

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

Навчально-науковий інститут «Інститут інформаційних технологій в економіці»

Кафедра системного аналізу та кібербезпеки

ОСВІТНЬО-ПРОФЕСІЙНА ПРОГРАМА

«Системний аналіз»

СПЕЦІАЛЬНІСТЬ

124 «Системний аналіз»

Форма навчання: денна

КВАЛІФІКАЦІЙНА МАГІСТЕРСЬКА РОБОТА

На тему: «Математичне моделювання алгоритмів прогнозування довічної цінності користувачів у digital-кампаніях на основі багатовимірних даних»

Здобувача: Бородатюка Дмитра Михайловича

(підпис)

Науковий керівник: д.ф.-м.н., проф. Колєчкіна Л. М.

(підпис)

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

Завідувач кафедри: д.ф.-м.н., проф. Джалладова І.А.

(підпис)

Київ 2026

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

Навчально-науковий інститут «Інститут інформаційних технологій в економіці»

Кафедра системного аналізу та кібербезпеки

ОСВІТНЬО-ПРОФЕСІЙНА ПРОГРАМА

«Системний аналіз»

СПЕЦІАЛЬНІСТЬ

124 «Системний аналіз»

ПОГОДЖЕНО

Керівник проєктної групи (гарант)
освітньо-професійної програми

ЗАТВЕРДЖУЮ

Завідувач кафедри системного аналізу та
кібербезпеки

(підпис)

2026 р.

(підпис)

Джалладова І.А.

2026 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

здобувачу вищої освіти Бородатюку Дмитру Михайловичу

денної форми навчання

на підготовку кваліфікаційної магістерської роботи

На тему: «Математичне моделювання алгоритмів прогнозування довічної цінності користувачів у digital-кампаніях на основі багатовимірних даних»

Тему затверджено наказом ректора Університету від «__» _____ 2026р. №__

Кваліфікаційна магістерська робота виконується на матеріалах

План кваліфікаційної магістерської роботи

Розділ 1	Теоретико-методичні засади прогнозування довічної цінності користувачів у сучасному digital-маркетингу.
Розділ 2	Аналітико-методологічні основи прогнозування довічної цінності користувачів в digital-кампаніях.
Розділ 3	Програмна реалізація та експериментальне дослідження ефективності предиктивних моделей.

Об'єкт дослідження:	Об'єктом дослідження є процес формування, зміни та предиктивної аналітики довічної цінності користувачів у цифрових маркетингових кампаніях на основі багатовимірних даних.
Предмет дослідження:	Предметом дослідження є математичні та комп'ютерні моделі, алгоритми і методи прогнозування довічної цінності користувачів цифрових платформ із застосуванням технологій машинного навчання..
Мета кваліфікаційної магістерської роботи:	Метою кваліфікаційної магістерської роботи є розробка та математичне моделювання алгоритмів прогнозування довічної цінності користувачів на основі багатовимірних поведінкових даних для підвищення ефективності маркетингових стратегій у digital-кампаніях.

Конкретні завдання, які здобувач повинен виконати для досягнення поставленої мети:

У розділі 1

1. Проаналізувати економічну сутність та стратегічне значення довічної цінності користувача для оптимізації рекламних бюджетів у цифрових кампаніях.
2. Здійснити ґрунтовний огляд класичних, імовірнісних та сучасних підходів до математичного моделювання показників.
3. Дослідити роль багатовимірних поведінкових масивів даних та інструментарію машинного навчання в задачах предиктивного моделювання та аналітики дій користувачів.

У розділі 2

1. Формалізувати та здійснити математичну постановку задачі прогнозування з урахуванням проблеми екстремальної асиметрії розподілу та надмірної кількості нульових значень.
2. Провести системний аналіз багатовимірного простору ознак (транзакційних, поведінкових, контекстуальних) та дослідити методи відбору найбільш інформативних предикторів.

-
3. Обґрунтувати математичні архітектури моделей машинного навчання, а також розробити методологію їх валідації в умовах часових рядів

У розділі 3

1. Описати архітектуру предиктивного конвеєра даних та сформувати репрезентативний багатовимірний датасет на основі реальних деперсоніфікованих логів маркетингової кампанії.
-
2. Здійснити практичну програмну реалізацію розроблених предиктивних алгоритмів та бар'єрної моделі мовою Python на базі бібліотеки NumPy.
-
3. Провести комплексний порівняльний аналіз моделей, виконати стрес-тестування алгоритмів на стійкість до інформаційного шуму і сформулювати практичні рекомендації щодо їх впровадження в CRM-системи

**Завдання підготував
науковий керівник**

(підпис)

Колєчкіна Л. М.

(ініціали, прізвище)

«__»_____2026 р.

**Завдання одержав
здобувач**

(підпис)

Бородатюк Д. М.

(ініціали, прізвище)

«__»_____2026 р.

Реферат

Кваліфікаційна магістерська робота містить: 66 сторінок, 2 таблиці, 9 рисунків, список літератури з 39 найменувань.

Об'єкт дослідження – процес формування, зміни та предиктивної аналітики довічної цінності користувачів (Lifetime Value) у цифрових маркетингових кампаніях на основі багатовимірних даних.

Предмет дослідження – математичні та комп'ютерні моделі, алгоритми і методи прогнозування довічної цінності користувачів цифрових платформ із застосуванням технологій машинного навчання.

Мета кваліфікаційної магістерської роботи – розробка та математичне моделювання алгоритмів прогнозування довічної цінності користувачів (LTV) на основі багатовимірних поведінкових даних для підвищення ефективності маркетингових стратегій у digital-кампаніях.

Відповідно до поставленої мети були поставлені та вирішені такі завдання:

- проаналізувати економічну сутність LTV, здійснити огляд методів його моделювання та дослідити проблему екстремальної асиметрії даних (Zero-Inflation);
- формалізувати математичну модель прогнозування як задачу регресії та провести системний аналіз простору ознак із відбором найбільш інформативних предикторів;
- розробити та програмно реалізувати архітектури машинного навчання: ансамблеві методи (Random Forest, XGBoost) та кастомну Двохетапну бар'єрну модель (Two-Stage Hurdle Model);
- провести порівняльний аналіз моделей з використанням методології Out-of-Time Validation;
- оцінити стійкість розроблених алгоритмів до інформаційного шуму та динамічних змін у даних шляхом стрес-тестування.

Теоретична значущість кваліфікаційної магістерської роботи полягає у систематизації математичних методів прогнозування LTV та обґрунтуванні переваг використання композитних моделей для роботи з асиметричними

розподілами даних. У роботі поглиблено розуміння механізмів впливу поведінкових ознак на довгострокову цінність клієнта через призму методів просторового відбору (Feature Importance).

Методична значущість полягає у розробці цілісного конвеєра обробки даних (Data Pipeline): від препроцесингу до фінальної валідації прогнозів у часових розрізах. Детально описано архітектуру об'єктно-орієнтованої побудови алгоритмів машинного навчання мовою Python «з нуля», що забезпечує глибокий контроль алгоритмічної логіки на кожному етапі обчислень.

Практичне значення отриманих результатів визначається можливістю безпосередньої інтеграції розробленого предиктивного контуру у високонавантажені CRM-системи для динамічного перерозподілу маркетингових бюджетів. На основі реальних даних доведено, що розроблена Hurdle Model забезпечує найкращу точність прогнозування ($MAE = 31.62$ у.о.) на чистих даних, тоді як базова модель Random Forest демонструє найвищу стійкість в умовах інформаційного шуму. Впровадження системи дозволяє ідентифікувати окупність когорти та вимикати збиткову рекламу вже на 7-й день її життя.

Рік виконання кваліфікаційної магістерської роботи: 2026

Рік захисту роботи: 2026

Ключові слова: прогнозування LTV, машинне навчання, багатовимірні дані, Random Forest, XGBoost, Hurdle Model, предиктивна аналітика, digital-маркетинг, асиметрія даних, Zero-Inflation.

