає стандартне відхилення для і ∈ {P,Q}.

2.3.6 Попарний зв'язок з доходом

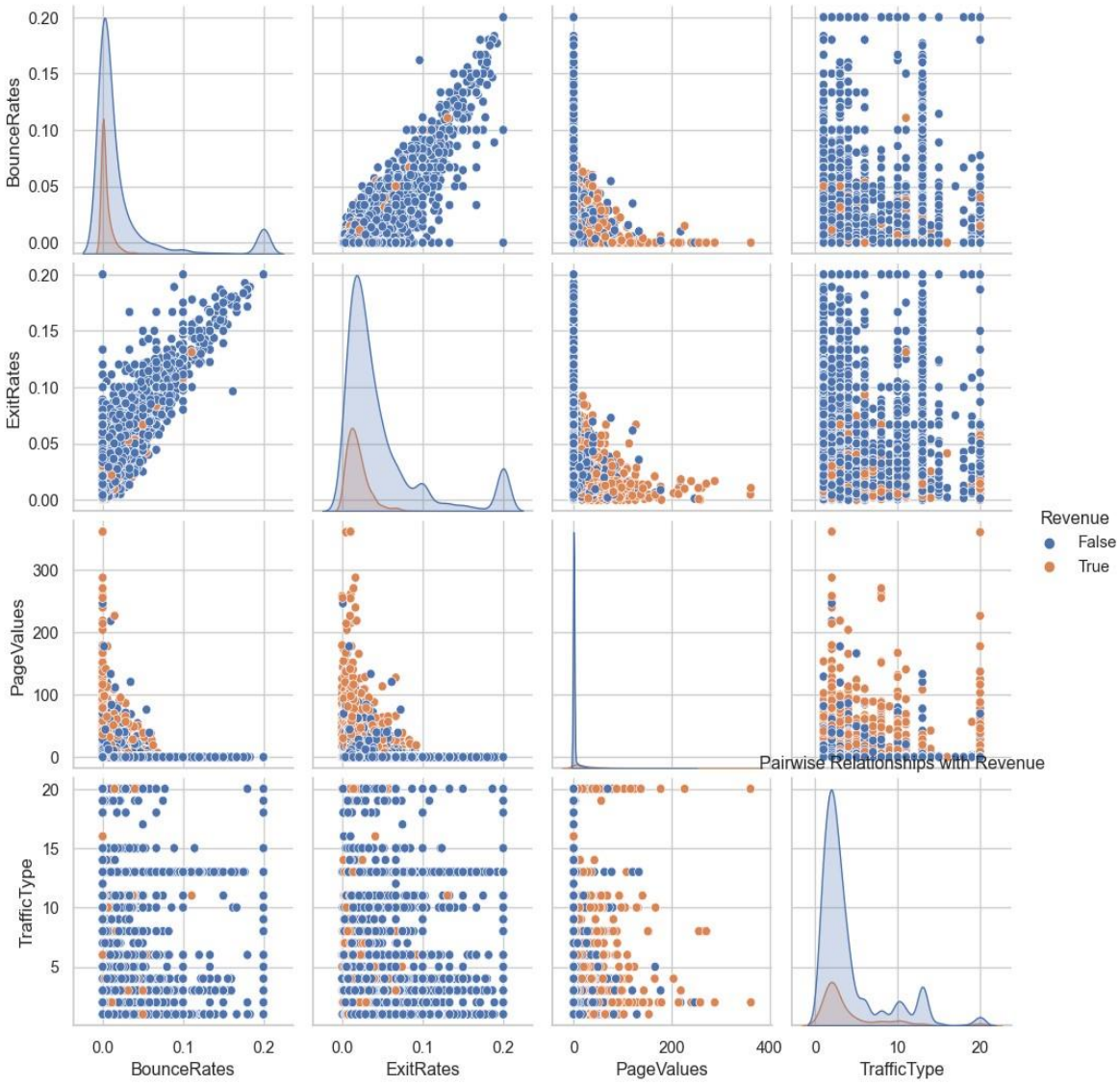


Рисунок 7 - Попарний зв'язок з доходом

Парний аналіз показує взаємозв’язки між BounceRates, ExitRates, PageValues та TrafficType щодо цільової змінної класу - Revenue.

У цьому розділі дослідження датасету було спрямоване на глибоке розуміння поведінки користувачів та її впливу на бізнес-стратегії. Детальний аналіз таких показників, як *BounceRates* та *ExitRates*, дав змогу виявити тонкі закономірності у взаємодії з веб-сайтом. Отримана інформація забезпечила цінні орієнтири для оптимізації користувацької взаємодії та вдосконалення маркетингових стратегій.

Крім того, вивчення шаблонів, пов’язаних із доходом, виявило суттєвий дисбаланс класів, що стало підставою для розгляду передових методів, як-от *oversampling* та ансамблеві моделі. Ефективне вирішення цих проблем сприятиме покращенню продуктивності моделі та отриманню більш точних

бізнес-рішень.

Більш того, дослідження охопило аналіз категорій користувацької взаємодії, впливів особливих днів і взаємозв’язків між ознаками. Ця інформація пропонує практичні підходи до поліпшення клієнтського досвіду та адаптації маркетингових стратегій для максимізації ефективності бізнесу. Глибоке занурення в структуру датасету є фундаментальним етапом для усвідомлення його складності та стратегічного використання з метою досягнення бізнес-цілей.

2.4 Обробка та очищення даних

На етапі попередньої обробки та очищення даних було виконано низку ключових дій для підготовки датасету до навчання та оцінювання моделей машинного навчання.

Перш за все, у датасеті містилися стовпці з булевими значеннями, які були перетворені у цілочисельний тип. Таке перетворення є важливим, оскільки дає змогу виконувати математичні операції з цими стовпцями, що необхідно для їх коректного використання в моделях машинного навчання. Аналогічним чином, стовпці з типом *object*, які зазвичай містять текстові або категоріальні дані, були перетворені у числові представлення. Кожен унікальний елемент таких стовпців був замінений відповідним унікальним числовим значенням. Цей крок є необхідним для того, щоб алгоритми могли ефективно опрацьовувати та навчатися на категоріальних даних.

Датасет було розділено на дві підмножини: тренувальну та тестову. Розподіл здійснювався на основі стовпця *Month*, щоб забезпечити навчання моделі на різноманітних умовах та сценаріях. Тестова частина залишалася окремою для точної оцінки продуктивності моделі та її здатності до узагальнення.

Важливою частиною підготовчого етапу було розв’язання проблеми дисбалансу класів у даних. Тренувальна вибірка була оброблена за допомогою методу SMOTE (Synthetic Minority Over-sampling Technique). SMOTE створює

синтетичні зразки міноритарного класу, використовуючи параметри, зокрема кількість найближчих сусідів (k), щоб збалансувати розподіл класів. Такий підхід значно підвищує ефективність моделі, особливо коли прикладів міноритарного класу недостатньо. Проте застосування SMOTE потребує обережності, оскільки цей метод підходить не для всіх типів задач чи датасетів.

Фінальним кроком було масштабування ознак за допомогою методу StandardScaler. Ця процедура стандартизує всі ознаки таким чином, щоб вони мали середнє значення 0 та стандартне відхилення 1. Стандартизація є критично важливою, оскільки гарантує рівний внесок усіх ознак у процес навчання моделі й запобігає домінуванню ознак із більшими числовими діапазонами. Масштабування було застосоване як до збалансованої тренувальної вибірки, так і до тестової, щоб забезпечити узгодженість представлення даних у всіх етапах побудови моделі.

Завдяки виконанню цих кроків попередньої обробки датасет був якісно підготовлений до наступних етапів розробки моделей машинного навчання, що забезпечило оптимальне представлення даних та підвищило ймовірність отримання точних і надійних прогнозів.

Набір даних вимагає ретельного поводження, надаючи пріоритет конфіденційності та гідності осіб, яких він містить. Він є прикладом відповідального управління даними через такі принципи, як анонімність, згода та прозорість, демонструючи трансформаційну силу аналітики на основі даних в електронній комерції, водночас дотримуючись етичних стандартів.

РОЗДІЛ 3 МЕТОДОЛОГІЯ ДОСЛІДЖЕННЯ

У цьому розділі подано огляд методології дослідження, використаної в роботі. Описано системний підхід — від початкового етапу проектування до збору та попередньої обробки даних, методів аналізу, технік валідації та оцінювання, етичних аспектів, а також програмного забезпечення й інструментів, що були застосовані. Методологія була ретельно розроблена для ґрунтовного дослідження намірів онлайн-покупців. Для отримання глибшої інформації було використано різноманітні класифікатори машинного навчання.

3.1 Збір та моделювання даних

У цьому дослідженні застосовано кількісний підхід для аналізу поведінки онлайн-покупців та прогнозування їхніх намірів здійснити покупку. Унікальною особливістю роботи є її порівняльна методологія, яка дає змогу оцінити ефективність різних класифікаторів машинного навчання. Такий підхід не лише виявляє сильні та слабкі сторони кожного класифікатора, але й допомагає визначити їхню практичну цінність у контексті реальних даних електронної комерції.

У межах цього дослідження значна увага була приділена отриманню даних із джерел відкритих датасетів [30]. Використаний набір даних охоплює широкий спектр взаємодій користувачів, історій покупок та демографічних характеристик. Він містить різноманітні метрики, зокрема тривалість сесій, відвідані веб сторінки, показники виходів, а також бінарну змінну, що відображає ймовірність здійснення покупки. Цей комплексний датасет слугує фундаментом аналітичного дослідження, забезпечуючи цінне розуміння щодо особливостей поведінки споживачів у сфері онлайн-торгівлі.

Для підготовки датасета до аналізу було застосовано ретельний підхід. Спочатку проведено детальне дослідження даних, щоб гарантувати їхню повноту та цілісність шляхом всебічного усунення пропущених значень.

Категоріальні змінні були перетворені у формат, придатний для обробки моделями машинного навчання, що забезпечило їхню безперебійну інтеграцію в аналітичний конвеєр. Крім того, числові ознаки були нормалізовані до стандартизованої шкали, щоб зменшити непропорційний вплив окремих змінних на продуктивність моделей. Потім були застосовані сучасні методи відбору ознак для уточнення датасета, з пріоритетом провісник, що демонструють сильний зв'язок із намірами здійснення покупки. Цей ітеративний процес удосконалення оптимізував аналіз, зосереджуючись на найбільш впливових факторах, що формують поведінку споживачів.

3.2 Моделі та функції аналізу

Логістична регресія є базовою моделлю в машинному навчанні, яка особливо цінується за свою ефективність у задачах бінарної класифікації. Вона обчислює ймовірність того, що певний вхідний приклад належить до конкретного класу, що робить її незамінною в умовах, де результат має два можливі значення.

У центрі логістичної регресії знаходиться логістична (сигмоїдна) функція, яка є ключовою для оцінки ймовірності того, що спостереження належить до певної категорії. У математичному вигляді це виражається наступним чином:

$$P(y = 1|x) = \frac{1}{(1+e\text{^}(-(\beta\_0+\beta\_1\, x)))} \quad (4)$$

Де "P(y = 1 | x)" представляє ймовірність того, що спостереження x буде віднесене до класу 1, а "β0" та "β1" є відповідно коефіцієнтами зсуву та нахилу. Ця модель встановлює зв'язок між лінійними та логарифмічними шансами бінарного результату, що додатково пояснюється через логіт-функцію.

$$\log\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right) = \beta_0 + \beta_1 x \quad (5)$$

Функції вартості та помилки:

Оптимізація логістичної регресії зосереджується на мінімізації функції вартості, також відомої як логарифмічна втрата (log loss), яка оцінює помилку прогнозування моделі.

$$J(\beta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(P(y^{(i)}|x^{(i)})) + (1-y^{(i)})\log(1-P(y^{(i)}|x^{(i)}))] \quad (6)$$

Ця функція вартості штрафує за відхилення між передбаченими ймовірностями та фактичними класовими результатами, спонукаючи модель до точніших прогнозів і підвищуючи її здатність ефективно виявляти приховані закономірності в даних.

KNN-класифікатор:

K-Nearest Neighbors (KNN) — це непараметричний метод, який використовується для задач класифікації та регресії. У задачах класифікації алгоритм KNN визначає клас нового зразка на основі більшості класів серед *k* найближчих сусідів.

KNN використовує відстань у просторі ознак, щоб знайти *k* найближчих сусідів до точки-запиту. Відстань може вимірюватися різними метриками, такими як: евклідова відстань, манґеттенська (міська) відстань та відстань Чебишева.

Кожна метрика по-різному визначає подібність між точками даних.
Евклідова відстань між двома точками x та y, n-вимірному просторі визначається формулою:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{\{ik\}} - x_{\{jk\}})^2} \quad (7)$$

Манхеттенська відстань, або L1-відстань, визначає дистанцію між двома точками таким чином:

$$d(x_i, x_j) = \sum_{k=1}^{n}|x_{\{ik\}} - x_{\{jk\}}| \quad (8)$$

Вона відображає суму абсолютних різниць їхніх координат.

Відстань Чебишева між двома точками задається так:

$$d(x_i, x_j) = \max_k \left| x_{\{ik\}} - x_{\{jk\}} \right| \quad (9)$$

Що представляє собою максимальну абсолютну різницю між координатами точок у всіх вимірах.

Візуалізація показників відстані:

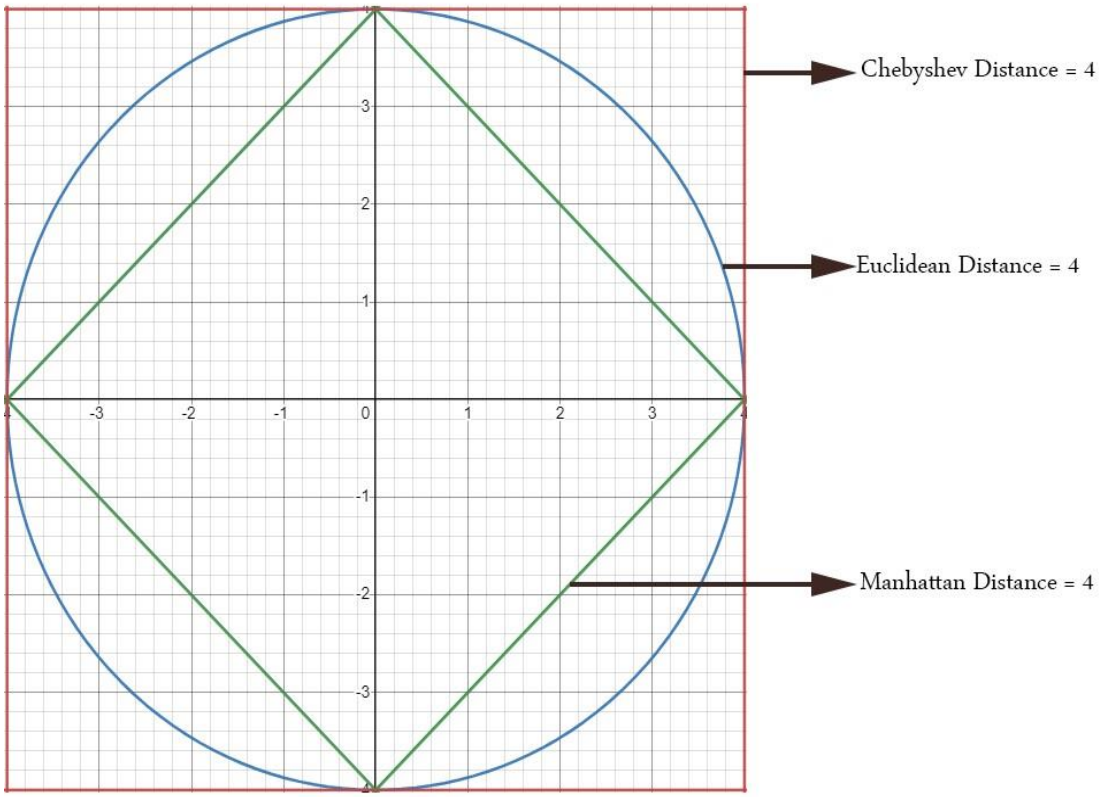


Рисунок 8 - Візуалізація евклідової, манхеттенської та чебишевської відстаней

- Коло позначає евклідову відстань (пряма лінія).
- Ромб — відстань Чебишева (максимальна різниця по осях).
- Квадрат представляє мангеттенську відстань (рух за сіткою).