Відгук
на кваліфікаційну магістерську роботу
навчально-наукового інституту інформаційних технологій в економіці
освітньо-професійної програми «Системний аналіз»

Бородатюка Дмитра Михайловича

(прізвище, ініціали)

на тему: Математичне моделювання алгоритмів прогнозування довічної цінності користувачів у digital-кампаніях на основі багатовимірних даних

(назва теми)

1. Актуальність теми: визначається зростаючою роллю предиктивної аналітики в управлінні відносинами з клієнтами та необхідністю оптимізації маркетингових бюджетів. В умовах високої конкуренції та постійного збільшення вартості залучення нових користувачів (CAC), точне прогнозування довічної цінності клієнта (LTV) на основі багатовимірних даних із застосуванням алгоритмів машинного навчання стає критично важливим для забезпечення рентабельності та життєздатності сучасних digital-кампаній.
2. Позитивні риси кваліфікаційної магістерської роботи: автором всебічно досліджено економічну сутність показника LTV та математичні засади його прогнозування. Глибоко проаналізовано специфіку багатовимірного простору ознак та методи вирішення проблеми екстремальної асиметрії даних (Zero-Inflation), характерної для цифрового маркетингу. Проведено чіткий та структурований порівняльний аналіз трьох предиктивних архітектур: базової (Random Forest), просунутої (XGBoost) та кастомної системної двохетапної бар'єрної моделі (Two-Stage Hurdle Model) за метриками MAE, RMSE та R^2 .
3. Наявність самостійних розробок автора: у Розділі 3 та Додатку А представлено унікальну практичну реалізацію алгоритмів машинного навчання «з нуля» мовою Python із використанням об'єктно-орієнтованого підходу, без залучення високорівневих фреймворків. Автором самостійно розроблено архітектурний контур системи, виконано попередню обробку даних, а також проведено стрес-тестування моделей на стійкість до дрейфу даних шляхом симуляції інформаційного шуму. Здобувач продемонстрував високий рівень інженерної та аналітичної роботи.
4. Цінність теоретичних висновків та практичних рекомендацій: результати дослідження мають безпосереднє практичне значення для бізнесу в сфері e-commerce та мобільних застосунків. Розроблена композитна бар'єрна модель ефективно вирішує проблему неактивного («нульового») трафіку та може бути інтегрована в сучасні CRM-системи для динамічного перерозподілу рекламних бюджетів у режимі реального часу. Висновки щодо стійкості моделей дають

бізнесу чіткі рекомендації щодо вибору алгоритмів залежно від якості наявної аналітичної інфраструктури.

5. Наявність недоліків: робота виконана на високому науковому та практичному рівні і позбавлена суттєвих недоліків, здатних вплинути на якість її виконання або достовірність отриманих результатів.

6. Загальна оцінка кваліфікаційної магістерської роботи та її допущення до захисту перед ЕК: кваліфікаційна магістерська робота Бородатюка Дмитра Михайловича робота виконана на достатньому рівні, має редакційні помилки і опечатки, але може бути допущена до захисту.

Науковий керівник д.ф.-м.н., професор, доктор фізико-математичних наук

_____	Колєчкіна Л. М..
(підпис)	(прізвище, ініціали)

“ ” _____ 2025 р.

РЕЦЕНЗІЯ

на кваліфікаційну магістерську роботу здобувача вищої освіти
Бородатюка Дмитра Михайловича

Тема «Математичне моделювання алгоритмів прогнозування довічної цінності користувачів у digital-кампаніях на основі багатовимірних даних / Mathematical modeling of customer lifetime value prediction algorithms in digital campaigns based on multidimensional data»

Актуальність теми кваліфікаційної магістерської роботи і доцільність її розроблення

Актуальність роботи визначається стрімким розвитком алгоритмічної економіки та зростаючим значенням предиктивної аналітики в задачах цифрового маркетингу. В умовах постійного підвищення вартості залучення нових клієнтів (CAC), традиційні підходи до усереднення фінансових показників (ARPU) втрачають ефективність. Тема дослідження є особливо актуальною в контексті системного аналізу багатовимірних поведінкових даних та розробки інтелектуальних систем автоматизованого прогнозування довічної цінності користувача (LTV), що дозволяє компаніям проактивно керувати рекламними бюджетами в режимі реального часу.

Якість проведеного дослідження

Робота вирізняється чіткою постановкою мети, послідовною логічною структурою та високим рівнем опрацювання сучасних наукових джерел. Здобувач продемонстрував глибокі знання методів системного аналізу, математичної статистики та машинного навчання, а також вміння застосовувати їх до реальних бізнес-задач. У дослідженні проведено комплексне порівняння ефективності трьох обчислювальних стратегій: базової моделі (Random Forest), просунутого градієнтного бустингу (XGBoost) та кастомної системної двохетапної бар'єрної моделі (Two-Stage Hurdle Model). Здійснено повний цикл реалізації системи — від інтеграції деперсоніфікованих комерційних логів та Feature Engineering до Out-of-Time валідації та стрес-тестування моделей на стійкість до дрейфу даних.

Позитивні риси кваліфікаційної магістерської роботи

До ключових сильних сторін роботи варто віднести строге теоретико-методологічне обґрунтування простору ознак та інноваційний підхід до вирішення проблеми надмірної кількості нульових значень (Zero-Inflation) у вибірці. Особливої уваги заслуговує самостійна програмна реалізація складних матричних алгоритмів та кастомного ядра моделей «з нуля» мовою Python (from scratch) з використанням об'єктно-орієнтованого підходу. Грамотний аналіз Feature Importance дозволив автору виявити приховані нелінійні закономірності формування LTV, а експериментальна оцінка деградації метрик в умовах інформаційного шуму додає роботі вагомому науково-практичного значення. Здобувач успішно поєднав апарат системного аналізу з навичками прикладного програмування у середовищі Jupyter Notebook (Google Colab) із залученням NumPy та Pandas.

Зауваження

У процесі виконання дослідження здобувачу доцільно було б ширше висвітлити питання хмарної інфраструктури розгортання предиктивних моделей та інтеграції конвеєра даних (Data Pipeline) із потоковими шинами повідомлень. Разом із тим, робота повністю позбавлена суттєвих недоліків, які могли б вплинути на її загальну високу якість, цілісність, наукову глибину та достовірність отриманих результатів.

Практична значимість висновків і рекомендацій

Результати дослідження мають безпосереднє прикладне значення для підприємств у сфері e-commerce, мобільних застосунків та SaaS-сервісів. Запропонована двохетапна бар'єрна архітектура дозволяє з високою точністю прогнозувати фінансовий потенціал користувачів уже в перший тиждень їхньої активності. Зведені правила проектування інфраструктури та рекомендації щодо використання мета-маршрутизації (Dynamic Ensemble) можуть бути безпосередньо впроваджені в діючі CRM-системи для мінімізації фінансових ризиків та підвищення рентабельності інвестицій у маркетинг (ROI/ROAS).

Місце роботи та посада рецензента:

Товариство з обмеженою відповідальністю “КОМ'ЮНІТІ ФІЛЬМ ПРОДАКШН”,
гіг-спеціаліст, Аналітик продукту

Рецензент: _____ /Гордієнко Олена Володимирівна /
(підпис) (прізвище, ім'я, по батькові)

Підпис засвідчую: _____ Директор
(посада, підпис)

Місце печатки організації, де працює рецензент

ЗМІСТ

ВСТУП.....	3
РОЗДІЛ 1	
ТЕОРЕТИКО-МЕТОДИЧНІ ЗАСАДИ ПРОГНОЗУВАННЯ LTV У	
СУЧАСНОМУ DIGITAL-MАРКЕТИНГУ.....	6
1.1 Економічна сутність та значення довічної цінності користувача (LTV) у цифрових кампаніях.....	6
1.2. Огляд класичних та сучасних підходів до математичного моделювання LTV.....	11
1.3. Роль багатовимірних даних та машинного навчання в задачах предиктивної аналітики поведінки користувачів.....	16
РОЗДІЛ 2	
АНАЛІТИКО-МЕТОДОЛОГІЧНІ ОСНОВИ ПРОГНОЗУВАННЯ LTV В	
DIGITAL-KАМПАНІЯХ.....	23
2.1. Формалізація та математична постановка задачі прогнозування довічної цінності користувача.....	23
2.2. Системний аналіз багатовимірному простору ознак та методи відбору предикторів.....	27
2.3. Математичні моделі машинного навчання для предиктивної аналітики LTV.....	32
2.4. Методологія валідації моделей та специфіка оцінювання в умовах часових рядів.....	34
РОЗДІЛ 3	
ПРОГРАМНА РЕАЛІЗАЦІЯ ТА ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ	
ЕФЕКТИВНОСТІ ПРЕДИКТИВНИХ МОДЕЛЕЙ.....	39
3.1. Архітектура системи прогнозування довічної цінності користувачів..	39
3.2. Опис експериментального середовища та формування багатовимірному датасету digital-кампаній.....	41
3.3. Реалізація алгоритмів машинного навчання для моделювання LTV мовою Python.....	44
3.4. Вибір та обґрунтування метрик ефективності предиктивних моделей	48
3.5. Аналіз результатів моделювання та порівняння з базовими (історичними) методами розрахунку LTV.....	51
3.6. Оцінка стійкості моделей до динамічних змін у поведінкових даних користувачів.....	55
ВИСНОВКИ.....	61
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	63
ДОДАТОК А.....	67

ВСТУП

У сучасній цифровій економіці дані стали ключовим активом, а здатність бізнесу ефективно взаємодіяти з клієнтом визначає його життєздатність. В умовах глобальної цифровізації та стрімкого розвитку рекламних технологій вартість залучення нового користувача (CAC) постійно зростає, що змушує компанії переходити від стратегії масового охоплення до індивідуалізованого управління клієнтським досвідом. Центральним показником у цій парадигмі є довічна цінність користувача (Customer Lifetime Value – LTV), яка відображає прогнозований чистий прибуток від стосунків із клієнтом протягом усього періоду взаємодії [24].

Прогнозування LTV є складним математичним завданням через високу асиметрію даних, наявність великої кількості «нульових» значень (користувачів, що не здійснили покупку) та динамічну зміну поведінкових патернів. Традиційні імовірнісні моделі (як-от BG/NBD) часто не враховують багатовимірність ознак, які надають сучасні системи аналітики, Google Analytics тощо). Це зумовлює потребу у розробці складних архітектур машинного навчання, здатних обробляти великі масиви даних та забезпечувати високу точність прогнозів у реальному часі.

Актуальність роботи полягає у відсутності універсальних і робастних систем прогнозування LTV, які б ефективно працювали в умовах інформаційного шуму та різної активності клієнтських сегментів. Наявні рішення часто або занадто спрощені, або не мають механізмів адаптації до «брудних» даних. Це дослідження пропонує розробку та математичне обґрунтування дворівневих моделей (Hurdle Models) та ансамблів (Random Forest, XGBoost), що дозволяють відокремити ймовірність конверсії від прогнозу обсягу витрат, забезпечуючи бізнесу гнучкий інструмент для оптимізації маркетингових бюджетів.

Метою дослідження є розробка та математичне моделювання алгоритмів прогнозування LTV на основі багатовимірних поведінкових даних для підвищення ефективності маркетингових стратегій у digital-кампаніях.

Для досягнення цієї мети необхідно вирішити такі завдання:

- Провести аналіз економічної сутності LTV та класифікувати існуючі математичні підходи до його оцінки.
- Дослідити методи попередньої обробки багатовимірних даних та методи відбору найбільш інформативних ознак (Feature Selection).
- Розробити та реалізувати архітектури машинного навчання: ансамблеві методи (Random Forest, XGBoost) та спеціалізовану дворівневу бар'єрну модель (Two-Stage Hurdle Model).
- Провести порівняльний аналіз моделей за метриками точності (MAE, RMSE, R^2) та оцінити їхню здатність до генералізації.
- Здійснити стрес-тестування моделей на стійкість до дрейфу даних та інформаційного шуму для визначення меж їхнього практичного застосування.

Теоретична значущість дослідження полягає у систематизації математичних методів прогнозування LTV та обґрунтуванні переваг використання композитних моделей для роботи з асиметричними розподілами даних. У роботі поглиблено розуміння механізмів впливу поведінкових ознак на довгострокову цінність клієнта через призму методів Feature Importance.

Методична значущість полягає у розробці цілісного конвеєра (Data Pipeline) обробки даних: від препроцесингу та подолання мультиколінеарності до фінальної валідації прогнозів. Запропонований підхід до побудови Hurdle-моделі «з нуля» (manual implementation) дозволяє детально відстежувати логіку прийняття рішень на кожному етапі моделювання [14].

Практична значущість роботи полягає у створенні готового програмного інструментарію для предиктивної аналітики, який може бути інтегрований у CRM-системи та рекламні кабінети. Використання розроблених моделей дозволяє компаніям здійснювати прецизійний таргетинг, ідентифікувати VIP-клієнтів на ранніх етапах та мінімізувати фінансові ризики при розподілі рекламних бюджетів.

Інформаційну базу дослідження становлять деперсоніфіковані логи активності користувачів, що містять транзакційні та поведінкові ознаки, наукові публікації у галузі Data Science та маркетингу, документація до бібліотек

Scikit-learn, XGBoost, Pandas, а також результати експериментів, проведених у середовищі розробки Python.

Структура роботи складається з чотирьох розділів:

- Перший розділ присвячено аналізу теоретичних засад LTV, його ролі в digital-маркетингу та огляду класичних математичних моделей.
- Другий розділ описує методичний апарат: від підготовки багатовимірних даних до математичного опису архітектур Random Forest, XGBoost та Hurdle Model.
- Третій розділ містить опис експериментальної частини, програмної реалізації алгоритмів та порівняльний аналіз результатів за допомогою статистичних метрик.
- Четвертий розділ фокусується на тестуванні моделей в умовах невизначеності (стрес-тести), аналізі стабільності та формулюванні практичних рекомендацій щодо впровадження.

РОЗДІЛ 1

ТЕОРЕТИКО-МЕТОДИЧНІ ЗАСАДИ ПРОГНОЗУВАННЯ LTV У СУЧАСНОМУ DIGITAL-MАРКЕТИНГУ

1.1 Економічна сутність та значення довічної цінності користувача (LTV) у цифрових кампаніях

В умовах глобальної цифровізації економіки та стрімкого розвитку інформаційних технологій відбувається фундаментальна трансформація підходів до ведення бізнесу, зокрема у сфері взаємодії підприємств зі споживачами. Перехід від традиційних моделей масового маркетингу до персоналізованих комунікацій у цифровому середовищі зумовив необхідність перегляду ключових показників ефективності діяльності компаній. Сучасний digital-маркетинг характеризується не лише високим рівнем конкуренції, але й значним зростанням вартості залучення нових користувачів. З огляду на це, парадигма управління маркетинговими бюджетами зміщується від короткострокової транзакційної вигоди до побудови довгострокових, економічно обґрунтованих відносин із клієнтом [21]. У цьому контексті центральним поняттям стає довічна цінність користувача (Customer Lifetime Value - LTV або CLV).

Економічна сутність довічної цінності користувача полягає у визначенні поточної дисконтованої вартості всіх майбутніх чистих грошових потоків, які компанія очікує отримати від взаємодії з конкретним клієнтом протягом усього періоду його співпраці з бізнесом. Тобто, з економічної точки зору, LTV розглядає клієнта не просто як джерело одноразового доходу, а як довгостроковий інвестиційний актив компанії. Формування цього показника базується на розумінні того, що утримання існуючого клієнта, як правило, вимагає значно менших фінансових витрат порівняно із залученням нового, а лояльні споживачі схильні генерувати більший обсяг прибутку за рахунок повторних покупок, крос-продажів (cross-sell) та додаткових продажів (upsell).

З теоретичної точки зору, концепція LTV спирається на теорію маркетингу взаємовідносин (relationship marketing), яка постулює, що фінансовий успіх

підприємства є прямо пропорційним тривалості та якості його відносин зі споживачами. У цифровому середовищі ця концепція набуває нових масштабів завдяки можливості детального відстеження цифрового сліду користувача (digital footprint). Багатовимірні дані, що генеруються в результаті взаємодії користувача з веб-ресурсами, мобільними додатками, рекламними банерами та електронними розсилками, створюють унікальне інформаційне поле. Це поле дозволяє не лише ретроспективно оцінювати прибутковість клієнта, але й будувати складні математичні моделі для прогнозування його майбутньої поведінки.

Значення показника LTV у цифрових кампаніях неможливо переоцінити, оскільки він виконує одразу кілька критично важливих функцій в системі управління маркетингом. По-перше, LTV виступає базовим критерієм для визначення економічно доцільної вартості залучення клієнта (Customer Acquisition Cost - CAC). В економічній теорії та практиці цифрового бізнесу загальноприйнятим є правило: бізнес-модель вважається життєздатною та масштабованою лише за умови, що LTV значно перевищує CAC. Оптимальним співвідношенням LTV до CAC зазвичай вважається пропорція 3:1. Розуміння цього балансу дозволяє маркетологам встановлювати ліміти ставок в аукціонах контекстної чи таргетованої реклами, уникаючи ситуацій, коли витрати на залучення користувача перевищують прибуток, який він здатен генерувати за весь свій «життєвий цикл» у продукті.

По-друге, прогнозування LTV відіграє ключову роль у процесах сегментації клієнтської бази. У цифрових кампаніях аудиторія є вкрай неоднорідною: частина користувачів робить одну дрібну покупку і зникає, інша частина стає амбасадорами бренду і здійснює регулярні транзакції. Оперуючи прогнозованими значеннями довічної цінності, компанія отримує можливість розподіляти свою аудиторію на кластери (наприклад, високомаржинальні, середні та низькомаржинальні клієнти). Такий підхід дає змогу оптимізувати розподіл ресурсів: спрямовувати основну частину маркетингового бюджету, персоналізовані пропозиції та ресурси клієнтської підтримки на утримання

найбільш цінних сегментів, замість рівномірного і неефективного розпорошення коштів на всю масу користувачів.

По-третє, показник LTV є необхідним інструментом для оцінювання стратегічної ефективності самих цифрових кампаній. Традиційні метрики, такі як показник клікабельності (CTR), ціна за клік (CPC) або навіть рентабельність інвестицій у рекламу за першою транзакцією (ROAS), часто дають викривлену картину. Кампанія може демонструвати високу вартість залучення ліда, що на перший погляд здається економічно не вигідним. Проте, якщо когортний аналіз та математичне моделювання показують, що залучені через цей канал користувачі мають аномально високий рівень утримання (Retention Rate) і високий LTV, така кампанія в довгостроковій перспективі виявляється найбільш рентабельною [17]. Таким чином, перехід до управління на основі LTV нівелює проблему «короткозорості» в цифровому маркетингу.

Слід також зазначити, що в умовах сучасного інформаційного суспільства та мінливості споживчих уподобань життєвий цикл клієнта може несподівано перериватися. Відтік клієнтів (Churn) стає однією з головних загроз для накопиченого потенціалу довічної цінності. Тому економічне значення LTV також виражається у здатності системи предиктивної аналітики завчасно ідентифікувати користувачів, які демонструють ознаки зниження залученості, та вживати проактивних заходів для їх реактивації.