Метод KNN широко застосовується в різноманітних галузях завдяки своїй простоті та ефективності, особливо у випадках, коли межа класифікації не є лінійно відокремлюваною. У цьому дослідженні KNN було використано для

класифікації намірів онлайн-покупок, продемонструвавши здатність методу адаптуватися до складних закономірностей у даних без необхідності явного навчання моделі. Різноманітність можливостей та потенціал застосування KNN детально описані в [31].

Класифікатор на основі дерева рішень:

Структура дерева рішень зазвичай нагадує блок-схему:
внутрішні вузли відповідають ознакам або атрибутам, гілки відображають правила прийняття рішень, а листові вузли представляють кінцеві результати класифікації.

Кореневий вузол розташований у верхній частині дерева та ініціює процес навчання, розділяючи дані на основі значень атрибутів. Його завдання — розбити дані так, щоб максимально підвищити відмінність між сформованими групами.

Для побудови структури дерева рішень використовуються поняття ентропії та приросту інформації.

Ентропія, що є мірою невпорядкованості або "нечистоти" вибірки, обчислюється за формулою:

$$H(X) = -\sum_{i=1} P(x_i)\log_2 P(x_i) \qquad (10)$$

Де H(X) – ентропія множини X, P(xi) – частка прикладів у класі i в межах X. Інформаційний приріст потім використовується для визначення атрибута для розділення, розраховуючи його шляхом віднімання зважених ентропії кожного розділу від початкової ентропії

Дерева рішень є універсальними алгоритмами, які можна використовувати як для класифікації, так і для регресії. У цьому дослідженні, на основі набору змінних, Класифікатор дерева рішень було застосовано для розмежування клієнтів, які з високою ймовірністю здійснять покупку, та тих, хто цього не зробить. Ця модель особливо цінується за інтерпретованість та здатність

ефективно працювати з категоріальними даними. Основні концепції та сфера застосування дерев рішень детально висвітлені в [31].

Класифікатор Random Forest:

Класифікатор Random Forest працює як ансамблевий метод навчання, створюючи під час тренування множину дерев рішень і використовуючи їх спільно для прогнозування класу нового зразка. Підсумкове передбачення формується шляхом об’єднання індивідуальних рішень кожного дерева, зазвичай через вибір найпоширенішого класу серед них. Такий підхід істотно підвищує якість процесу прийняття рішень і точність, завдяки зменшенню перенавчання. Використовуючи різноманіття кількох дерев та усереднюючи їх результати, класифікатор Random Forest ефективно виявляє закономірності.

Random Forest удосконалює простоту дерев рішень шляхом внесення випадковості у двох аспектах: бутстрепінг (bootstrapping) даних та використання випадкових підмножин ознак для поділу вузлів.

Рівень похибки моделі можна описати на основі кореляції між деревами та «силою» (точністю) кожного окремого дерева в лісі:

$$\text{Error} = \text{function}(\text{tree correlation}, \text{tree strength}) \quad (11)$$

Метою є підвищення різноманітності між деревами, зменшення кореляції між ними та одночасне збереження «сили» кожного окремого дерева.

Random Forest особливо корисний під час роботи з великими наборами даних із високою розмірністю. Алгоритм здатний опрацьовувати пропущені значення, зберігати точність навіть за значної частки відсутніх даних, а також надає внутрішній механізм оцінки важливості ознак. У цьому дослідженні Класифікатор Random Forest було застосовано для прогнозування намірів покупок в онлайн-середовищі, продемонструвавши здатність моделі виявляти складні взаємодії без потреби в детальному налаштуванні параметрів. Надійність підходу та різноманітність застосувань Random Forest детально розглянуто в

[31].

Класифікатор Gradient Boosting:

GradientBoostingClassifier — це ансамблевий метод, який будує моделі послідовно, причому кожна нова модель виправляє помилки попередніх. Цей підхід передбачає об'єднання кількох дерев рішень для створення потужної прогностичної моделі.

Gradient Boosting ґрунтується на трьох ключових елементах:

- функція втрат - яку необхідно мінімізувати;
- слабкий учень (weak learner) - що формує початкові прогнози;
- адитивна модель - яка поетапно додає слабкі моделі для зменшення значення функції втрат.

Алгоритм будує модель поетапно за таким принципом:

$$F_{m+1}(x) = F_m(x) + \alpha_m h_m(x) \quad (12)$$

Де $F_m(x)$— це прогностична модель на ітерації m,

$h_m(x)$— слабкий учень, доданий на цій ітерації,

а α_m— крок (ваговий коефіцієнт), обраний на ітерації m.

Метою є знайти набір α_mта $h_m(x)$, які мінімізують функцію втрат.

Gradient Boosting Classifier високо цінується за ефективність у задачах класифікації, зокрема у випадках із незбалансованими даними. Цей метод відомий високою точністю прогнозування, але потребує ретельного налаштування параметрів, щоб уникнути перенавчання.

У даному дослідженні Gradient Boosting Classifier був застосований для покращення прогнозування поведінки онлайн-покупців, спираючись на його здатність працювати зі складними наборами даних і підвищувати продуктивність моделі шляхом ітеративного вдосконалення. Метод чудово виявляє приховані та слабко виражені закономірності в даних. Методологію та практичні аспекти застосування Gradient Boosting детально описано в [31].

Класифікатор AdaBoost:

Класифікатор AdaBoost (Adaptive Boosting) — це ансамблевий метод, який об’єднує кілька слабких моделей у одну сильну. Він послідовно змінює ваги неправильно класифікованих прикладів, щоб наступні слабкі моделі приділяли більше уваги складним випадкам.

Алгоритм AdaBoost починає роботу з рівними вагами для всіх прикладів та ітеративно оновлює їх залежно від точності прогнозів слабкого учня. Після кожної ітерації ваги неправильно класифікованих прикладів збільшуються, завдяки чому наступний слабкий учень отримує більший вплив саме на помилкові випадки.

Підсумковий прогноз AdaBoost є зваженою сумою прогнозів слабких моделей:

$$\mathrm{F}(x) = \sum_{m=1}^{M} \alpha_m h_m(x) \quad (13)$$

Де M – кількість слабких учнів, hm(x) – прогноз m-го учня, а αm – вага, присвоєна прогнозу цього учня, що визначається його точністю.

AdaBoost відомий покращенням точності алгоритму навчання та є дуже ефективним у задачах класифікації. Він менш схильний до перенавчання, ніж складніші моделі, особливо з шумними даними. У цьому дослідженні AdaBoost було застосовано для підвищення точності прогнозування намірів щодо покупок в Інтернеті, що ілюструє його корисність у підвищенні продуктивності прогностичних моделей у сценаріях зі складними даними. Методологія та застосування AdaBoost детально пояснюються в [31].

Класифікатор додаткових дерев:

Класифікатор додаткових дерев (надзвичайно випадкових дерев) – це метод ансамблювального навчання, подібний до випадкового лісу, але він вносить більше випадковості в модель. Під час побудови дерев він випадковим чином вибирає точки розрізу для розділення ознак, а не шукає найкращі можливі

пороги, як у випадковому лісі.

Метод працює шляхом побудови кількох дерев. Для кожного вузла, замість обчислення оптимального розподілу між усіма ознаками, Extra Trees вибирає випадкову підмножину ознак і робить найкращий розподіл з цієї підмножини. Такий підхід може значно зменшити дисперсію моделі, хоча й з можливим збільшенням зміщення:

$$F(x) = \frac{1}{M} \sum_{m=1}^{M} h_m(x) \quad (14)$$

Де M – кількість дерев, а hm(x) – прогноз m-го дерева. Остаточний прогноз – це середнє значення всіх прогнозів дерев, що підвищує стійкість моделі.

Класифікатор додаткових дерев хвалять за його ефективність та простоту використання, часто потребуючи менше часу на навчання, ніж випадковий ліс, через випадкову природу розбиття. Він особливо корисний у сценаріях, де розмірність даних висока, а зв'язок між ознаками та ціллю складний. У цьому дослідженні класифікатор додаткових дерев був використаний для прогнозування поведінки онлайн-покупців, використовуючи його здатність обробляти високо розмірні дані та надавати глибокі інтерпретації важливості різних ознак. Для поглибленого дослідження теоретичних основ та практичного застосування класифікатора додаткових дерев зверніться до [32].

SVM-лінійний класифікатор:

Лінійний класифікатор SVM – це алгоритм машинного навчання з учителем, який використовується для задач класифікації. Він намагається знайти оптимальну гіперплощину для найкращого розділення класів.

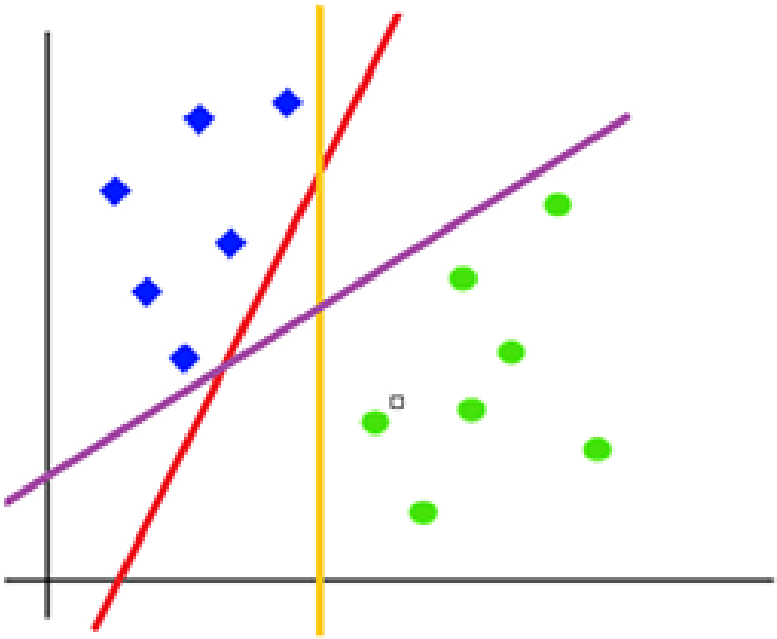


Рисунок 9 - Оптимальна гіперплощина та опорні вектори

Лінійний класифікатор SVM будує гіперплощину в n-вимірному просторі, яка категоріально розділяє різні мітки класів. Оптимальна гіперплощина характеризується найбільшою відстанню між двома класами. Ця відстань представляє відстань між гіперплощиною та найближчою точкою даних з будь-якого класу.

Функцію прийняття рішення можна сформулювати наступним чином:

$$f(x) = \mathrm{w}^T\mathrm{x} + b \tag{15}$$

Де **w** представляє вектор ваг, **x** позначає вектор ознак, а **b** означає член зміщення. Класифікація **x** до певного класу визначається знаком $f(x)$

Лінійний класифікатор SVM особливо ефективний у просторах з високою вимірністю та коли дані лінійно роздільні. Його стійкість у випадках, коли кількість вимірів перевищує кількість вибірок, робить його безцінним для певних задач класифікації, зокрема для уникнення перенавчання. У цьому дослідженні лінійний класифікатор SVM використовувався для розрізнення потенційних онлайн-покупців та тих, хто не купує, використовуючи його точність у створенні лінійних меж прийняття рішень у складних наборах даних. Для поглибленого

вивчення теоретичних основ та практичного застосування методів опорних векторів (SVM) зверніться до [32].

Класифікатор SVM-RBF:

SVM (метод опорних векторів) з ядром радіальної базисної функції (RBF) – це надійний класифікатор, відомий своєю ефективністю в обробці складних наборів даних. Це досягається шляхом перетворення вхідних ознак у простір вищих вимірів, де не лінійні роздільні точки даних стають лінійно роздільними. Цей процес дозволяє SVM виконувати складні класифікації та ефективно розрізняти закономірності в даних.

У контексті класифікаторів SVM центральною метою є визначення оптимальної гіперплощини, яка максимально розділяє класи, концепція, візуально представлена на рисунку 9. У той час як лінійний SVM шукає цю гіперплощину безпосередньо у вхідному просторі, класифікатор SVM-RBF розширює цю ідею, використовуючи функцію ядра для перетворення вхідних ознак у простір вищих вимірів. Це перетворення сприяє розділенню класів, які не є лінійно роздільними у вихідному просторі, таким чином ефективно вирішуючи проблеми нелінійності.

Ядро RBF є поширеним вибором для класифікатора SVM і визначається:

$$K(x, x') = exp(-\gamma \parallel x - x' \parallel)^2 \quad (16)$$

Де K – це функція ядра, x та x′ – два вектори ознак. Змінна ”γ” – це параметр, що визначає ширину ядра RBF. Ця функція вимірює подібність між векторами, що дозволяє SVM обробляти нелінійне розділення даних.

Класифікатор SVM-RBF відомий своєю гнучкістю в обробці випадків, коли зв'язок між мітками класів та атрибутами є нелінійним. Він особливо корисний у складних проблемних областях з недостатніми лінійними межами прийняття рішень. У цьому дослідженні класифікатор SVM-RBF було застосовано для класифікації намірів онлайн-покупок, використовуючи його

здатність фіксувати складні закономірності в наборі даних, які не одразу помітні за допомогою лінійних моделей. Для глибокого занурення в методи ядра та їх використання, зокрема SVM-RBF, зверніться до [33].

Класифікатор SVM-Poly:

SVM (метод опорних векторів) з поліноміальним (Poly) ядром – це варіант SVM, який використовує функцію поліноміального ядра для перетворення вхідного простору у простір більшої вимірності, що дозволяє йому фіксувати складні зв'язки між ознаками за допомогою поліноміальних рівнянь.

Подібно до інших класифікаторів SVM, класифікатор SVM-Poly прагне знайти оптимальну гіперплощину для розділення класів, як показано на рисунку 10. Поліноміальне ядро варіанту SVM-Poly покращує цей процес, ефективно обробляючи нелінійні дані за допомогою перетворень більшої вимірності. Цей підхід дозволяє фіксувати складні закономірності в даних, демонструючи адаптивність SVM до різних сценаріїв класифікації.

Поліноміальне ядро визначається як:

$$K(x, x') = (\gamma\langle x, x'\rangle + r)^d \quad (17)$$

Де K – це функція ядра, x та x′ – два вектори ознак, γ – масштабний коефіцієнт, r – константа, а d представляє степінь полінома. Це ядро дозволяє SVM моделювати лінійні та нелінійні зв'язки між точками даних.

Класифікатор SVM-Poly демонструє ефективність у сценаріях, що характеризуються нелінійними кореляціями між ознаками та цільовою змінною. Його застосування особливо корисне в задачах класифікації, де складність даних вимагає складної моделі, яка може адаптуватися до різних ступенів взаємодії між ознаками. У цьому дослідженні класифікатор SVM-Poly використовувався для виявлення закономірностей у поведінці онлайн-покупців, використовуючи його здатність моделювати складні поліноміальні зв'язки в даних. Для повного розуміння методів SVM, що охоплюють використання поліноміальних ядер, зверніться до джерела [34].

Класифікатор з екстремальним градієнтним підвищенням:

Класифікатор з екстремальним градієнтним підвищенням (XGB) виділяється як ефективна та масштабована реалізація методів градієнтного підвищення. Завдяки включенню більш упорядковане формулювання моделі, вона ефективно зменшує ризик перенавчання, тим самим забезпечуючи надійну продуктивність у широкому спектрі завдань та викликів у галузі науки про дані.

XGB оптимізує алгоритм машинного навчання з градієнтним підвищенням, включаючи розв'язувач лінійної моделі та алгоритми навчання деревами. Його цільова функція складається з функції втрат та члена упорядкування:

$$Obj(\Theta)\ \sum_i l(yi, \hat{}yi) + \sum_k \Omega(fk) \qquad (18)$$

Де l представляє функцію втрат, яка вимірює розбіжність між прогнозованими та фактичними значеннями ˆyi, а Ω позначає член упорядкування, що знижує складність моделі, де fk позначає k-те дерево.

Здатність XGB легко обробляти великі набори даних, його ефективність у виконанні паралельної обробки та здатність до постійного вдосконалення шляхом перехресної перевірки роблять його потужним інструментом для складних завдань класифікації. У цьому дослідженні класифікатор XGB був застосований для точного прогнозування поведінки онлайн-покупців, використовуючи його розширені функції для ефективного моделювання складнощів споживчих даних.

Класифікатор LightGBM:

LightGBM (Light Gradient Boosting Machine) постає як високоефективна та масштабована реалізація фреймворку градієнтного підвищення, що використовує алгоритми навчання на основі дерев. Розроблений для розподіленого та ефективного навчання, особливо на великих наборах даних, LightGBM впроваджує інноваційні методи, такі як GOSS (градієнтна

одностороння вибірка) та EFB (ексклюзивне об'єднання ознак). Завдяки цим методологіям LightGBM оптимізує використання ресурсів та підвищує швидкість навчання, що робить його кращим вибором для вирішення великомасштабних завдань аналізу даних з точністю та гнучкістю.

LightGBM підвищує ефективність навчання моделей та зменшує використання пам'яті завдяки своїм інноваційним методам. Метод GOSS гарантує, що алгоритм зосереджується на найбільш інформативних екземплярах, тоді як EFB зменшує розмірність даних шляхом об'єднання взаємовиключних ознак. Мету моделі можна представити так:

$$Obj = \sum_{i=1}^{N} l(yi, \hat{}yi) + \sum_{k=1}^{K} \Omega(fk) \quad (19)$$

Де N – кількість точок даних, l позначає функцію втрат, яка порівнює прогнозоване ˆyi з фактичним yi, K представляє кількість дерев, а Ω – член регуляризації, що знижує складність моделі.

LightGBM, відомий своєю ефективністю в управлінні завданнями, що потребують багато даних, чудово справляється з обробкою великих наборів даних, значно скорочуючи час навчання, зберігаючи при цьому високу точність у завданнях класифікації [35]. У цьому дослідженні LightGBM використовувався для прогнозування намірів щодо онлайн-покупок, використовуючи його можливості швидкої обробки даних та стабільну продуктивність у класифікації.

Класифікатор CatBoost:

CatBoost (Categorical Boosting) – це алгоритм градієнтного підвищення на деревах рішень, розроблений Яндексом. Цей метод адаптований для обробки категоріальних змінних, зазвичай без ретельної попередньої обробки даних, яку вимагають багато інших алгоритмів машинного навчання. Він забезпечує ефективну реалізацію, яка зменшує перенавчання та покращує точність прогнозування.

CatBoost оптимізує традиційне градієнтне підвищення, реалізуючи впорядковане підвищення, альтернативу класичному методу, керовану

перестановками, та впроваджуючи новий алгоритм для обробки категоріальних ознак. Цільова функція подібна до інших методів градієнтного бустингу, але включає покращення для зменшення перенавчання:

$$Obj = \sum_{i=1}^{N} l(yi, \hat{}yi) + \sum_{k=1}^{K} \Omega(fk) \quad (20)$$

Ідентично формулі (19), але тепер N представляє кількість вибірок, а Ω символізує складність моделі дерева.

CatBoost відомий своєю здатністю забезпечувати високу продуктивність зі стандартними налаштуваннями параметрів, що робить його доступним як для початківців, так і для досвідчених фахівців з обробки даних. Його ефективність у різних завданнях прогнозного моделювання, особливо тих, що стосуються категоріальних даних та складних наборів даних з кількома взаємодіями ознак, робить його потужним інструментом у цій галузі. У цьому дослідженні CatBoost використовувався для прогнозування намірів щодо покупок в Інтернеті, демонструючи його чудову обробку категоріальних ознак та стійкість до перенавчання [36].

Класифікатор нейронних мереж:

Класифікатори нейронних мереж використовують взаємопов'язані нейрони, що охоплюють шари, для аналізу даних та створення прогнозів. Відомі своєю адаптивністю, ці моделі чудово вловлюють складні нелінійні зв'язки в даних завдяки своїй структурованій архітектурі, що складається з вхідного шару, кількох прихованих шарів та вихідного шару.

Нейрон у нейронній мережі обробляє вхідні дані, обчислюючи зважену суму та застосовуючи нелінійну функцію активації. Цей процес для одного шару можна концептуалізувати наступним чином:

$$y = \sigma(Wx + b) \quad (21)$$

Де x – вхідний вектор, W – матриця ваг, b – вектор зміщення, σ – функція

активації, а y – вихідний вектор. Нейронні мережі вивчають оптимальні значення W та b під час процесу навчання.

Завдяки своїй здатності сприймати абстрактні поняття з даних, класифікатори нейронних мереж знаходять широке застосування в численних галузях, таких як розпізнавання зображень та мовлення, обробка природної мови тощо. У цьому дослідженні використовувався класифікатор нейронної мережі для моделювання поведінки онлайн-покупок, використовуючи його можливості глибокого навчання для виявлення закономірностей та отримання інформації зі складних та багатовимірних наборів даних [37].

Багатошаровий перцепторнний класифікатор:

Багатошаровий перцепторнний (MLP) класифікатор, форма архітектури нейронної мережі, відрізняється своїм складом з початкового вхідного шару, за яким слідує один або кілька прихованих шарів, і завершується кінцевим вихідним шаром. Така структурована конструкція дозволяє MLP вміло фіксувати складні закономірності та зв'язки даних шляхом ітеративного коригування ваг та зміщень за допомогою зворотного поширення під час навчання. Використовуючи цей механізм, MLP досягає успіху в навчанні нелінійних моделей, що робить його універсальним та потужним інструментом для широкого спектру завдань машинного навчання.

На рисунку 10 показано базову структуру MLP-NN, від вхідного шару через приховані шари до вихідного шару, демонструючи, як вона обробляє та навчається на основі складних шаблонів даних.

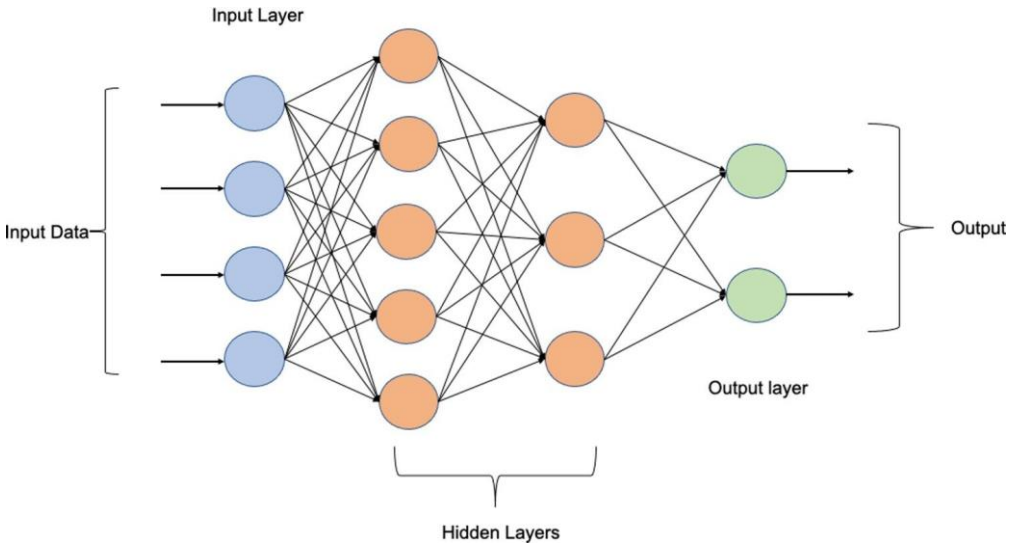


Рисунок 10 - Базова архітектура MLP-NN [37]

MLP зображує зв'язок між входами та виходами через кілька шарів вузлів, розташованих в орієнтованому графі, де кожен вузол символізує нейрон, що використовує нелінійну функцію активації. Вихід кожного нейрона визначається:

$$y = \sigma(\sum_{i=1}^{n} wixi + b) \quad (22)$$

Де xi представляє вхідні значення, wi позначає ваги, b означає зміщення, а σ представляє функцію активації. Мережа коригує ваги та зміщення, щоб мінімізувати різницю між прогнозованими та фактичними результатами, і найкращі параметри необхідно вивчити.

Класифікатори MLP здатні вирішувати складні задачі класифікації, де межі лінійного прийняття рішень не відповідають вимогам. Завдяки своїй здатності фіксувати складні шаблони у багатовимірних наборах даних, вони виділяються в завданнях, що включають розпізнавання шаблонів, розпізнавання мовлення та класифікацію зображень. У цьому дослідженні класифікатор MLP був використаний для ретельного вивчення та прогнозування намірів щодо покупок в Інтернеті, використовуючи його систему глибокого навчання для ефективної обробки та моделювання нюансованої природи даних про поведінку споживачів. Для ретельного аналізу архітектур нейронних мереж, що охоплюють MLP та

інші моделі, описані у [37].

Багатошарова модель MLP:

Багатошарова модель MLP (багатошаровий перцептрон) або ансамбль стекування – це вдосконалена стратегія машинного навчання, яка об'єднує кілька класифікаторів MLP для створення складної прогнозної моделі. Поєднуючи прогнози різних моделей MLP, ця техніка використовує їхні сильні сторони та компенсує їхні слабкі сторони, підвищуючи загальну точність та стійкість. Процес включає навчання базових моделей MLP на наборі даних та використання їхніх прогнозів як вхідних даних для метамоделі або метанавчання. Цей метанавчання оцінює результати базових моделей, щоб визначити найкращу комбінацію для точних прогнозів. Завдяки цьому ієрархічному підходу, багатошарова модель MLP пропонує більш нюансований та потужний засіб розуміння складних даних, значно перевершуючи окремі класифікатори MLP.

Накладання поєднує прогнози з кількох базових моделей шляхом навчання нової моделі. Моделі базового рівня навчаються на всьому наборі даних, а їхні прогнози служать вхідними даними для моделі другого рівня для остаточного прогнозу. Цей процес можна підсумувати наступним чином:

$$y = fmeta(f1(x), f2(x), \ldots, fn(x)) \quad (23)$$

де f1, f2,..., fn – базові моделі MLP, fmeta – мета-модель навчання, x – вектор вхідних ознак, а ˆy – прогнозований вихід.

Складені моделі MLP виявляються безцінними в сценаріях, коли одна модель може лише частково охоплювати деякі аспекти даних. У цьому дослідженні було реалізовано складну модель MLP для підвищення точності прогнозування намірів щодо онлайн-покупок. Архітектура моделі включала: - П'ять базових моделей, кожна з послідовністю шарів, призначених для початкової незалежної обробки набору даних. Ці моделі містили вхідний шар,

приховані шари з різною кількістю нейронів (16, 12, 8 та 4) та сигмоїдний вихідний шар для двійкової класифікації. - Проміжна модель MLP, навчена на об'єднаних прогнозах базових моделей, додатково уточнювала представлення даних за допомогою своїх шарів (10 та 8 нейронів у прихованих шарах). - Мета-модель MLP діяла як кінцевий агрегат, використовуючи свою структуру (10 нейронів у прихованому шарі) для асиміляції вихідних даних проміжної моделі в кінцевий прогноз.

Цей багаторівневий, багатомодульний підхід значно покращив прогностичну ефективність, використовуючи різні можливості навчання кожної моделі MLP, демонструючи виняткову ефективність у розпізнаванні складних закономірностей у даних про поведінку онлайн-покупців.

3.3 Валідація та оцінка

У цьому розділі буде обговорено вирішальний процес валідації та оцінки ефективності моделей. Для ретельної оцінки ефективності моделей використовується набір показників, таких як точність, прецизійність, повнота та бал F1. Ці показники пропонують безцінне розуміння здатності моделей робити точні прогнози для різних класів та сценаріїв. За допомогою суворих методів валідації, таких як перехресна валідація, гарантується, що моделі є надійними та можуть добре узагальнюватися на нові дані. Давайте розглянемо ці показники оцінки та їх значення у вимірюванні ефективності моделей класифікації.

Точність, повнота, акуратність та бал F1 критично оцінюють моделі класифікації, підкреслюючи ефективність та надійність. Ці показники, що відображають результати прогнозування, пропонують детальне уявлення про продуктивність моделі та спрямовують її на покращення. Аналіз цих ключових показників допомагає визначити області для вдосконалення та покращення, що призводить до підвищеної точності моделі. Така ретельна оцінка забезпечує розробку ефективніших та надійніших моделей класифікації.

Акуратність (Accuracy)

Акуратність, що визначається як частка точно передбачених спостережень серед усього набору даних, забезпечує повний огляд правильності моделі.

$$\text{Акуратність} = \frac{\text{Істинні позитивні} + \text{Істинні негативні}}{\text{Загальна кількість спостережень}} \quad (24)$$

Точність (Precision)

Точність, також відома як позитивне прогностичне значення, обчислює частку правильно передбачуваних позитивних спостережень відносно загальної кількості передбачуваних позитивних результатів.

$$\text{Точність} = \frac{\text{Істинно позитивні результати}}{\text{Істинно позитивні результати} + \text{Хибнопозитивні результати}} \quad (25)$$

Повнота (Recall)

Повнота, або Чутливість, обчислює відношення правильно передбачених позитивних спостережень до всіх спостережень, що справді належать до позитивного класу (Так).

$$\text{Повнота} = \frac{\text{Істинно позитивні результати}}{\text{Істинно позитивні результати} + \text{Хибнопозитивні результати}} \quad (26)$$

Оцінка F1 (F1-Score)

Оцінка F1 визначається як середньозважене значення точності та повноти. Отже, вона враховує хибнопозитивні та хибнонегативні результати, що є перевагою в ситуаціях, коли вартість цих помилок відрізняється

$$\text{Оцінка } F1 = 2 \times \frac{\text{Точність} \times \text{Повність}}{\text{Точність} + \text{Повність}} \quad (27)$$

Ці показники надають комплексне уявлення про продуктивність моделі,

виділяючи різні аспекти. Точність дає швидке уявлення про загальну продуктивність, але в незбалансованих класах,

Precision (точність), Recollect (повністю) та F1 Score (оцінка F1) пропонують глибше розуміння здатності моделі правильно передбачати кожен клас.