Підсумовуючи вищевикладене, можна стверджувати, що економічна сутність довічної цінності користувача зводиться до трансформації інформації про його поведінку у фінансовий еквівалент довгострокового прибутку. Значення LTV у сучасному цифровому маркетингу полягає у зміщенні фокусу з тактичної оптимізації вартості кліка чи конверсії на стратегічне управління капіталом клієнтських відносин. Це створює міцний фундамент для побудови стійких бізнес-моделей, де інвестиції у маркетинг розглядаються не як безповоротні витрати, а як вкладення у формування портфеля високорентабельних цифрових активів - лояльних користувачів. Здатність компанії точно моделювати та

прогнозувати цей показник за допомогою багатовимірних даних стає ключовою конкурентною перевагою в епоху алгоритмічної економіки.

Варто також наголосити, що розуміння економічної природи LTV нерозривно пов'язане з еволюцією методів його обчислення, які пройшли шлях від простої історичної агрегації до складної предиктивної аналітики. Традиційно компанії оперували ретроспективним (історичним) LTV, який розраховувався як фактична сума прибутку, отриманого від клієнта за минулий період. Хоча такий підхід є простим у реалізації, він має суттєвий економічний недолік: історичний метод констатує вже dokonаний факт і не враховує зміни у споживчій поведінці, вплив нових маркетингових стимулів чи макроекономічні фактори. Натомість сучасна парадигма digital-маркетингу вимагає переходу до прогностичного (предиктивного) LTV. Прогностичний підхід розглядає майбутню цінність клієнта як імовірнісну величину, що динамічно змінюється під впливом безлічі факторів. Саме це зумовлює необхідність застосування методів системного аналізу та математичного моделювання для завчасного визначення потенціалу кожного залученого користувача.

Формування показника довічної цінності у сучасному цифровому просторі відбувається в умовах омніканальності (omnichannel). Сучасний споживач рідко приймає рішення про цільову дію (покупку, підписку) після першого контакту з брендом. Його шлях до конверсії (Customer Journey) пролягає через безліч точок дотику: перегляд таргетованої реклами у соціальних мережах, клік по банеру в пошуковій системі, перехід з email-розсилки, взаємодія з мобільним додатком тощо. Кожна така взаємодія залишає масив атрибутивних даних: час на сайті, глибина перегляду сторінок, реакція на push-сповіщення, частота додавання товарів до кошика без подальшої оплати. Цей масив формує багатовимірний простір поведінкових характеристик (багатовимірні дані), який стає сировиною для прогнозування LTV. Економічна цінність користувача в цьому розрізі залежить не лише від його платоспроможності, але й від ступеня його цифрової залученості (engagement).

Крім того, інтеграція LTV у систему управління цифровими кампаніями є невід'ємною складовою юніт-економіки (Unit Economics) продукту. Юніт-економіка дозволяє оцінити прибутковість бізнес-моделі на рівні базової одиниці, якою в digital-сфері найчастіше виступає один користувач. У ситуаціях, коли компанія працює за моделлю підписки (SaaS) або з орієнтацією на високу частоту повторних покупок (e-commerce, мобільні ігри), перша транзакція користувача може навіть не покривати витрат на його залучення (CAC). Тобто, на початковому етапі бізнес свідомо несе збитки, інвестуючи в клієнтську базу (стратегія loss leader). І лише завдяки накопиченому LTV через утримання (retention) ці інвестиції окупаються в довгостроковій перспективі. Математичне прогнозування моменту перетину графіків накопиченого доходу з витратами на залучення (Payback Period) є критично важливим для забезпечення фінансової стабільності та уникнення касових розривів.

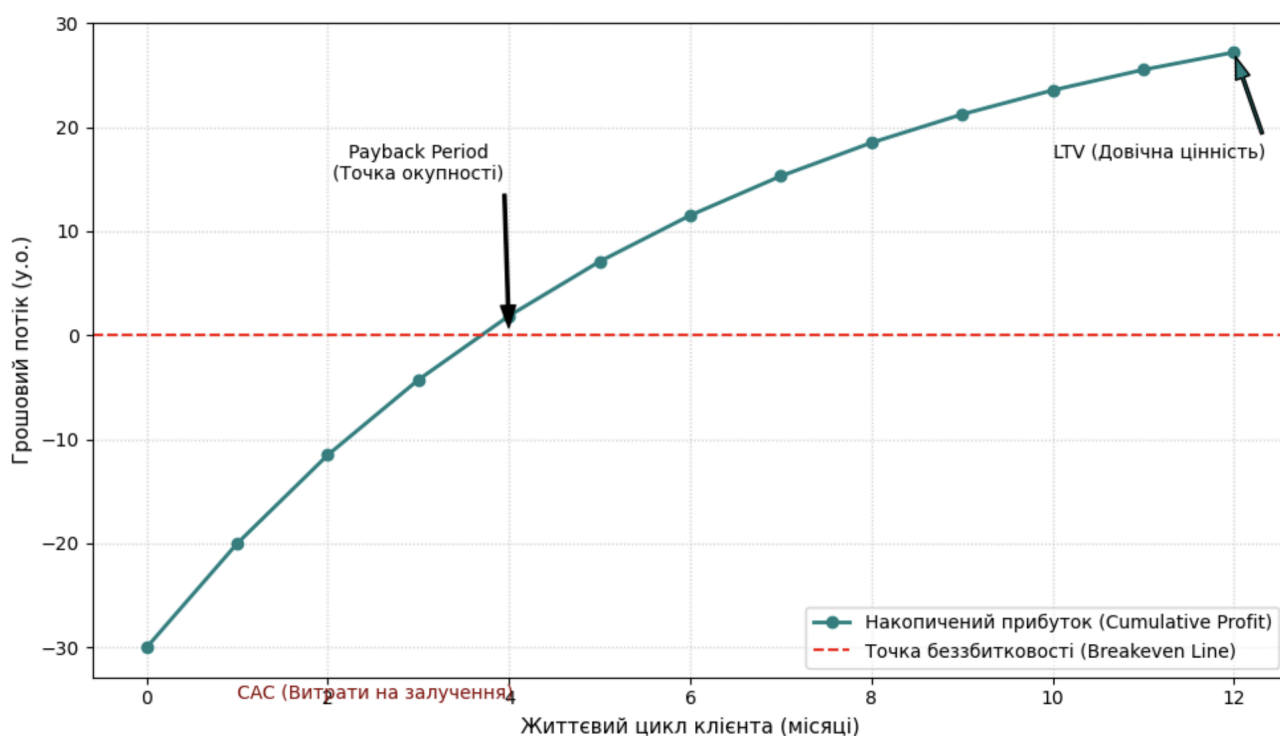


Рисунок 1.1 - Динаміка формування LTV та точки окупності

Джерело: розроблено автором

Таким чином, довічна цінність користувача виступає не просто фінансовою метрикою, а комплексним індикатором життєздатності цифрового бізнесу. Вона

синтезує в собі ефективність рекламних каналів, якість самого продукту, рівень сервісу та лояльність аудиторії. Перехід від інтуїтивного маркетингу до управління на основі математично обґрунтованого LTV дозволяє компаніям уникати цінових війн за «одноразових» покупців, концентруючи ресурси на формуванні стійкої бази прихильників бренду. Водночас складність нелінійних зв'язків між поведінкою користувачів та їхньою кінцевою цінністю вимагає розроблення формалізованих алгоритмічних рішень, що актуалізує потребу в глибокому огляді та систематизації методів математичного моделювання даного показника.

1.2. Огляд класичних та сучасних підходів до математичного моделювання LTV

Проблема математичного моделювання довічної цінності користувача не є новою для економічної та системної науки. Разом із тим, еволюція підходів до обчислення LTV нерозривно пов'язана з розвитком обчислювальних потужностей та накопиченням великих масивів цифрових даних (Big Data). Історично розвиток методів прогнозування цінності клієнта можна умовно поділити на три великі класи: класичні (евристичні та історичні) підходи, імовірнісні (стохастичні) моделі та сучасні алгоритми на базі машинного навчання (Machine Learning). Кожен із цих класів має власні математичні припущення, обмеження та сферу оптимального застосування в digital-кампаніях.

Класичні підходи до моделювання LTV здебільшого базуються на ретроспективному аналізі та агрегованих показниках. Найпростішою математичною абстракцією в цьому класі є розрахунок LTV як добутку середнього доходу на користувача (Average Revenue Per User, ARPU) та середньої тривалості його «життя» в продукті (Average Customer Lifespan). Хоча така модель є легкою для імплементації, з точки зору системного аналізу вона має критичні недоліки. Вона передбачає лінійність і стаціонарність поведінки клієнтів, що абсолютно не відповідає реаліям сучасного цифрового маркетингу. У динамічних системах, якими є digital-кампанії, користувачі демонструють високу дисперсію у своїх

витратах, а їхня поведінка зазнає впливу сезонності, змін у ринковій кон'юнктурі та маркетингових стимулів. Відповідно, використання усереднених метрик призводить до систематичної похибки (bias) у прогнозах, нівелюючи можливість персоналізації.

Більш просунутим класичним методом є RFM-аналіз (Recency, Frequency, Monetary), який сегментує користувачів на основі давності останньої транзакції, частоти покупок та грошового обсягу витрат. RFM-модель стала першим кроком до розуміння того, що минула поведінка клієнта є найкращим предиктором його майбутньої активності. Однак RFM залишається радше евристичним інструментом ранжування, ніж повноцінною предиктивною математичною моделлю, оскільки не дає точної кількісної оцінки майбутніх грошових потоків у неперервному часі.

Розуміння обмежень історичних методів зумовило перехід до другого класу - імовірнісних (стохастичних) моделей, також відомих у науковій літературі як моделі сімейства BTYD (Buy 'Till You Die). Фундаментальною працею в цій галузі стала модель Pareto/NBD (Negative Binomial Distribution), розроблена Д. Шмітцляйном та співавторами. Математичний апарат цієї моделі базується на припущенні, що кожен клієнт має два приховані (латентні) процеси, які неможливо спостерігати безпосередньо: процес здійснення покупок та процес «відмирання» (відтоку).

У моделях BTYD припускається, що поки клієнт є активним, кількість його транзакцій є розподілена за законом Пуассона з певним параметром інтенсивності λ . Водночас момент часу, коли клієнт назавжди припиняє взаємодію з компанією, моделюється як експоненційний розподіл із параметром μ . Оскільки інтенсивність покупок та схильність до відтоку є гетерогенними (різними для різних користувачів), параметри λ та μ моделюються через гамма-розподіли. Математичне очікування кількості майбутніх транзакцій $E[Y(t)]$ у таких стохастичних системах зазвичай подається у вигляді складних інтегральних рівнянь, де враховується ймовірність того, що клієнт все ще «живий» у момент часу t :

$$E[Y(t)] = \int_0^t P(\text{alive}|\tau) * E[v]d\tau$$

де $P(\text{alive}|\tau)$ - імовірність активності клієнта, а $E[v]$ - очікувана грошова цінність однієї транзакції.

Пізніше математичний апарат був спрощений та вдосконалений у вигляді моделі BG/NBD (Beta-Geometric/NBD), яка замінила експоненційний процес відтоку на дискретний процес з бета-геометричним розподілом. Хоча імовірнісні моделі мають міцне математичне підґрунтя та добре описують транзакційні процеси, вони мають суттєвий недолік у контексті сучасних digital-кампаній. Вони здатні враховувати лише історію транзакцій (ті ж самі Recency та Frequency) і дуже погано масштабуються для включення екзогенних коваріат (наприклад, даних про демографію, кліки по рекламі, час проведений у мобільному додатку, канали залучення).

Саме ця проблема зумовила перехід до третього, найсучаснішого класу - моделей на основі машинного навчання (Machine Learning, ML). В епоху багатовимірних даних, коли після першого контакту користувача з рекламою генеруються сотні параметрів (фіч), алгоритми машинного навчання демонструють найвищу ефективність. У цьому контексті задача прогнозування LTV зводиться до задачі регресії з учителем (Supervised Learning), де метою є мінімізація функції втрат між прогнозованим та фактичним значенням довічної цінності.

Сучасні підходи на базі ML варіюються від множинної лінійної регресії, регулярних моделей (Lasso, Ridge) до ансамблевих методів. Зокрема, алгоритми на основі дерев рішень, такі як Random Forest та градієнтний бустинг (XGBoost, LightGBM, CatBoost), стали індустріальним стандартом у digital-аналітиці. Їхня перевага над імовірнісними моделями полягає у здатності автоматично виявляти складні нелінійні залежності та взаємодії між десятками поведінкових характеристик користувача. Крім того, алгоритми бустингу ефективно працюють із розрідженими даними та пропусками, що є типовою ситуацією для баз даних веб-аналітики.

Ще одним інноваційним напрямом у моделюванні LTV є використання штучних нейронних мереж (Deep Learning). Оскільки взаємодія клієнта з digital-кампаніями є послідовним процесом у часі, дослідники все частіше застосовують рекурентні нейронні мережі (RNN), зокрема архітектури з довгою короткостроковою пам'яттю (LSTM). Такі математичні моделі здатні аналізувати не просто агреговані дані (загальну кількість кліків), а саму послідовність дій (наприклад: переглянув рекламу, зайшов на сайт, залишив кошик, отримав email, здійснив покупку). Використання багатошарових перцептронів дозволяє витягувати латентні ознаки з неструктурованих даних, створюючи високоточні предиктивні вектори.

Таким чином, огляд існуючих підходів до математичного моделювання довічної цінності користувача демонструє очевидний парадигмальний зсув. Якщо раніше бізнес задовільнявся описовою статистикою та жорсткими ймовірнісними розподілами, то сьогодні, в умовах високої розмірності даних digital-маркетингу, ключовим інструментарієм системного аналітика стають гнучкі та потужні алгоритми машинного навчання. Саме вони здатні врахувати багатовимірність клієнтського досвіду та забезпечити необхідну точність прогнозу для ефективної оптимізації цифрових рекламних бюджетів.

Поряд із розглянутими ймовірнісними моделями та алгоритмами машинного навчання, вагоме місце в еволюції підходів до обчислення LTV займає аналіз виживання (Survival Analysis). Цей математичний апарат, початково розроблений для медичних та демографічних досліджень з метою прогнозування тривалості життя, знайшов своє успішне застосування у системному аналізі клієнтської поведінки. У контексті digital-маркетингу «смерть» пацієнта інтерпретується як відтік користувача (churn), а час до настання цієї події визначає тривалість його життєвого циклу.

Фундаментальною моделлю в цьому класі є модель пропорційних ризиків Кокса (Cox Proportional Hazards Model). Її головна математична перевага полягає у здатності оцінювати вплив різних незалежних змінних (коваріат) на функцію

ризик (hazard function) відтоку клієнта в будь-який момент часу. Математично функція ризику $h(t|x)$ у моделі Кокса описується наступним рівнянням:

$$h(t|x) = h_0(t) \exp\left(\sum_{i=1}^n \beta_i x_i\right)$$

де $h_0(t)$ - базова функція ризику (ймовірність відтоку за умови, що всі коваріати дорівнюють нулю), x_i - набір предикторів (наприклад, активність користувача, канал залучення, демографічні дані), а β_i - коефіцієнти, що визначають вагу кожного предиктора. Інтегруючи цю функцію виживання з очікуваною монетарною цінністю транзакцій, аналітики отримують можливість обчислювати динамічний LTV. Незважаючи на високу статистичну надійність, моделі аналізу виживання мають певні обмеження: вони добре передбачають тривалість взаємодії, але гірше справляються з прогнозуванням частоти покупок і суми чека, тому їх найчастіше використовують для підписних (subscription-based) бізнес-моделей.

Ще одним потужним інструментом стохастичного моделювання, який історично передував складним нейронним мережам, є застосування марковських ланцюгів (Markov Chains). З точки зору теорії масового обслуговування та дослідження операцій, життєвий цикл клієнта розглядається як набір дискретних станів. Наприклад, користувач може перебувати у станах: «новий», «активний», «сплячий», «під загрозою відтоку» та «втрачений». Переходи між цими станами описуються матрицею ймовірностей переходів (Transition Matrix).

Довічна цінність у такій системі обчислюється як сума очікуваних винагород (грошових потоків), зважених на ймовірність перебування клієнта в певному стані на кожному майбутньому кроці (ітерації). Використання марковських процесів дозволяє компаніям не лише розраховувати сумарний LTV, але й моделювати сценарії «що, якщо» (what-if analysis). Наприклад, як зміниться загальний LTV бази користувачів, якщо завдяки новій маркетинговій кампанії ймовірність переходу зі стану «сплячий» у стан «активний» зросте на 5%. Втім, класичні марковські моделі володіють властивістю відсутності пам'яті (Markov

property), що означає залежність майбутнього стану лише від поточного, ігноруючи весь попередній шлях користувача. Це є суттєвим мінусом при аналізі довгих ланцюжків взаємодій у digital-кампаніях.

На сучасному етапі розвитку предиктивної аналітики спостерігається тенденція до синергії класичних, імовірнісних та оптимізаційних методів - створення гібридних моделей (Hybrid Models). Системні аналітики дійшли висновку, що жоден із наведених класів не є абсолютно універсальним. Імовірнісні моделі (як-от BG/NBD) чудово вловлюють приховану транзакційну динаміку, але ігнорують багатовимірні ознаки (фічі) користувача. Натомість моделі машинного навчання (наприклад, градієнтний бустинг) ідеально працюють із багатовимірними даними, але можуть видавати нестабільні результати на довгострокових горизонтах прогнозування через схильність до перенавчання (overfitting).

Суть гібридного підходу полягає у використанні вихідних параметрів імовірнісних моделей або моделей виживання як додаткових (синтетичних) предикторів для алгоритмів машинного навчання. Таким чином, нейронна мережа або ансамбль дерев рішень отримує на вхід не лише сирі дані про поведінку (кліки, сесії, транзакції), але й розраховану стохастичну ймовірність того, що клієнт все ще «живий». Це дозволяє об'єднати строгість математичної статистики з обчислювальною потужністю сучасного Data Science.