3.4 Огляд ефективності моделей

Матриця плутанини – це корисний інструмент для оцінки ефективності моделі класифікації на тестовому наборі даних з відомими істинними значеннями. Вона допомагає візуалізувати продуктивність алгоритму.

Таблиця 4: Матриця плутанини

Актуальність	**Позитивно**	**Негативно**
Позитивно	**TP**	**FN**
Негативно	**FP**	**TN**

• TP (істинно позитивний) – це правильно передбачені позитивні спостереження.

• FN (хибнонегативний) – це позитивні спостереження, помилково передбачені як негативні.

• FP (хибнопозитивний) – це негативні спостереження, помилково передбачені як позитивні.

• TN (істинно негативний) – це правильно передбачені негативні спостереження.

Ця матриця допомагає візуалізувати продуктивність моделі класифікації, зокрема її здатність передбачати справжні позитивні та негативні результати, уникаючи хибнопозитивних та негативних.

Дослідження проводилося суворо відповідно до етичних принципів, забезпечуючи суворі заходи щодо конфіденційності даних. Дані користувачів були анонімізовані, а протокол дослідження схвалено інституційною

експертною радою.

У цьому дослідженні використовувалися різноманітні програмні інструменти для проведення аналізу даних та розробки моделей. Рішення використовувати Python як основну мову програмування було обумовлено її адаптивністю та широким спектром бібліотечних ресурсів. Для маніпулювання даними та обчислювальних завдань використовувалися основні бібліотеки, що дозволяло ефективно обробляти числові дані та завдання маніпулювання даними.

Інформацію про дані візуалізували за допомогою потужних бібліотек побудови графіків, що сприяло створенню візуально привабливих та інформативних діаграм і графіків. У машинному навчанні були впроваджені різні алгоритми та моделі для дослідження прогностичних закономірностей у даних. Завдання глибокого навчання вирішувалися за допомогою передових фреймворків, що забезпечувало гнучкість та масштабованість для побудови та навчання архітектур нейронних мереж.

Були включені додаткові утиліти для оптимізації процесу аналізу, забезпечуючи безперебійний робочий процес від дослідження даних до оцінки моделі. Загалом, цей комплексний набір програмних інструментів дозволив застосувати ретельний та багатогранний підхід до аналізу даних, сприяючи отриманню корисної інформації та розробці надійних прогностичних моделей.

Оцінювання ефективності моделей машинного навчання у реальних умовах є вирішальним для визначення їхньої практичної цінності. У цьому розділі проведено детальний аналіз продуктивності різних класифікаційних моделей, застосованих до набору даних. Аналіз охоплює широкий спектр метрик, зокрема accuracy, precision, recall, F1-score та ROC-AUC. Ці показники забезпечують всебічне уявлення про здатність моделей правильно класифікувати приклади та розрізняти позитивні й негативні випадки.

Таблиця 5: Показники ефективності

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.642	0.256	0.856	0.394	0.732

KNN	0.755	0.313	0.671	0.427	0.720
Decision Tree	0.845	0.449	0.617	0.520	0.749
Random Forest	0.906	0.636	0.716	0.673	0.826
Gradient Boosting	0.900	0.616	0.698	0.654	0.815
AdaBoost	0.882	0.545	0.778	0.641	0.838
Extra Trees	0.896	0.604	0.686	0.642	0.808
SVM-Linear	0.879	0.537	0.784	0.637	0.839
SVM-RBF	0.875	0.532	0.644	0.583	0.777
SVM-Poly	0.781	0.356	0.754	0.484	0.770
XGBoost	0.904	0.641	0.662	0.651	0.802
LightGBM	0.906	0.644	0.689	0.666	0.814
CatBoost	0.913	0.681	0.671	0.676	0.811
Neural Network	0.862	0.495	0.725	0.588	0.804
Multilayer Perceptron	0.851	0.462	0.581	0.515	0.737
Stacked MLP Model	0.910	0.703	0.581	0.636	0.771

Детальний аналіз ефективності кожної моделі подано після представлення метрик продуктивності в Таблиці 5. Цей аналіз охоплює сильні сторони, слабкі сторони та загальну результативність моделей у розв’язанні поставленої задачі. Завдяки ретельному оцінюванню отриманих результатів ставиться за мету зрозуміти можливості кожної моделі й визначити потенційні напрями для їх поліпшення. Така критична оцінка слугує основою для ухвалення обґрунтованих рішень щодо вибору та впровадження моделей машинного навчання у практичних застосуваннях.

3.5 Інформація з візуалізації даних

Цей розділ зосереджується на висновках, отриманих завдяки візуалізації даних. Через різні графічні представлення вдалося виявити приховані закономірності, дослідити кореляції та отримати цінне розуміння поведінки користувачів на платформі електронної комерції.

Візуалізація показників *bounce* та *exit* дає цікаві результати щодо залученості користувачів і їхньої навігаційної поведінки на сайті. Аналізуючи Рисунок 2, який демонструє розподіли показників bounce та exit, можна помітити

частоту та закономірності взаємодій користувачів. Зокрема, ширший розподіл *exit rate* порівняно з *bounce rate* натякає на варіативність рівнів задоволеності та зацікавленості залежно від категорій сторінок. Це розуміння, отримане за допомогою метрик вебаналітики, є важливим для вдосконалення користувацького досвіду та оптимізації контенту задля підвищення залученості та утримання відвідувачів.

Розуміння розподілу класів доходу в межах датасету має важливе значення для виявлення трендів і закономірностей, пов'язаних із результатами транзакцій, але більшість сеансів не приводять до транзакцій. Такий дисбаланс класів створює певні труднощі під час навчання моделей і підкреслює важливість використання надійних методів класифікації. Аналітичні інструменти надають цінні результати щодо ефективності маркетингових стратегій, привабливості товарів і загальної взаємодії користувачів із сайтом. Використовуючи ці дані, платформи можуть краще адаптувати свої підходи до потреб клієнтів, оптимізувати коефіцієнти конверсій та забезпечити сталий розвиток.

Дослідження взаємозв'язку між показниками bounce, exit, цінністю сторінки та типом трафіку щодо статусу доходу показує взаємозв'язок з факторами, що впливають на ухвалення рішення про покупку, а також демонструє взаємодію між змінними та їхній комбінований вплив на формування доходу. Розуміння таких кореляцій дозволяє бізнесу стратегічно вдосконалювати структуру сайту й адаптувати маркетингові підходи для підвищення залученості користувачів та збільшення конверсії.

Аналіз розподілу типів відвідувачів надає цінну інформацію про різні моделі поведінки користувачів. Рисунок 4 показує поділ користувачів на нових відвідувачів, постійних клієнтів і гостьових користувачів. Помітно, що значну частку трафіку формують саме постійні клієнти, що підкреслює важливість підтримки лояльності та утримання через персоналізований досвід і цільові маркетингові кампанії. Розуміння динаміки взаємодії відвідувачів є важливим для адаптації стратегій під унікальні потреби різних сегментів, підвищення задоволеності та стимулювання бізнес-зростання.

Аналіз транзакцій у спеціальні дні забезпечує комплексне розуміння поведінки споживачів у періоди пікових покупок та свят. Інфографіка в Рисунку 6 ілюструє дані про пошукові запити, пов’язані з товарами, що приводять до транзакцій, підкреслюючи вплив особливих подій. Дослідження того, як споживачі витрачають кошти в такі періоди, дає корисну інформацію щодо мотивацій покупок і товарних преференцій. Використання цих знань може допомогти у стратегічному плануванні та розробці рекламних кампаній, спрямованих на максимізацію сезонних можливостей, збільшення продажів і підвищення задоволеності клієнтів.

Аналіз карти кореляцій, наведеної в Рисунку 7, виявляє помітні асоціації між різними числовими атрибутами. Розуміння цих кореляцій є важливим для визначення потенційних синоптиків поведінки користувачів і вибору ознак для побудови моделей машинного навчання. Деякі атрибути демонструють сильні позитивні чи негативні зв’язки, що вказує на напрямки для глибшого аналізу та оптимізації.

3.6 Порівняння та вибір моделей

У цьому розділі порівнюється продуктивність шістнадцяти різних моделей машинного навчання, застосованих до датасету. Їх ефективність оцінюється на основі метрик accuracy, precision, recall, F1-міри та ROC-AUC. Подальший аналіз надає інформацію щодо сильних і слабких сторін кожної моделі, а також підкреслює компроміси між різними показниками продуктивності.

Логістична регресія

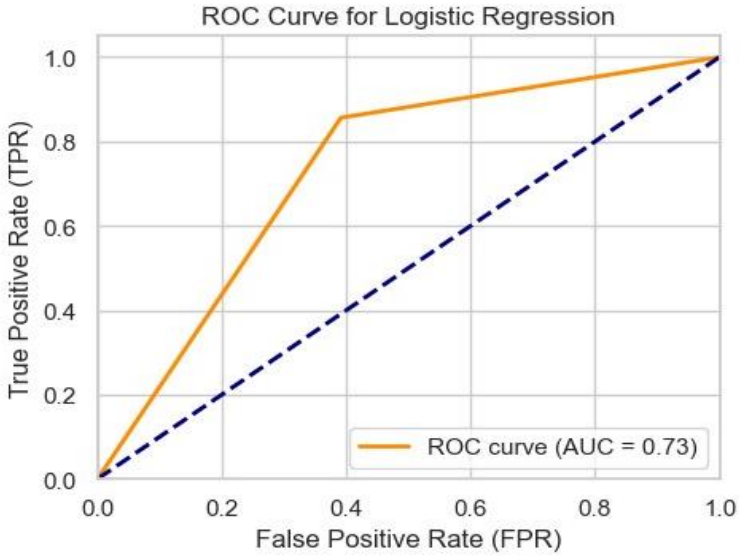


Рисунок 11 - Логістична регресія

Класифікатор KNN

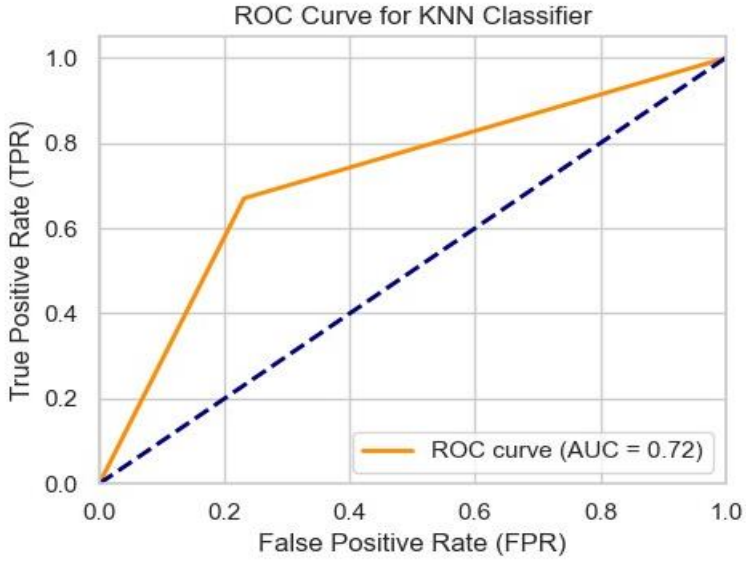


Рисунок 12 - Класифікатор KNN

Класифікатор дерева рішень

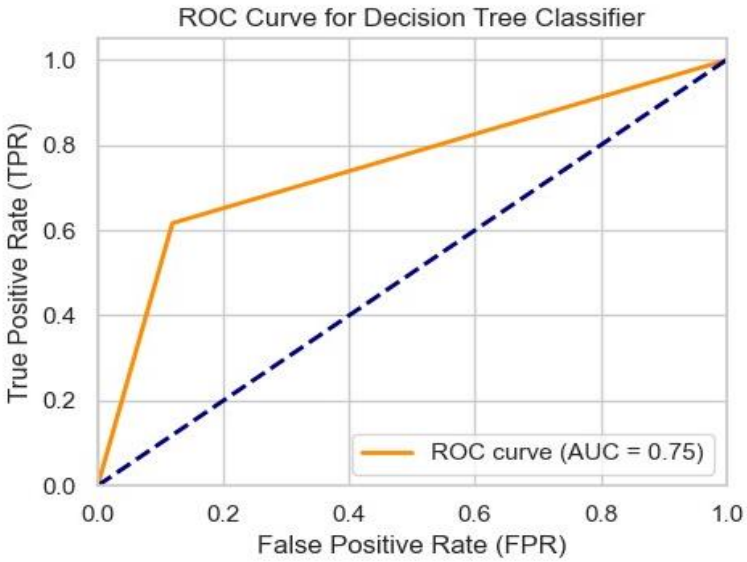


Рисунок 13 - Класифікатор дерева рішень

Класифікатор випадкового лісу

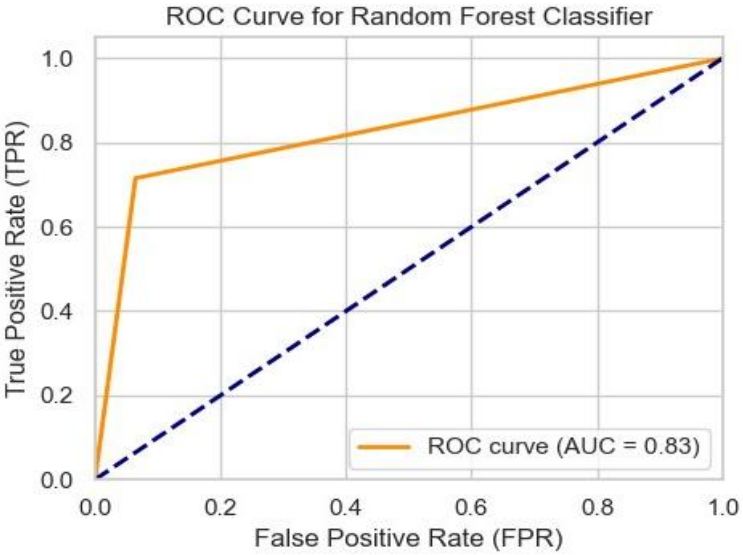


Рисунок 14 - Класифікатор випадкового лісу

Класифікатор градієнтного підвищення

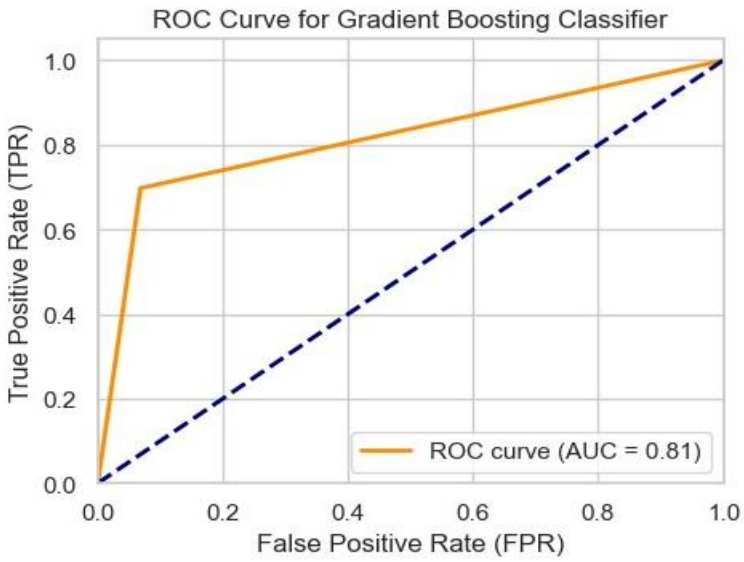


Рисунок 15 - Класифікатор градієнтного підвищення

Класифікатор AdaBoost

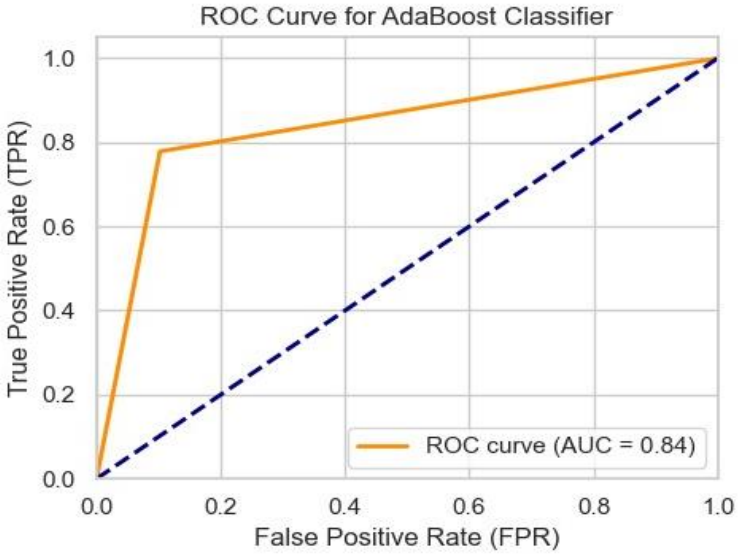


Рисунок 16 - Класифікатор AdaBoost

Класифікатор додаткових дерев

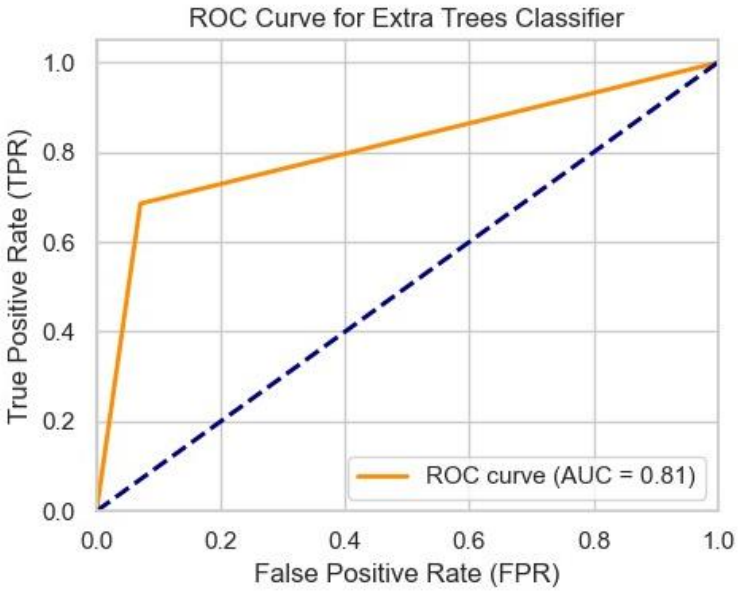


Рисунок 17 - Класифікатор додаткових дерев

SVM-лінійний класифікатор

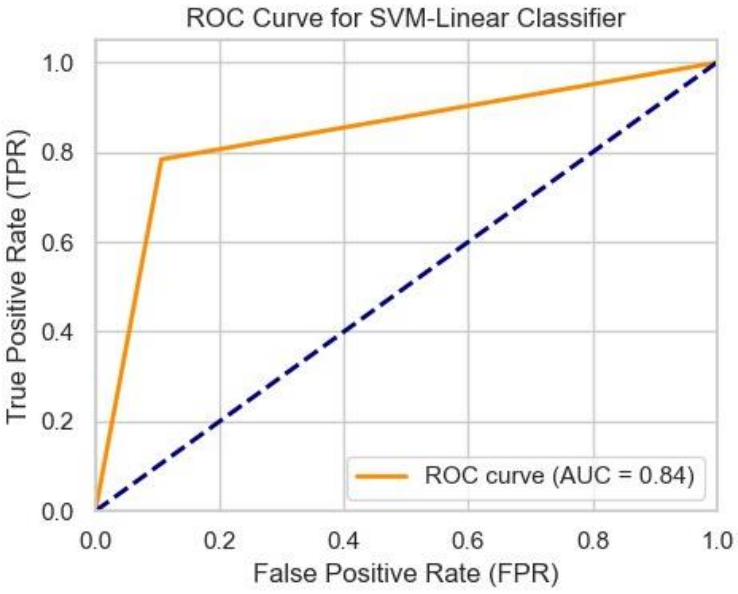


Рисунок 18 - SVM-лінійний класифікатор

Класифікатор SVM-RBF

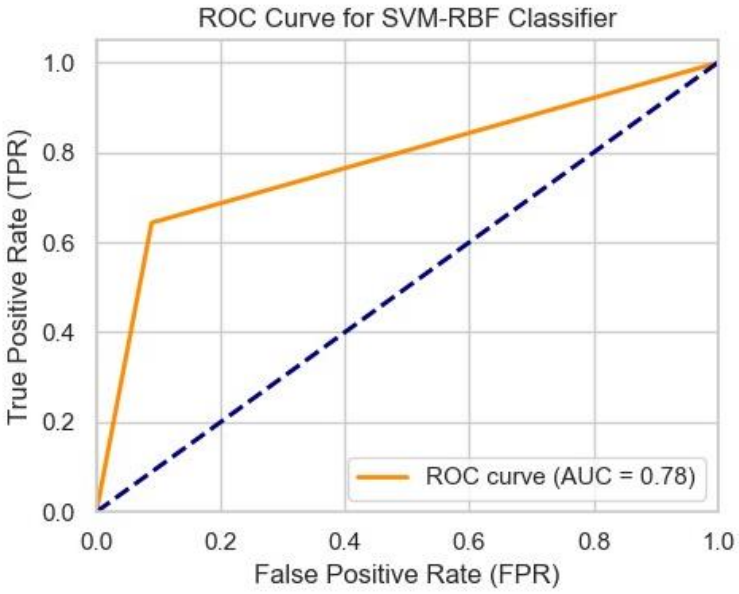


Рисунок 19 - Класифікатор SVM-RBF

Класифікатор SVM-Poly

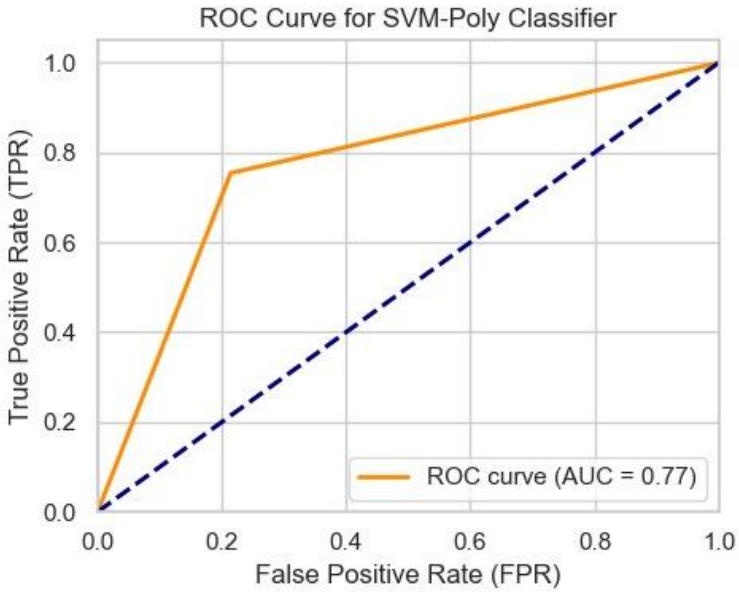


Рисунок 20 - Класифікатор SVM-Poly

Класифікатор Екстремального Підсилення Градієнта

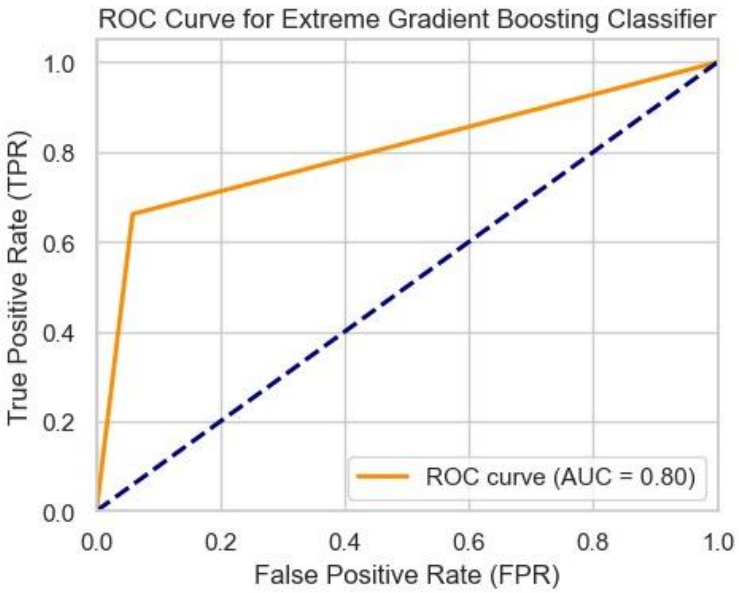


Рисунок 21 - Класифікатор Екстремального Підсилення Градієнта

Класифікатор LightGBM

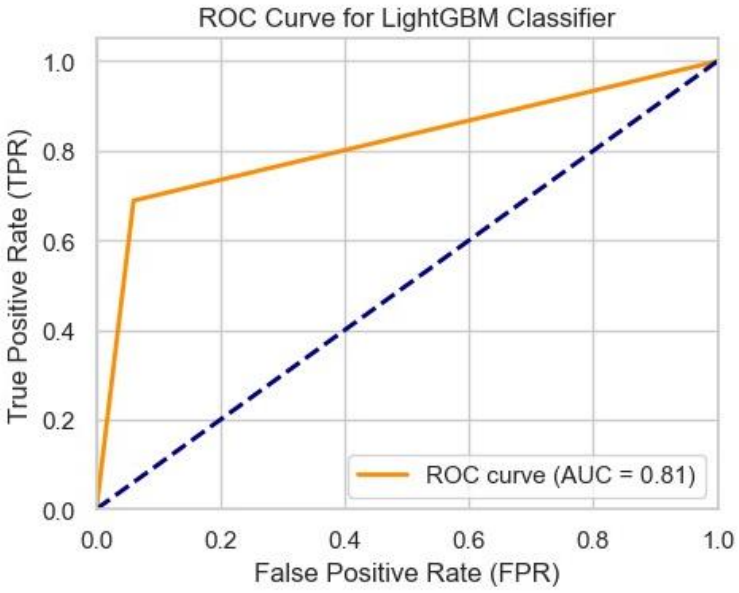


Рисунок 22 - Класифікатор LightGBM

Класифікатор CatBoost

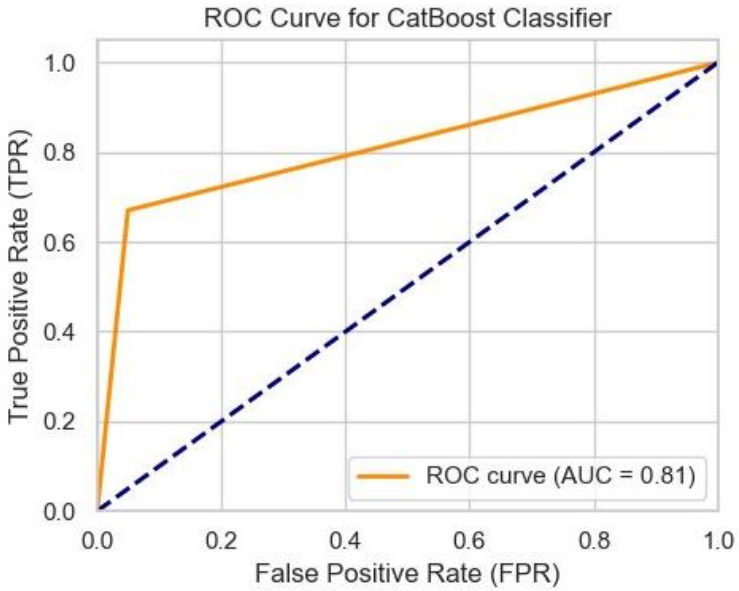


Рисунок 23 - Класифікатор CatBoost

Класифікатор нейронних мереж

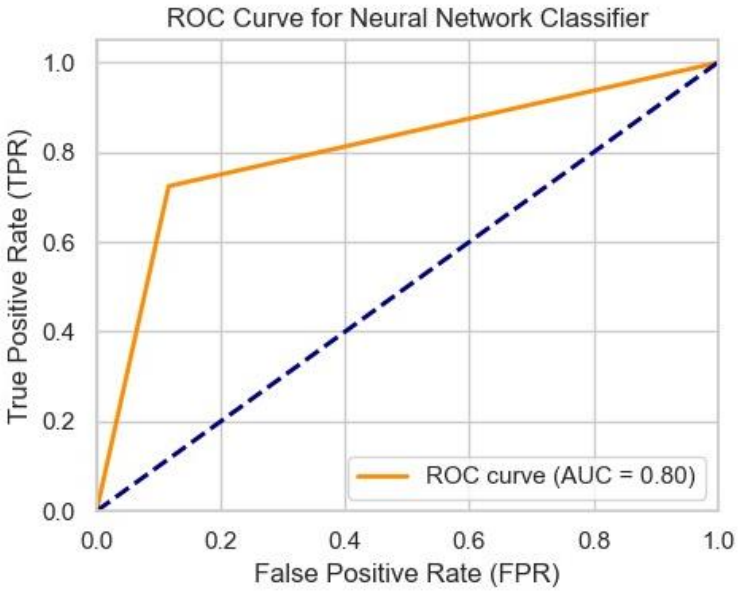


Рисунок 24 - Класифікатор нейронних мереж

Багатошаровий класифікатор перцепторів

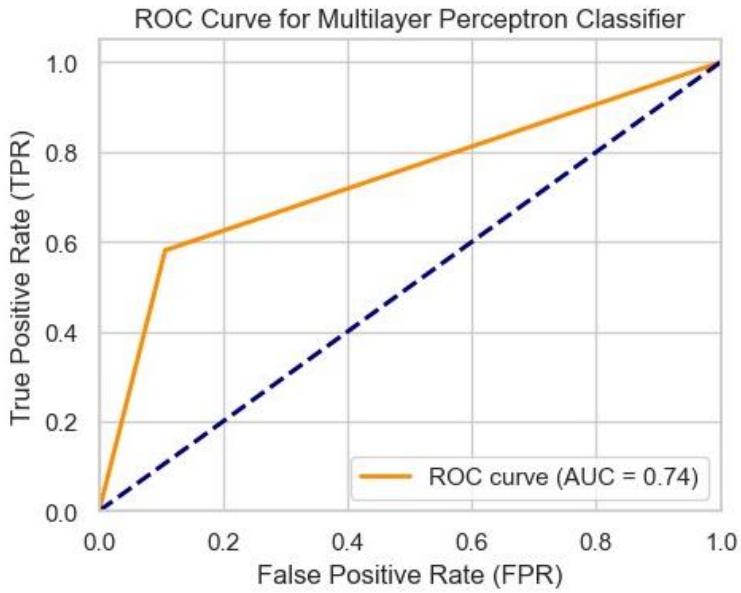


Рисунок 25 - Багатошаровий класифікатор перцепторів

Модель MLP зі стеком

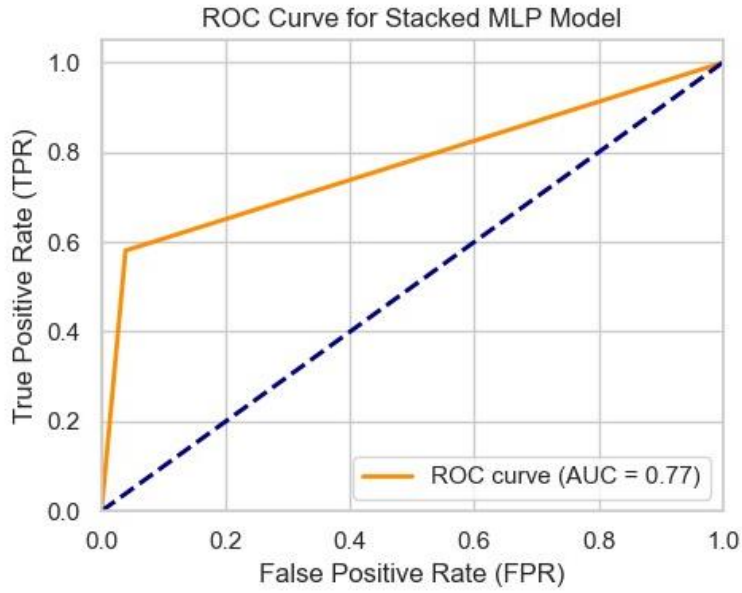


Рисунок 26 - Модель MLP зі стеком

Для надання всебічного огляду продуктивності оцінених класифікаторів представлено візуалізації, що порівнюють їхню точність та показники ROC-AUC. Перша візуалізація, наведена на Рисунку 27, безпосередньо порівнює якість класифікації, досягнуту кожною моделлю. Друга візуалізація на Рисунку

28 зіставляє точність і значення ROC-AUC, щоб підкреслити компроміси між цими ключовими метриками. Ці візуалізації забезпечують зручне резюме продуктивності класифікаторів, що полегшує визначення найефективніших моделей і вибір найбільш придатного класифікатора для даного набору даних.

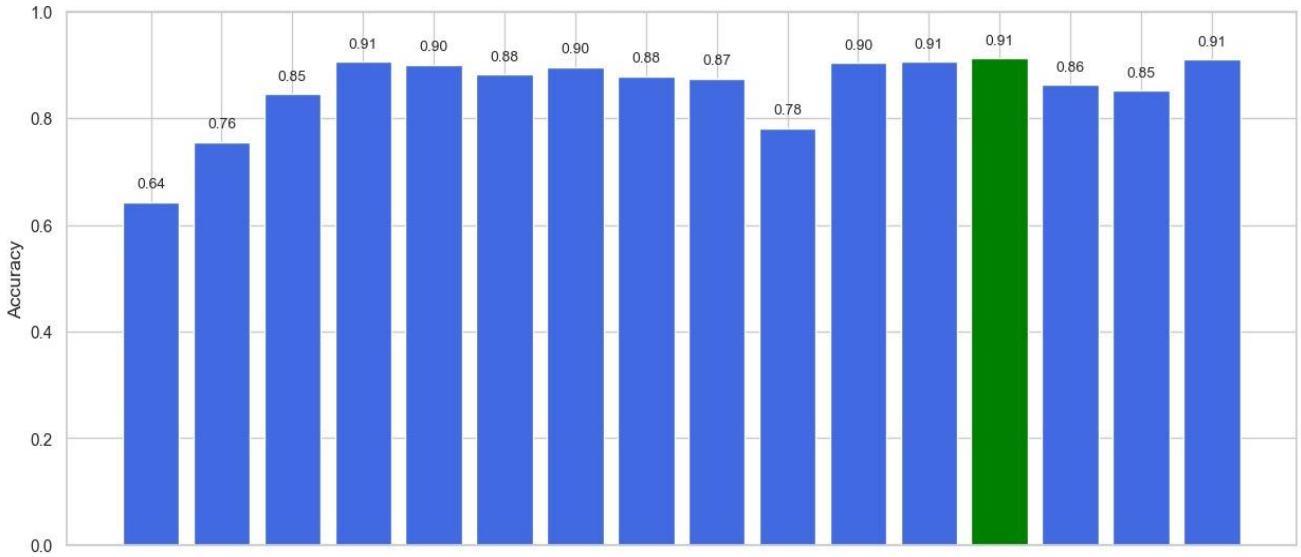


Рисунок 27 - Порівняння точності для класифікаційних моделей

3.7 Узагальнення результатів

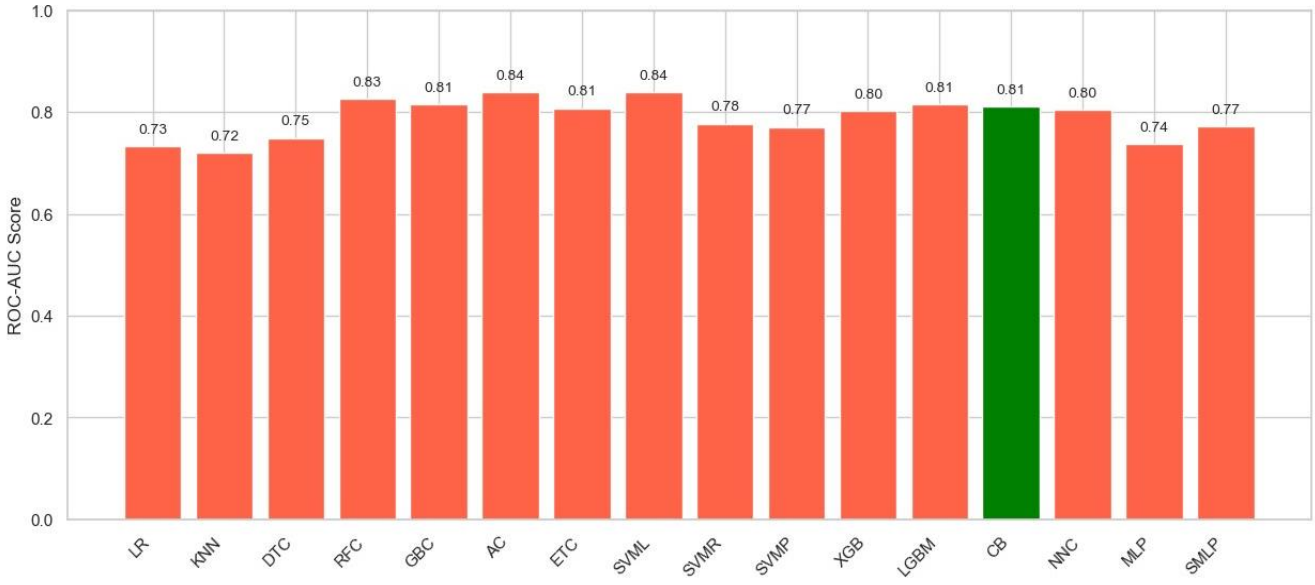


Рисунок 28 - Акуратність і ROC-AUC

Під час оцінювання ефективності різних моделей машинного навчання для прогнозування поведінки користувачів в онлайн-шопінгу стало очевидно, що їхня масштабованість має значний потенціал для застосувань у сфері великих даних. Хоча набір даних, використаний у цьому дослідженні, є досить

всеосяжним, його розмір і структура відображають характер більших та складніших масивів даних, властивих екосистемі big data в електронній комерції. Методи та моделі, застосовані в роботі, є адаптивними та такими, що масштабуються, що свідчить про можливість їх розширення для аналізу великих обсягів різнорідних даних, типовим для сценаріїв big data. Така потенційна масштабованість означає, що отримані результати можуть бути застосовані й у ширших контекстах великих даних, де вміння обробляти та аналізувати дані в реальному часі є критично важливим для отримання практичних інсайтів та точного прогнозування поведінки споживачів.

Попередні розділи надали всебічну оцінку моделей машинного навчання для аналізу поведінки користувачів в онлайн-шопінгу. Це узагальнення відіграло ключову роль у визначенні найбільш ефективних моделей шляхом використання відповідного набору даних та застосування різних класифікаторів з оцінюванням на основі таких метрик, як точність, прецизійність, повнота, F1-міра та ROC-AUC. Отримані результати забезпечили глибоке розуміння здатності цих моделей і їхньої придатності для аналізу та прогнозування поведінки споживачів в онлайн-середовищі.

Аналіз охопив широкий спектр моделей машинного навчання — від базових класифікаторів, таких як логістична регресія та KNN, до більш складних ансамблевих методів, зокрема Random Forest та Gradient Boosting. Серед них особливо вирізнявся CatBoost, який продемонстрував найвищу ефективність. Додатково були розглянуті нейронні мережі, для яких було виявлено значну варіативність результатів у контексті прогнозування поведінки онлайн-покупців. Основні висновки:

- Random Forest і Gradient Boosting продемонстрували високу ефективність за різними метриками. Їхня здатність обробляти складні задачі онлайн-прогнозування була очевидною завдяки високій точності та здатності добре розмежовувати класи.

- CatBoost показав себе найкраще, що зумовлено його алгоритмічними

перевагами в роботі з категоріальними даними та великими просторами ознак. Він став одним з провідних моделей у дослідженні.

- Logistic Regression та KNN - забезпечили початкове уявлення про структуру даних, проте стабільно поступалися складнішим алгоритмам за показниками точності та F1-міри, що підкреслює потребу у використанні просунутих методів для вирішення подібних складних задач.