1.3. Роль багатовимірних даних та машинного навчання в задачах предиктивної аналітики поведінки користувачів

Сучасний етап розвитку інформаційного суспільства характеризується експоненційним зростанням обсягів даних, що генеруються в результаті життєдіяльності людини у цифровому просторі. Для системного аналітика digital-середовище постає як надскладна динамічна система, в якій кожен об'єкт (користувач) залишає безперервний потік інформаційних сигналів. Традиційні підходи до аналізу маркетингових показників, що оперували обмеженим набором агрегованих змінних, сьогодні виявляються неспроможними забезпечити

необхідну точність управління. Це обумовлює перехід до використання багатовимірних масивів даних (multidimensional data) та інструментарію машинного навчання як фундаменту предиктивної аналітики поведінки споживачів.

Багатовимірність даних у контексті прогнозування LTV означає, що кожна одиниця аналізу описується не одним чи двома показниками, а вектором ознак у просторі високої розмірності. Ці дані можна класифікувати за кількома ключовими доменами. По-перше, це транзакційні дані: історія платежів, типи придбаних товарів чи послуг, середній чек, використання промокодів. По-друге, це поведінкові (івент-орієнтовані) дані: логи взаємодії з інтерфейсом вебсайту чи мобільного додатку, кількість та тривалість сесій, глибина перегляду сторінок, швидкість скролінгу, реакція на push-повідомлення. По-третє, це контекстуальні та технічні дані: тип пристрою, операційна система, геопозиція, джерело переходу (реферальна ссылка, пошуковий запит, соціальна мережа). По-четверте, це демографічні та психографічні характеристики, що збираються через профілі користувачів.



Рисунок 1.2 - Структурні компоненти багатовимірних даних у digital-маркетингу

Джерело: розроблено автором

З точки зору системного аналізу, така багатовимірність створює так зване «прокляття розмірності» (curse of dimensionality), коли зі збільшенням кількості ознак дані стають розрідженими, а традиційні статистичні моделі втрачають свою передбачувальну силу. Саме тут виникає критична роль машинного навчання. Алгоритми ML за своєю природою розроблені для роботи у багатовимірних просторах, де вони здатні виявляти приховані закономірності та нелінійні залежності між ознаками, які неможливо ідентифікувати за допомогою класичних методів регресії. Роль машинного навчання у предиктивній аналітиці полягає у трансформації сирих, часто зашумлених даних у структуроване знання про майбутній стан системи.

Одним із ключових аспектів застосування ML у задачах LTV є етап конструювання ознак (feature engineering). Багатовимірні дані самі по собі є лише сировиною; завдання системного аналітика полягає у створенні таких похідних змінних, які максимально повно відображають латентні наміри користувача. Наприклад, замість того, щоб просто враховувати кількість кліків, алгоритм може аналізувати динаміку зміни активності за останні сім днів порівняно з середньомісячним показником. Така деривативна ознака є значно сильнішим предиктором відтоку чи майбутньої покупки. Машинне навчання дозволяє автоматизувати цей процес, використовуючи методи вибору ознак (feature selection) та зменшення розмірності (наприклад, метод головних компонент (PCA) або t-SNE), що дає змогу сконцентрувати обчислювальні ресурси на найбільш значущих факторах впливу.

Важливу роль у предиктивній аналітиці відіграє парадигма навчання з учителем (Supervised Learning). У задачах прогнозування LTV ми маємо справу з об'єктом, чия цільова мітка (фактичний дохід) стає відомою лише через значний проміжок часу. Багатовимірні дані, накопичені на ранніх етапах «життя» користувача, виступають як вхідний вектор X , а підсумкова цінність — як цільова функція y . Алгоритми машинного навчання, такі як випадкові лісові (Random Forest) або градієнтний бустинг (Gradient Boosting), будують складну апроксимаційну функцію $f(X) \approx y$. Особлива роль тут належить ансамблевим

методам, які поєднують прогнози багатьох «слабких» моделей для отримання одного потужного та стабільного результату, що мінімізує як зміщення, так і дисперсію прогнозу.

Окрім чистої регресії, велике значення має класифікаційне машинне навчання. В умовах багатовимірності часто доцільно спочатку розв'язати задачу класифікації користувачів за рівнем потенційної цінності (наприклад, «High-Value», «Medium-Value», «Low-Value»). Це дозволяє системі маркетингу застосовувати різні стратегії взаємодії для різних сегментів ще до того, як вони здійснять першу транзакцію. Роль багатовимірних даних тут проявляється у здатності алгоритму враховувати мікро-сигнали: наприклад, користувач, який прийшов з органічного пошуку в нічний час з пристрою останньої моделі, може мати статистично вищу ймовірність потрапляння в сегмент «High-Value».

Необхідно також відзначити роль глибинного навчання (Deep Learning) в аналізі часових рядів поведінкових даних. Сучасні нейромережеві архітектури, зокрема рекурентні нейронні мережі (RNN) та трансформери, розглядають шлях користувача не як набір статичних ознак, а як послідовність подій у часі. Кожна дія користувача впливає на внутрішній стан моделі, що дозволяє робити прогнози LTV у режимі реального часу (real-time prediction). Це створює умови для предиктивного маркетингу, коли система може автоматично змінити параметри рекламної кампанії для конкретного користувача безпосередньо в момент його перебування на сайті, максимізуючи його майбутню цінність.

Системна інтеграція багатовимірних даних та машинного навчання призводить до виникнення нової якості аналітики — переходу від описових моделей (що сталося?) до предиктивних (що станеться?) і, зрештою, до прескриптивних моделей (що зробити, щоб сталося?). У задачах LTV це означає не просто пасивне очікування прибутку, а активне формування поведінки клієнта через алгоритмічні підказки. Наприклад, якщо модель прогнозує високу ймовірність відтоку цінного клієнта на основі багатовимірного аналізу його останніх сесій, система автоматично генерує індивідуальну знижку чи бонус, рятуючи накопичений LTV.

Таким чином, багатовимірні дані виступають як інформаційний ресурс, а машинне навчання — як технологічний драйвер предиктивної аналітики. Їхня синергія дозволяє перетворити digital-кампанії з інтуїтивних процесів на точні інженерні системи, де кожне рішення базується на математично обґрунтованому прогнозі. Розвиток цих методів є ключовим напрямком системного аналізу в маркетингу, оскільки саме вони забезпечують здатність бізнесу адаптуватися до високої невизначеності та динамізму сучасного цифрового ринку. Впровадження таких алгоритмів дозволяє не лише оптимізувати витрати, а й створювати додану вартість через глибинне розуміння потреб та поведінкових паттернів кожного окремого користувача.

Окрім класичних структурованих масивів інформації, концепція багатовимірних даних у сучасній предиктивній аналітиці неминуче охоплює і неструктуровані дані. Цифровий слід користувача формується не лише через кліки чи транзакції, але й через текстові відгуки, коментарі в соціальних мережах, звернення до служби підтримки та навіть голосові запити. Для системного аналітика перетворення таких неструктурованих масивів у машинозчитуваний формат є одним із найскладніших етапів моделювання. Інтеграція методів обробки природної мови (Natural Language Processing, NLP) та аналізу тональності (Sentiment Analysis) дозволяє квантифікувати емоційне ставлення клієнта до бренду. Наприклад, семантичний аналіз історії звернень клієнта до чат-бота може виявити приховані патерни незадоволеності ще до того, як транзакційна активність почне знижуватися. У результаті, профіль користувача розширюється новими вимірами, які суттєво підвищують точність імовірнісних та регресійних алгоритмів при прогнозуванні його довічної цінності.

З математичної точки зору, інтеграція структурованих та неструктурованих даних призводить до того, що кожен користувач U_i репрезентується як складний вектор у n -вимірному просторі ознак: $U_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$, де кожна компонента x_{ij} відображає певну поведінкову, демографічну або семантичну метрику. Застосування алгоритмів машинного навчання, зокрема методів кластеризації

(наприклад, k-means або DBSCAN), дозволяє обчислювати відстані між векторами користувачів у цьому просторі (наприклад, використовуючи евклідову або косинусну відстань). Це дає змогу виявляти неочевидні мікросегменти аудиторії. Наприклад, алгоритм може встановити математичну подібність між поведінкою нового користувача та історичними профілями найбільш лояльних клієнтів (look-alike modeling), що автоматично класифікує новачка як потенційно високоприбуткового об'єкта впливу.

Впровадження таких складних математичних моделей вимагає відповідної модернізації IT-архітектури підприємства. Традиційні системи пакетної обробки даних (Batch Processing), які оновлюють прогнози раз на добу або тиждень, стають неефективними в умовах високодинамічного digital-середовища. Системний аналіз сучасних маркетингових інфраструктур вказує на необхідність переходу до Лямбда- (Lambda) або Каппа-архітектур (Kappa Architecture), що поєднують історичний аналіз із потоковою обробкою даних у режимі реального часу (Stream Processing). Використання таких технологій, як Apache Kafka або Apache Spark Streaming, дозволяє алгоритмам машинного навчання миттєво перераховувати прогнозований LTV користувача безпосередньо під час його навігації по сайту. Зміна значення LTV в реальному часі слугує тригером для автоматизованих систем персоналізації, які миттєво адаптують рекламні пропозиції, динамічне ціноутворення (dynamic pricing) або алгоритми рекомендацій під поточний стан клієнта.