Інсайти та наслідки:

- Успіх ансамблевих методів і складних нейронних мереж у прогнозуванні онлайн-поведінки користувачів свідчить про їхню здатність ефективно моделювати складні, нелінійні залежності в даних.

- Різні моделі продемонстрували різні сильні сторони, що підкреслює важливість ретельного вибору моделі відповідно до вимог e-commerce, включаючи характеристики даних, масштабованість, інтерпретованість, можливість роботи в реальному часі, вимоги до безпеки й адаптивності.

- Існує очевидний компроміс між показниками precision і recall. Бізнесам потрібно правильно врівноважити ці метрики під час вибору моделі, враховуючи свої стратегічні цілі та особливості доступних даних.

3.7.1 Розширена предикативна система для електронної комерції

Методологічні підходи, описані в працях [27] та [29], відіграли ключову роль у формуванні запропонованої системи. Ця система є прикладом інтеграції стратегій машинного навчання, що поєднують проактивні та реактивні можливості впливу на онлайн-поведінку споживачів. Використовуючи аналітику в режимі реального часу, система динамічно персоналізує взаємодію з користувачем, адаптуючись до його миттєвих поведінкових сигналів.

Рисунок 29 демонструє детальний робочий процес нової системи, відображаючи інтеграцію ключових дослідницьких компонентів із сучасними технологіями для створення динамічної, предикативної платформи електронної комерції.

Спираючись на аналітичні можливості моделі CatBoost пропонується вдосконалена система для прогнозування намірів онлайн-покупців у режимі реального часу.

Ця система поєднує методи, отримані з наукових досліджень, сучасні досягнення в галузі предикативної аналітики, алгоритми штучного інтелекту, розроблені для роботи в реальному часі.

Запропонована архітектура спеціально розроблена для задоволення зростаючих вимог e-commerce, враховуючи потребу в швидкій адаптації до поведінки користувача, масштабованій обробці даних, високоточному прогнозуванні, персоналізації та автоматизації взаємодії на вебплатформі.

3.7.2 Деталізація робочого процесу прогнозуючої системи

Запропонована система оптимально поєднує аналітичну точність класифікатора CatBoost із здатністю LSTM-RNN розпізнавати часові шаблони, утворюючи високорозвинену прогнозуючу архітектуру. Нижче подано покрокове пояснення роботи системи:

1. **Авторизація користувача та отримання даних:**
   Робота системи починається з моменту входу користувача, після чого негайно здійснюється збір сесійних та персоналізованих даних. Ці дані є ключовими для формування індивідуалізованого користувацького досвіду.

2. **Виявлення намірів у режимі реального часу:**
   Модель CatBoost миттєво оцінює ймовірність здійснення покупки на основі отриманих даних. Метою є підвищення залученості користувача

шляхом ідентифікації високої ймовірності покупки та демонстрації відповідного контенту.

3. **Стратегія взаємодії та контенту:**
   Якщо явного наміру здійснити покупку не виявлено, система переходить у режим спостереження, відслідковуючи дії користувача для фіксації нових поведінкових сигналів. У разі появи таких сигналів система готова миттєво надати релевантний персоналізований контент для стимулювання взаємодії.

4. **Аналіз імовірності переривання сесії:**
   Модель LSTM-RNN аналізує поведінкові шаблони для прогнозування ризику завершення сесії. Це дозволяє системі завчасно ініціювати заходи для утримання користувача від виходу.

5. **Персоналізація за допомогою сучасних технологій:**
   Для користувачів із чітко вираженим наміром придбати товар система підсилює взаємодію за допомогою технологій AI, AR та голосового/візуального пошуку, що підвищує якість користувацького досвіду.

6. **Цілісність та безпека транзакцій:**
   Під час завершення транзакції запускається механізм перевірки на основі блокчейну, який гарантує безпеку, прозорість та цілісність операції, підвищуючи довіру користувачів.

7. **Безперервне навчання та вдосконалення системи:**
   Ключовим компонентом системи є циклічний механізм зворотного зв’язку, що навчається на кожній взаємодії. Це забезпечує постійне вдосконалення точності прогнозів та якості користувацького досвіду.

Цей ретельно спроектований робочий процес обіцяє динамічний та чутливий користувацький досвід, який постійно розвивається через взаємодію користувача. У центрі уваги — прагнення до безперервного вдосконалення, що забезпечує стійку точність прогнозування та залучення користувачів.

3.8 Пропонована інтеграція та застосування

Інтеграція цієї системи в платформу електронної комерції може стати каталізатором трансформаційних змін в еру динамічних та адаптивних онлайн-шопінг-досвідів. Інновація адаптує шлях покупця до тонких поведінкових особливостей клієнтів, фіксуючи відгук на негайні взаємодії та одночасно враховуючи загальні ринкові тенденції та змінні споживчі вподобання. Здатність системи використовувати прогностичну аналітику в реальному часі забезпечує більш захоплюючий досвід користувача, застосовуючи дані для утримання уваги та лояльності клієнтів. Вона прагне переглянути галузеві стандарти використання прогностичної аналітики в секторі електронної комерції, роблячи процес шопінгу більш персоналізованим та передбачуваним щодо майбутніх споживчих тенденцій.