Водночас, розглядаючи роль багатовимірних даних, неможливо оминути увагою системні обмеження, що виникають через глобальні зміни у сфері захисту конфіденційності (Data Privacy). Введення суворих регуляторних норм, таких як GDPR (Загальний регламент про захист даних у ЄС) або CCPA (Закон Каліфорнії про конфіденційність споживачів), а також поступова відмова провідних технологічних корпорацій від використання сторонніх файлів cookie (third-party cookies), докорінно змінюють ландшафт збору даних. Системні аналітики стикаються з явищем «фрагментації ідентичності», коли побудувати безперервний

ланцюжок поведінки користувача між різними пристроями та платформами стає все складніше.

У цих умовах машинне навчання набуває ще однієї критичної функції — предиктивного відновлення пропущених даних (Data Imputation) та імовірнісного метчингу (Probabilistic Matching). Коли прямий детермінований збір багатовимірних даних неможливий через налаштування приватності, алгоритми ML здатні апроксимувати відсутні змінні на основі наявних патернів (First-party data). Компанії змушені зміщувати фокус з агресивного збору сторонніх даних на максимально глибокий математичний аналіз тих даних, які користувачі надають добровільно під час прямої взаємодії з брендом. Таким чином, машинне навчання не лише підвищує точність маркетингових прогнозів, але й забезпечує стійкість (resilience) аналітичних систем до зовнішніх регуляторних шоків.

РОЗДІЛ 2

АНАЛІТИКО-МЕТОДОЛОГІЧНІ ОСНОВИ ПРОГНОЗУВАННЯ LTV В DIGITAL-КАМПАНІЯХ

2.1. Формалізація та математична постановка задачі прогнозування довічної цінності користувача

Побудова будь-якої предиктивної моделі в межах методології системного аналізу вимагає чіткої математичної формалізації об'єкта дослідження, встановлення границь системи та визначення цільових критеріїв оптимізації. В умовах сучасного цифрового маркетингу взаємодія користувача з digital-кампанією не може розглядатися як статична подія. Навпаки, це надзвичайно складний, багатоетапний стохастичний процес, що динамічно розгортається у часі та просторі багатовимірних поведінкових ознак. З огляду на це, першочерговим завданням є переведення бізнес-проблеми у строгу математичну площину. Математична постановка задачі прогнозування довічної цінності користувача (LTV) зводиться до пошуку оптимального відображення між спостережуваним профілем поведінки клієнта на початкових етапах його взаємодії з системою та його сумарною фінансовою цінністю у довгостроковій перспективі.

Для строгої формалізації цього процесу введемо базові позначення. Нехай U — множина всіх унікальних користувачів, які були залучені в екосистему продукту в результаті проведення певної digital-кампанії. Для кожного окремого користувача $u \in U$ загальний процес його життєвого циклу можна розділити на два ключові часові інтервали, які розділені точкою прийняття рішення. Перший інтервал — це вікно спостереження (observation window), яке позначимо як $[0, T_{obs}]$, де точка $t = 0$ відповідає моменту першого контакту користувача з кампанією (наприклад, перехід за рекламним посиланням, реєстрація на платформі або встановлення мобільного додатка). Вибір оптимальної довжини цього вікна є класичною задачею компромісу (trade-off) у системному аналізі: занадто коротке вікно не дозволить зібрати достатньо інформації про патерни поведінки, а занадто довге — призведе до запізнення прогнозів, що унеможливить

оперативну оптимізацію рекламних бюджетів. Відповідно, другий інтервал визначається як горизонт прогнозування (prediction horizon) і позначається як $(T_{obs}, T_{pred}]$, де T_{pred} — це цільовий часовий рубіж, для якого бізнес прагне оцінити сумарну фінансову віддачу від користувача.

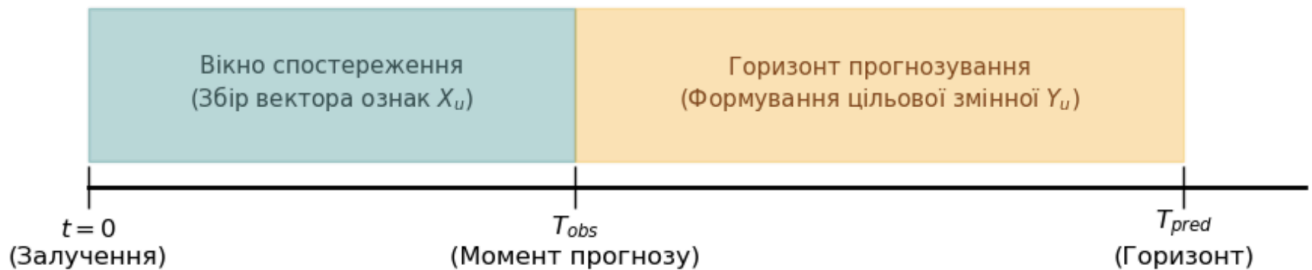


Рисунок 2.1 - Часова структура задачі прогнозування довічної цінності

Джерело: розроблено автором

Протягом вікна спостереження $[0, T_{obs}]$ аналітичні системи збирають та агрегують багатовимірний вектор ознак користувача $x_u \in R^d$, де d — розмірність простору ознак. Цей вектор є математичною репрезентацією цифрового сліду клієнта і включає в себе всі доступні історичні метрики: транзакційні показники, щільність сесійної активності, параметри взаємодії з інтерфейсом, соціодемографічні характеристики та атрибуційні ваги каналів залучення.

З точки зору фінансової математики та економічної теорії, фактична довічна цінність користувача y_u на заданому горизонті T_{pred} є сумою дисконтованих чистих грошових потоків, які компанія здатна згенерувати завдяки цьому користувачеві. У дискретному часі ця фундаментальна економічна величина формалізується наступним чином:

$$y_u = \sum_{t=1}^{T_{pred}} \frac{R_{u,t} - C_{u,t}}{(1+r)^t}$$

де $R_{u,t}$ — загальний дохід (revenue), отриманий від користувача u в період часу t ; $C_{u,t}$ — сукупні витрати (costs), пов'язані з обслуговуванням, технічною підтримкою, процесингом та маркетинговим утриманням цього користувача в

період t ; r — ставка дисконтування. Введення ставки дисконтування є критично важливим системним кроком, адже вона враховує не лише зміну вартості грошей у часі (інфляційні процеси), але й специфічні макроекономічні ризики та альтернативні витрати капіталу в турбулентному середовищі digital-бізнесу.

Оскільки в момент часу T_{obs} (коли приймається рішення щодо подальших маркетингових інвестицій) майбутні значення доходів і витрат є принципово невідомими, величина y_u виступає як прихована (латентна) змінна. За таких умов, завдання прогнозування LTV класифікується в термінології машинного навчання як задача контрольованого навчання (Supervised Learning), а саме — задача багатовимірної регресії на гетерогенних даних. Наша глобальна мета зводиться до знаходження такої функції апроксимації F , параметризованої набором невідомих ваг Θ , яка здатна з максимально можливою точністю відобразити вхідний вектор спостережуваних ознак у прогнозоване значення довічної цінності:

$$\hat{y}_u = F(x_u, \Theta)$$

де $\hat{y}_u$ — прогнозована оцінка LTV для користувача u .

Успішна формалізація цієї проблеми потребує введення цільової функції (об'єктивної функції), мінімізація якої дозволить алгоритму оптимізувати параметри Θ в гіперпараметричному просторі моделі. Згідно з класичною теорією статистичного навчання, для оцінки розбіжності між фактичним вектором цільових значень Y та прогнозованим вектором $\hat{Y}$ розраховується функція емпіричного ризику. Процес ітеративного навчання моделі на навчальній (тренувальній) вибірці обсягом N математично формулюється як мінімізація наступного функціоналу:

$$\Theta^* = \arg \min_{\Theta} \left(\frac{1}{N} \sum_{u=1}^N L(y_u, F(x_u, \Theta)) + \lambda \Omega(\Theta) \right)$$

де $L(.)$ — обрана функція втрат, що штрафує модель за помилки (наприклад, квадратична функція втрат $(y_u - \hat{y}_u)^2$), $\Omega(\Theta)$ — регуляризаційний

член, а λ — коефіцієнт, що контролює інтенсивність регуляризації. Включення регуляризаційного доданку є обов'язковою умовою при роботі з багатовимірними маркетинговими даними для запобігання ефекту перенавчання (overfitting). Наприклад, використання L_1 -норми (Lasso) сприяє відбору найважливіших ознак шляхом обнулення ваг незначущих змінних, тоді як L_2 -норма (Ridge) запобігає надмірному зростанню абсолютних значень коефіцієнтів, забезпечуючи загальну стабільність системи.

Проте, формалізація задачі в реальних умовах digital-кампаній стикається із серйозними структурними та статистичними викликами. Першим і найскладнішим бар'єром є екстремальна асиметрія та зміщеність розподілу цільової змінної Y до нуля (Zero-Inflated and Heavy-Tailed Distribution). На практиці лєвова частка залученого трафіку ніколи не конвертується у платоспроможних клієнтів. Таким чином, для масиву користувачів значення y_u залишається нульовим. Водночас мізерна частка високоякісних користувачів формує диспропорційно великий дохід (ефект Парето або наявність «довгого хвоста»). Спроба налаштувати класичну регресійну модель на таких даних неминуче призводить до того, що алгоритм починає занижувати прогнози, тяжіючи до медіанного значення, близького до нуля.