Переваги такої інтеграції численні: вдосконалення взаємодій користувачів завдяки аналітиці в реальному часі, оптимізація маркетингових зусиль для підвищення ефективності та значне збільшення коефіцієнта конверсії. Такий підхід дозволяє покращити маркетингові стратегії шляхом прогнозування майбутніх покупок, забезпечуючи більш цілеспрямовані та ефективні кампанії. Крім того, підвищення залученості користувачів через персоналізовані рекомендації сприяє формуванню більш лояльної клієнтської бази.

Очікувані вдосконалення передбачають перехід до середовища шопінгу, більш орієнтованого на індивідуального покупця та більш чутливого до змін у поведінці онлайн-споживачів. Цей перехід підкреслює важливість інтуїтивного, орієнтованого на дані підходу до електронної комерції, пропонуючи майбутнє, в

якому онлайн-шопінг інтегрований із особистими вподобаннями та поведінкою споживачів.

Водночас слід враховувати, що масштаби цих переваг залежать від ретельності впровадження системи та конкретного контексту її використання, включно з технологічною інфраструктурою та характеристиками цільової аудиторії.

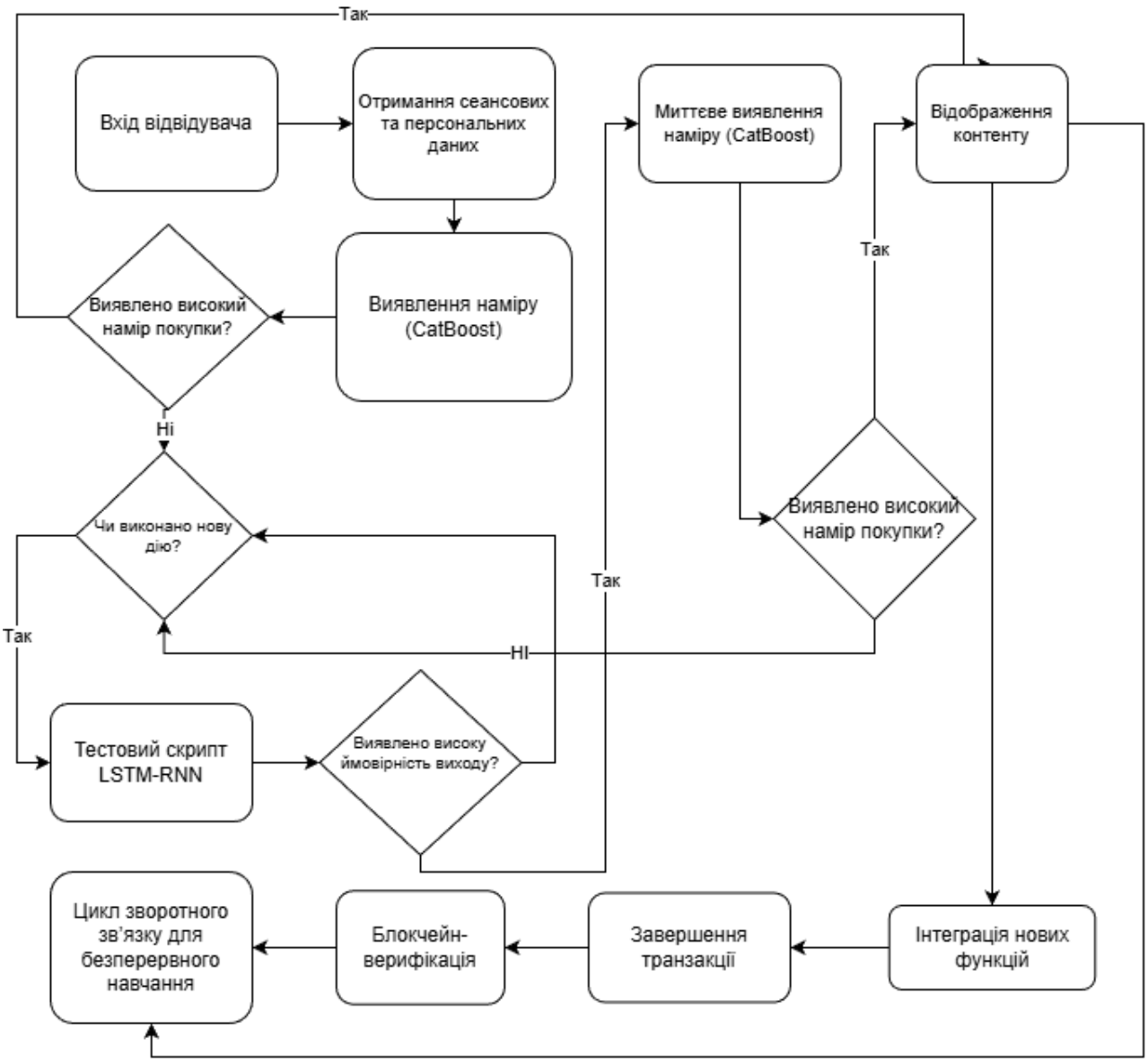


Рисунок 29 - Блок-схема інтеграції розширених методологій

У наступних розділах буде детально розглянуто практичні аспекти впровадження системи, її оцінку ефективності та перспективи масштабованості. Також буде проаналізовано потенційну інтеграцію системи з існуючими

технологіями електронної комерції для розширення функціональності та покращення користувацького досвіду.

Поєднання моделей машинного навчання, таких як класифікатор CatBoost, у системах прогнозування електронної комерції свідчить про трансформаційний зсув в онлайн-торгівлі. Впровадження цієї технології перевершує дуже технічні вдосконалення, сприяючи створенню інтелектуальних, адаптивних та високо персоналізованих середовищ для покупок. Наслідки такої інтеграції є далекосяжними та багатогранними, впливаючи на різні аспекти цифрового роздрібного простору:

Розширена персоналізація та користувацький досвід: Застосування поглибленої поведінкової аналітики дозволяє створювати персоналізований користувацький досвід. Це виходить за рамки простих рекомендацій щодо продуктів і охоплює маркетингові комунікації, динамічне ціноутворення та індивідуальні користувацькі інтерфейси. Таке точне розуміння та передбачення уподобань клієнтів передбачає нову еру в залученні та задоволеності клієнтів.

Операційна ефективність та ефективність ланцюга поставок: Прогнозна аналітика обмежується перевагами, орієнтованими на клієнтів, і поширюється на закулісні операції, такі як управління запасами та логістикою, підвищуючи ефективність ланцюга поставок. Точні прогнози попиту призводять до кращого рівня запасів, зменшення відходів та швидшої доставки, сприяючи сталому, економічно ефективному операційному моделі.

Обґрунтовані стратегічні рішення: Аналітика прогнозної аналітики надає підприємствам можливості прийняття рішень на основі даних. Це сприяє розумному та стратегічному реагуванню на зміни ринку, відгуки клієнтів та конкурентне середовище, забезпечуючи передову присутність на ринку.

Покращена безпека та запобігання шахрайству: Оскільки цифрові

транзакції стають більш поширеними, важливість безпеки зростає. Передові моделі машинного навчання, такі як CatBoost, покращують виявлення шахрайської діяльності, виявляючи нерегулярні закономірності та аномалії, захищаючи онлайн-транзакції та дані споживачів. Заглядаючи в майбутнє, подальша інтеграція цих моделей у платформи електронної комерції готова переосмислити ефективність, безпеку, користувацький досвід і, зрештою, майбутнє онлайн-шопінгу.

Незважаючи на прогрес у застосуванні машинного навчання в електронній комерції, існують певні обмеження, які слід подолати для повного реалізації потенціалу цих технологій. Майбутні дослідження відкривають численні можливості для подолання поточних викликів та розширення можливостей платформ електронної комерції:

Адаптивність між різними сферами: Наразі здатність моделей до узагальнення в різних секторах електронної комерції залишається викликом. Майбутні дослідження повинні зосередитися на розробці адаптивних моделей, здатних працювати в різних сферах без значного зниження продуктивності.

Обробка даних у реальному часі: Зі зростанням швидкості розвитку електронної комерції виникає потреба у вдосконалених аналітичних можливостях у реальному часі. Вивчення сучасних методів потокової обробки та аналізу даних буде критично важливим для надання миттєвих інсайтів і дозволить платформам реагувати на ринкові тенденції та поведінку споживачів по мірі їх виникнення.

Конфіденційність даних та етичні аспекти: Зростаюча залежність від даних вимагає збалансованого підходу до конфіденційності споживачів та етичного використання даних. Майбутні напрямки повинні включати створення рамок, що поважають приватність користувачів, одночасно використовуючи

потужності машинного навчання для персоналізованих сервісів.

Синергія з новими технологіями: Дослідження перетину машинного навчання з новими технологіями, такими як блокчейн, доповнена реальність та Інтернет речей (IoT), обіцяє вирішити давні проблеми електронної комерції та запропонувати нові підходи для покращення взаємодії з клієнтами та прозорості ланцюгів постачання.

Глобальна динаміка ринку: Глобальний характер електронної комерції вимагає моделей машинного навчання, чутливих до різноманітних культурних, регуляторних та економічних контекстів. Адаптація моделей до глобальних ринкових нюансів є важливою для майбутніх інновацій у сфері електронної комерції.

Вивчення цих обмежень та майбутніх напрямків підкреслює динамічний характер досліджень у електронній комерції. У міру того, як моделі машинного навчання стають більш досконалими, зростає потенціал революціонізувати всі аспекти електронної комерції — від взаємодії з клієнтами до операційної ефективності.

ВИСНОВКИ

Дослідження розпочалося з аналізу поведінки онлайн-споживачів, ретельно застосовуючи різні моделі машинного навчання для висвітлення складнощів взаємодій в електронній комерції. Порівняльний аналіз підтвердив ефективність ансамблевих методів, таких як Random Forest, Gradient Boosting та CatBoost, а також надав нові уявлення про їхні механізми роботи в контексті онлайн-шопінгу. Ці моделі продемонстрували виняткову здатність розкривати складні взаємозв'язки між метриками залучення користувачів, взаємодії з продуктами та, зрештою, процесами прийняття рішень, що призводять до покупок.

Поза підтвердженням прогнозної здатності моделей, це дослідження просунулося у нові напрями, виявивши критичну важливість динамічного та інтелектуального підходу до інженерії ознак. Цей процес виявився ключовим для підвищення чутливості моделей до тонких сигналів у даних, що значно покращує точність прогнозів. Крім того, дослідження підкреслило необхідність збалансованого підходу до вибору моделей, рекомендувавши стратегічне узгодження можливостей моделей із конкретними бізнес-цілями та операційними контекстами в електронній комерції.

Впроваджуючи та перевіряючи використання передових моделей машинного навчання, таких як CatBoost та ансамблеві методи, для прогнозування поведінки онлайн-споживачів, дослідження встановлює новий стандарт аналітичної строгості. Воно демонструє ключову роль складної обробки даних та стратегій оцінки моделей у видобуванні змістовних висновків із комплексних споживчих даних.

Висновки цього дослідження можуть бути корисними для підприємств електронної комерції, надаючи цінні рекомендації щодо вдосконалення їхніх операційних та стратегічних підходів. Завдяки впровадженню процесів прийняття рішень на основі даних, компанії можуть використовувати прогностичні можливості машинного навчання для підвищення залученості

клієнтів, оптимізації асортименту продуктів та підвищення операційної ефективності.

Дослідження підкреслює важливість розвитку культури, орієнтованої на дані, у організаціях електронної комерції. Використовуючи машинне навчання та аналітику як ключові компоненти бізнес-стратегії, компанії можуть залишатися попереду ринкових тенденцій, гнучко адаптуватися до потреб споживачів і підтримувати конкурентні переваги на швидкозмінному цифровому ринку. Воно також пропонує цінні аналітичні дані та практичні інструменти для фахівців з електронної комерції. Результати дослідження підкреслюють трансформаційний потенціал машинного навчання у зміні ландшафту електронної комерції, обіцяючи покращення споживчого досвіду, операційної ефективності та стратегічних інновацій. Оскільки ця галузь продовжує розвиватися, методології та аналітичні дані, представлені в цьому дослідженні, готові служити фундаментальним ресурсом для майбутніх досліджень, стимулюючи постійний розвиток застосувань машинного навчання в електронній комерції та за її межами.

Результати цього комплексного дослідження щодо застосування моделей машинного навчання для аналізу та прогнозування поведінки онлайн-покупців мають широкі наслідки для індустрії електронної комерції. Ефективно використовуючи дані в реальному часі та передову прогностичну аналітику, бізнес може відкрити нові можливості для взаємодії з клієнтами, оптимізації операцій та стратегічного планування.

Практичні наслідки цього дослідження підкреслюють значну роль машинного навчання та аналітики в режимі реального часу у формуванні майбутнього електронної комерції. Використовуючи ці технології, підприємства можуть покращити свою діяльність, отримати глибше розуміння поведінки клієнтів та адаптуватися до швидкозмінного цифрового середовища. Аналітика та методології, представлені в цьому дослідженні, пропонують цінні рекомендації щодо інтеграції передової аналітики в стратегії електронної комерції, сприяючи більш персоналізованим, ефективним та стратегічним

практикам онлайн-торгівлі. Оскільки підприємства приймають ці досягнення, вони готові перевершити очікування клієнтів у динамічному світі онлайн-комерції.

Широке дослідження впливу машинного навчання на електронну комерцію відкриває перспективний простір для подальших наукових пошуків та інновацій. Майбутні дослідження мають усувати наявні обмеження, досліджувати нові тенденції та використовувати сучасні технологічні досягнення. Основні напрями для подальших досліджень включають:

• **Інтеграція даних у реальному часі:** Майбутні дослідження повинні зосередитися на поєднанні аналітики в реальному часі з міжплатформеними даними про споживачів, щоб забезпечити більш повну картину їх поведінки в різних онлайн- та офлайн-каналах. Такий підхід може підвищити точність прогнозування та надати глибші результати щодо моделей багатоканального шопінгу.

• **Дослідження нових парадигм машинного навчання:** Із розвитком технологій машинного навчання зростає потреба у вивченні та перевірці ефективності нових підходів, таких як федеративне навчання, квантове машинне навчання у контексті електронної комерції. Ці інновації мають потенціал революціонізувати обробку даних, навчання моделей і забезпечення приватності.

• **Розвиток технологій обробки природної мови (NLP):** Зростання конверсійної комерції та контенту, створеного користувачами, підкреслює важливість NLP. Майбутні дослідження повинні розглядати передові методи для аналізу настроїв, чат-ботів та створення персоналізованого контенту, що сприятиме підвищенню залученості користувачів і полегшуватиме ухвалення рішень.

• **Сталий та етичний штучний інтелект в електронній комерції:** Збільшення використання ШІ підсилює потребу у сталих та етичних підходах. Майбутні дослідження повинні враховувати вплив масштабного використання ШІ в електронній комерції і розробляти рамки етичного ШІ, які ставлять у

пріоритет приватність споживачів та безпеку.

• **Інновації в UX/UI-дизайні:** Дослідження взаємодії між моделями машинного навчання та UX/UI-дизайном може дати значні покращення в навігації, доступності та персоналізації сайтів. Такі дослідження можуть сприяти створенню динамічних інтерфейсів на основі ШІ, що адаптуються до індивідуальних уподобань користувача у режимі реального часу.

• **Глобальні тренди електронної комерції та культурні особливості:** Зі світовим зростанням електронної комерції стає важливо глибоко розуміти регіональні та культурні особливості. Майбутні дослідження мають вивчати, як моделі машинного навчання можуть адаптуватися до різноманітних споживчих вподобань, платіжних методів і регуляторних вимог, сприяючи розвитку глобальної електронної комерції.

Підсумовуючи, шлях уперед для досліджень електронної комерції включає інновації, інтеграцію та відповідальність. Розглядаючи ці майбутні напрямки досліджень, науковці та практики можуть спільно розвивати сферу електронної комерції, забезпечуючи позитивний внесок технологічного прогресу в бізнес, споживачів та суспільство. Ключ до розкриття повного потенціалу електронної комерції полягає в глибокому розумінні поведінки споживачів та постійної еволюції машинного навчання та штучного інтелекту в цифрову епоху.

СПИСОК ЛІТЕРАТУРНИХ ДЖЕРЕЛ

[1] Хосе Рамон Саура, Ана Рейєс-Менендес, Нельсон Матос, Марісол Б. Коррейя та Педро Палос-Санчес. Поведінка споживачів у цифрову епоху. *Journal of Spatial and Organizational Dynamics*, стр. 189–199, 2020.

[2] Адела Бара, Сімона-Василіка Опреа, Крістіан Букур і Богдан-Джордж Тудоріча. Розкриття впливу локдаунів на електронну комерцію: емпіричний аналіз даних Google Analytics 2019–2022 роки. *Journal of Theoretical and Applied Electronic Commerce Research*, стр. 1481–1512, 2023.

[3] Том Албі. чи стала Google Analytics перешкодою для вебаналітики? *ACM Web Science Conference 2023*, стр. 301–312, 2023.

[4] Ґражина Сучацька, Магдалена Сколімовська-Куліг та Анета Потемпа. Класифікація сесій e-commerce-клієнтів на основі SVM. *ECMS*, 2015.

[5] Г'ю Вілсон і Умут Конуш. *Consumer Behaviour and Analytics (Mastering Business Analytics)*. Routledge, 1-е видання, 2019.

[6] Умаїр Акрам, Мелінда Тімеа Фюльоп, Адріана Тірон-Тудор, Дан Іоан Топор і Сорінел Капушнеану. Вплив цифровізації на добробут клієнтів у період пандемії: виклики та можливості для роздрібної торгівлі. *International Journal of Environmental Research and Public Health*, 2021

[7] Рікі Гунаван, Джеральді Ентоні, Марія Сьюзан Анггреайні та ін. Вплив дизайну інтерфейсу e-commerce на користувацький досвід. У *2021 6th International Conference on New Media Studies (CONMEDIA)*, стр. 90–100. 2021.

[8] Рушіл Чандра, Карун Санджая, А. Р. Аравінд, Ахмед Раді Аббас, Рузієва

Гульрух та ін. Алгоритмічна справедливість у системах машинного навчання. У *E3S Web of Conferences*, 399,. EDP Sciences, 2023

[9] Сюе Чжан, Фусен Гуо, Тао Чен, Лей Пан, Глеб Беляков і Цзянчжан Ву. Короткий огляд методів машинного та глибокого навчання для досліджень e-commerce. *Journal of Theoretical and Applied Electronic Commerce Research*, 18(4):2188–2216, 2023.

[10] Анікет Мадіха Квазі, Параг Крушанарадж Пімпалкар, Джанардан Гаджбхіє, Ріяз Кадар Саяд, Шріранг Шайлеш і Чопаде. Машинне навчання в вебдодатку e-commerce. *International Journal of Advanced Research in Science, Communication and Technology*, 2022.

[11] Александрос І. Мецай, Ірене-Марія Табакіс, Константінос Карамітсіос, Константінос Котротсіос, П. Чатзімісіос, Джордж Сталідіс і Костас Гуліанас. Застосування машинного навчання в e-commerce. У *Conference on Interactive Mobile Communication Technologies and Learning*, 2021.

[12] Джунчжі Ліу. Метод для підсилення купівельної здатності в e-commerce за допомогою технік машинного навчання. *Highlights in Business, Economics and Management*, стр. 329–336, 2023.

[13] Храг Джебамікйоус, Менглу Лі, Йога Сухас і Раша Кашеф. Використання машинного навчання та блокчейну в e-commerce та поза ним: переваги, моделі та застосування. *Discover Artificial Intelligence*, 3, 2023.

[14] Трілок Натх Пандей, Анірудх Васудев, Денніс Сагаянатан, Говінд Анджан, Даніш Аршад та Судгансу Шекхар Патра. Прогнозування задоволеності клієнтів в бразильській e-commerce: порівняльне дослідження методів машинного навчання. У 2023 International Conference on Computing, Communication, and

Intelligent Systems (ICCCIS), стр. 505–510, 2023.

[15] Мурат ГОЛЄРІ, Седат Джелік, Фатма Бозіїґіт та Деніз Килинч. Виявлення шахрайства в транзакціях e-commerce за допомогою методів машинного навчання. Artificial Intelligence Theory and Applications, 3(1):45–50, 2023.

[16] Самір Ядав, Ранджана Сінґх, Е. Маніґандан, Ману Васудеван Унні, С. Бгуванесварі та Нітін Ґірдгарвал. Дослідження факторів, що впливають на поведінку споживачів під час покупок на сайтах e-commerce у період пандемії COVID-19 на основі мережі RBF-SVM. *2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, стр. 370–380. 2023.

[17] Вей-Дун Лю та Сі-Шуй Ше. Застосування комп’ютерного зору на платформах електронної комерції та його вплив на прогнозування продажів. *Journal of Organizational and End User Computing (JOEUC),* 2024.

[18] Дженні Ян, Ендрю А. С. Солтан, Девід В. Ер та Девід А. Кліфтон. *Nature Machine Intelligence*, стр. 888–895, 2023.

[19] С. Нілакандан, В. Пракеш, М. С. ПранавКумар і Р. Баласубраманіам. Прогнозування продажів систем e-commerce за допомогою модифікованих глибинних нейронних мереж. У *2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC)*, стр. 1–5, 2023.

[20] Дікша Шравані, Праджвал Й. Р., Праджвал В. Атреяс і Шобха Г. VR Supermarket: платформа віртуальної реальності онлайн-покупок із динамічною системою рекомендацій. У *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, стр. 120–123, 2021.

[21] Бамшад Мобашер, Хонгхуа Дай та Тао Ло. Виявлення та оцінка агрегованих

профілів використання для вебперсоналізації. *Data Mining and Knowledge Discovery*, стр. 61–82, 2002.

[22] Венді В. Купівля, пошук чи перегляд: диференціація онлайн-покупців на основі навігаційних кліків. *Journal of Consumer Psychology*, 13(1–2):29–39, 2003.

[23] Ґражина Сучацька, Магдалена Сколімовська-Куліг та Анета Потемпа. Класифікація сесій e-commerce клієнтів на основі SVM. *ECMS*, 15:594–600, 2015.

[24] Германас Буднікас та ін. Комп’ютеризовані рекомендації щодо завершення електронних транзакцій за допомогою машинного навчання. *Statistics in Transition New Series*, 16(2):309–322, 2015.

[25] Окан Сакар, С. Ольджай Полат, Мете Катирджіоглу та Йомі Кастро. Прогнозування намірів покупців у реальному часі за допомогою MLP і LSTM. *Neural Computing and Applications*, 2019.

[26] Карім Бааті та Муаад Мохсіл. Прогнозування намірів покупців онлайн у реальному часі за допомогою Random Forest. У *AIAI 2020*, стр. 41–51. Springer, 2020.

[27] Лео Брейман. Random Forests. *Machine Learning*, стр. 5–32, 2001.

[28] Ієн Гудфеллоу, Йошуа Бенджіо і Аарон Курвілль. *Deep Learning*. MIT Press, 2016.

[29] Сепп Хохрайтер і Юрген Шмідхубер. LSTM. *Neural Computation*, 9(8):1735–1780, 1997.

[30] Online Shoppers Purchasing Intention Dataset URL:

https://www.kaggle.com/datasets/imakash3011/online-shoppers-purchasing-intention-dataset Дата звернення: 06.11.2025

[31] Тревор Хасті, Роберт Тібшірані та Джером Фрідман. *The Elements of Statistical Learning*. Springer, 2009.

[32] Панг-Нін Тан, Майкл Стейнбах і Віпін Кумар. *Introduction to Data Mining*. Addison-Wesley, 2006.

[33] Томас Дітеріч. Метод ансамблів у машинному навчанні. *Multiple Classifier Systems*, стр. 1–20, 2000.

[34] Нелло Крістіаніні та Джон Шоу-Тейлор. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[35] Гуолін Ке, Ці Менг, Томас Фінлі та ін. LightGBM: ефективні дерева бустингу. *NIPS*, 2017.

[36] Ієн Гудфеллоу, Йошуа Бенджіо та Аарон Курвілль. *Deep Learning*. MIT Press, 2016.

[37] Саймон Гейкін. *Neural Networks and Learning Machines*. Pearson, 3-є видання, 2009.

Додатки

Презинтація

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
Навчально-науковий Інститут інформаційних технологій
Кафедра інформаційних систем та технологій

Кваліфікаційна робота на тему:

АНАЛІТИЧНА МОДЕЛЬ СПОЖИВЧОЇ ПОВЕДІНКИ НА РИНКУ E-COMMERCE

Виконав – студент САДМ-61 Майборода Антон Олегович
Керівник – к.ф.-м.н, доцент Нафєєв Ровіл Касимович

Актуальність

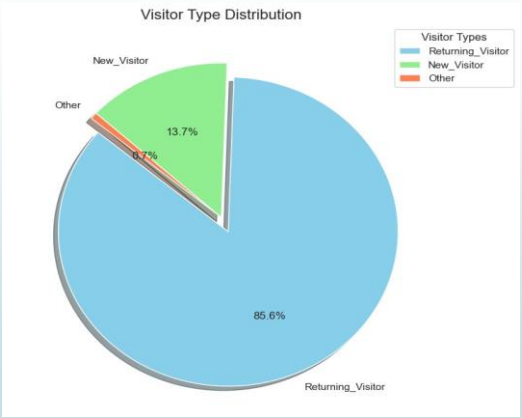
- Онлайн-шопінг і електронна комерція сьогодні є невід'ємною складовою глобальної економіки та повсякденного життя споживачів.
- Розуміння закономірностей та тенденцій поведінки споживачів в e-commerce дає змогу бізнесу ефективніше залучати аудиторію, підвищувати рівень довіри, оптимізувати коефіцієнт конверсії та, як результат, збільшувати прибутковість онлайн-платформ.

Структура та вміст датасету

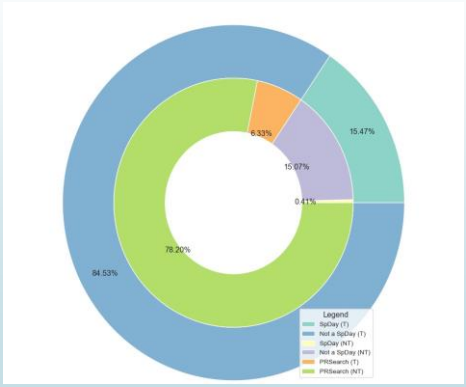
Числові атрибути	
АТРИБУТ	ОПИС
Administrative	Рахує кількість адміністративних сторінок, відвіданих під час сесії
Administrative Duration	Відображає час, проведений на адміністративних сторінках
Informational	Рахує кількість інформаційних сторінок, відвіданих користувачем
Informational Duration	Сукупний час, проведений на інформаційних сторінках
ProductRelated	Підсумовує перегляди сторінок, пов'язаних із продуктами
ProductRelated Duration	Загальний час, проведений на сторінках продуктів
BounceRates	Відсоток сесій, у яких було переглянуто лише одну сторінку
ExitRates	Частота виходів користувачів зі сторінки
PageValues	Середня цінність сторінок, відвіданих перед транзакцією
SpecialDay	Показує наближеність до особливих подій

Категоріальні атрибути	
АТРИБУТ	ОПИС
Month	Місяць сесії, критично важливий для аналізу сезонних тенденцій
OperatingSystems	Ідентифікує операційну систему, яку використовує користувач
Browser	Браузер, що використовувався для сесії, впливає на рендеринг сайту
Region	Розташування користувача для аналізу сегментації ринку
TrafficType	Джерело трафіку, що веде на вебсайт
VisitorType	Розрізняє нових і повернених відвідувачів
Weekend	Булевий сигнал, якщо візит відбувся у вихідні
Revenue	Булевий індикатор того, чи призвела сесія до транзакції

Аналіз типів користувачів



Інфографіка розподілу транзакцій за особливих умов



Моделі та функції аналізу

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.642	0.256	0.856	0.394
KNN	0.755	0.313	0.671	0.427
Decision Tree	0.845	0.449	0.617	0.520
Random Forest	0.906	0.636	0.716	0.673
Gradient Boosting	0.900	0.616	0.698	0.654
AdaBoost	0.882	0.545	0.778	0.641
Extra Trees	0.896	0.604	0.686	0.642
SVM-Linear	0.879	0.537	0.784	0.637
SVM-RBF	0.875	0.532	0.644	0.583
SVM-Poly	0.781	0.356	0.754	0.484
XGBoost	0.904	0.641	0.662	0.651
LightGBM	0.906	0.644	0.689	0.666
CatBoost	0.913	0.681	0.671	0.676
Neural Network	0.862	0.495	0.725	0.588
Multilayer Perceptron	0.851	0.462	0.581	0.515
Stacked MLP Model	0.910	0.703	0.581	0.636

Валідація та оцінка отриманих результатів

$$\text{Акуратність} = \frac{\text{Істинні позитивні + Істинні негативні}}{\text{Загальна кількість спостережень}}$$

$$\text{Точність} = \frac{\text{Істинно позитивні результати}}{\text{Істинно позитивні результати + Хибнопозитивні результати}}$$

$$\text{Повнота} = \frac{\text{Істинно позитивні результати}}{\text{Істинно позитивні результати + Хибнопозитивні результати}}$$

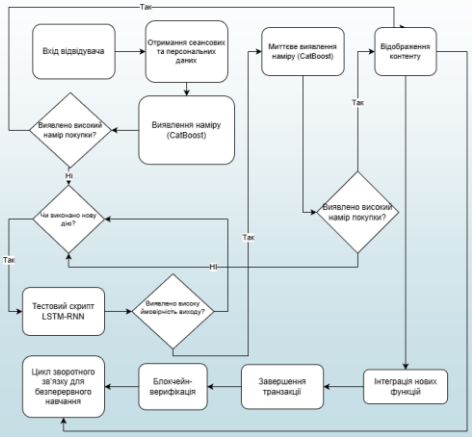
$$\text{Оцінка } F1 = 2 \times \frac{\text{Точність} \times \text{Повність}}{\text{Точність + Повність}}$$

Порівняння точності та узагальнення результатів

Порівняння точності для класифікаційних моделей

Акуратність і ROC-AUC (похибки)

Розширена предикативна система



Висновки

- CatBoost показав себе найкраще, що зумовлено його алгоритмічними перевагами в роботі з категоріальними даними та великими просторами ознак. Він став одним з провідних моделей у дослідженні
- Спираючись на аналітичні можливості моделі CatBoost пропонується вдосконалена система для прогнозування намірів онлайн-покупців у режимі реального часу.
- Запропонована система оптимально поєднує аналітичну точність класифікатора CatBoost із здатністю LSTM-RNN розпізнавати часові шаблони, утворюючи високорозвинену прогнозуючу архітектуру