Для подолання цієї системної проблеми доцільно застосовувати декомпозиційний підхід — так звану бар'єрну модель (Two-Stage Hurdle Model). Суть підходу полягає у розбитті процесу на два етапи. На першому етапі моделюється психологічний та фінансовий бар'єр першої транзакції шляхом розв'язання задачі бінарної класифікації. Ми оцінюємо ймовірність того, що конкретний користувач взагалі подолає цей бар'єр і принесе хоча б мінімальний дохід:

$$P(y_u > 0 | x_u) = \sigma(W * x_u + b)$$

де $\sigma(\cdot)$ — логістична функція активації (сигмоїда), яка трансформує лінійну комбінацію ознак у ймовірнісний простір $[0, 1]$. Лише на другому етапі, і тільки для тієї відфільтрованої підмножини користувачів, для яких $y_u > 0$,

застосовується регресійний апарат, який визначає безпосередню монетарну цінність. Фінальний прогноз $\hat{y}_u$ тоді обчислюється як добуток імовірності конверсії на очікувану суму доходу.

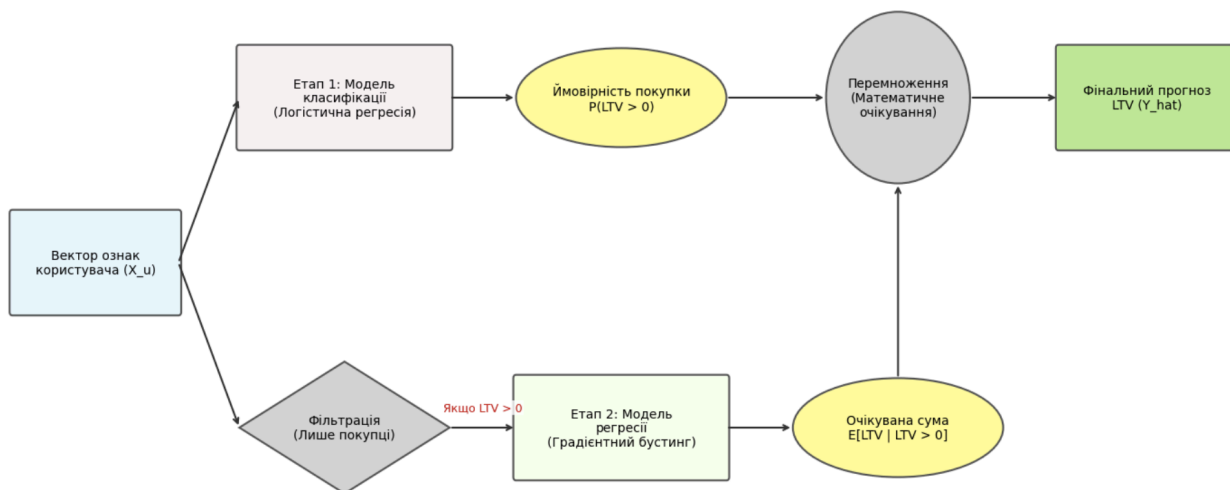


Рисунок 2.2 - Логічна архітектура Двохетапної бар'єрної моделі

Джерело: розроблено автором

Другим концептуальним викликом є проблема правого цензурування часових рядів (Right Censoring). В екосистемі безперервного сервісу життєвий цикл клієнта може бути тимчасово призупинений, а не остаточно завершений. На момент зрізу даних для побудови моделі частина користувачів все ще залишається активною, і їхній фактичний LTV продовжуватиме зростати в майбутньому. Ігнорування цього факту і прирівнювання поточного доходу до остаточного y_u вносить у модель систематичну негативну похибку (underestimation bias). Компенсація цього ефекту вимагає глибокого залучення методів аналізу виживання або формування суворих статичних когорт користувачів.

2.2. Системний аналіз багатовимірного простору ознак та методи відбору предикторів

У межах системного підходу до моделювання довічної цінності користувача, простір ознак (feature space) розглядається як структурована множина

змінних, що відображають різні аспекти взаємодії суб'єкта (користувача) та об'єкта (digital-платформи). Складність задач системного аналізу в маркетингових кампаніях зумовлена тим, що ці дані є не лише багатовимірними, а й часто мають ієрархічну структуру, високу степінь зашумленості та динамічність. Процес ідентифікації та аналізу такого простору є критично важливим етапом, оскільки від якості вхідних предикторів залежить здатність моделі до узагальнення та точність фінальних прогнозів.

Багатовимірний простір ознак у задачах прогнозування LTV можна декомпонувати на декілька функціональних підсистем:

1. Транзакційна підсистема: включає параметри фінансової активності — суми чеків, частоту покупок, використання дисконтних програм, середній прибуток на транзакцію. Ці ознаки є найбільш репрезентативними для короткострокових прогнозів, проте часто мають високу дисперсію.
2. Поведінкова підсистема: формується на основі «цифрового сліду» користувача — логів сесій, глибини перегляду сторінок, часу активності, швидкості скролінгу та клікабельності окремих елементів інтерфейсу. Системний аналіз цих даних дозволяє виявити латентні патерни залученості, які передують здійсненню покупки.
3. Технічна та контекстуальна підсистема: охоплює характеристики середовища взаємодії — тип пристрою, операційну систему, географічне розташування, час доби та джерело переходу. Ці параметри дозволяють врахувати зовнішні фактори впливу на LTV.
4. Соціодемографічна підсистема: вік, стать, професійна сфера та інші анкетні дані, що дозволяють будувати статичні профілі користувачів.

Проте, пряме використання всіх доступних змінних у математичному моделюванні стикається з фундаментальною проблемою — «прокляттям розмірності» (curse of dimensionality). Зі зростанням кількості ознак об'єм простору збільшується настільки стрімко, що наявні дані стають розрідженими (sparse). Це ускладнює пошук статистично значущих закономірностей і призводить до ефекту перенавчання моделі, коли алгоритм запам'ятовує шум у

навчальній вибірці замість реальних залежностей. Тому центральною задачею системного аналітика стає відбір оптимальної підмножини предикторів, яка б зберігала максимум інформаційної ємності при мінімальній кількості вимірів.

Процес підготовки та відбору ознак починається з етапу попереднього оброблення (preprocessing). На цьому етапі проводиться очищення даних від аномалій (outliers), які в digital-кампаніях часто спричинені технічними збоями аналітики або активністю ботів. Окремим завданням є обробка пропущених значень (missing data). Математично це реалізується або через видалення відповідних об'єктів, або через методи імпутації (imputation) — заміни пропусків середніми, медіанними значеннями або результатами прогнозів допоміжних моделей. Важливим елементом є також нормалізація та стандартування числових ознак, що дозволяє привести різні фізичні величини (наприклад, суму грошей та кількість секунд на сайті) до єдиної шкали, забезпечуючи рівний вплив ознак на процес оптимізації ваг моделі.

Особливу увагу в системному аналізі приділяють Feature Engineering — конструюванню нових синтетичних предикторів. Часто «сирі» дані не мають прямої кореляції з LTV, проте їхні комбінації можуть мати високу передбачувальну силу. Наприклад, відношення кількості сесій за останній тиждень до загальної кількості сесій за місяць (коефіцієнт динаміки залученості) може бути значно сильнішим предиктором майбутньої цінності, ніж просто загальна активність. Створення таких ознак вимагає глибокого розуміння предметної області та використання математичних функцій трансформації (логарифмування, поліноміальне розширення тощо).

Після формування розширеного простору ознак застосовуються методи відбору предикторів (feature selection), які можна класифікувати на три типи:

- Методи фільтрації (Filter methods): базуються на оцінці статистичних властивостей самих ознак незалежно від обраного алгоритму навчання. Використовуються критерії інформаційної ентропії (Mutual Information), коефіцієнти кореляції Пірсона або Спірмена, а також поріг варіативності (Variance Threshold). Ці методи є обчислювально ефективними та дозволяють швидко

видалити надлишкові або неінформативні змінні, що не мають зв'язку з цільовою міткою.

- Методи обгортки (Wrapper methods): розглядають підмножини ознак як входні параметри для моделі та оцінюють їхню якість на основі точності прогнозу. Найбільш відомим є метод рекурсивного виключення ознак (Recursive Feature Elimination, RFE). Незважаючи на високу точність відбору, ці методи вимагають значних обчислювальних витрат, особливо у багатовимірних просторах, через необхідність багаторазового перенавчання моделі.
- Вбудовані методи (Embedded methods): інтегрують процес відбору безпосередньо в алгоритм навчання. Найяскравішим прикладом є L1-регуляризація (Lasso), яка під час мінімізації функції втрат обнуляє коефіцієнти перед неважливими ознаками, фактично видаляючи їх із системи. Також широкого застосування набув аналіз важливості ознак (Feature Importance) у моделях на основі дерев рішень, де кожна змінна оцінюється за ступенем зниження неоднорідності (Gini Impurity) або помилки при розбитті вузлів дерева.

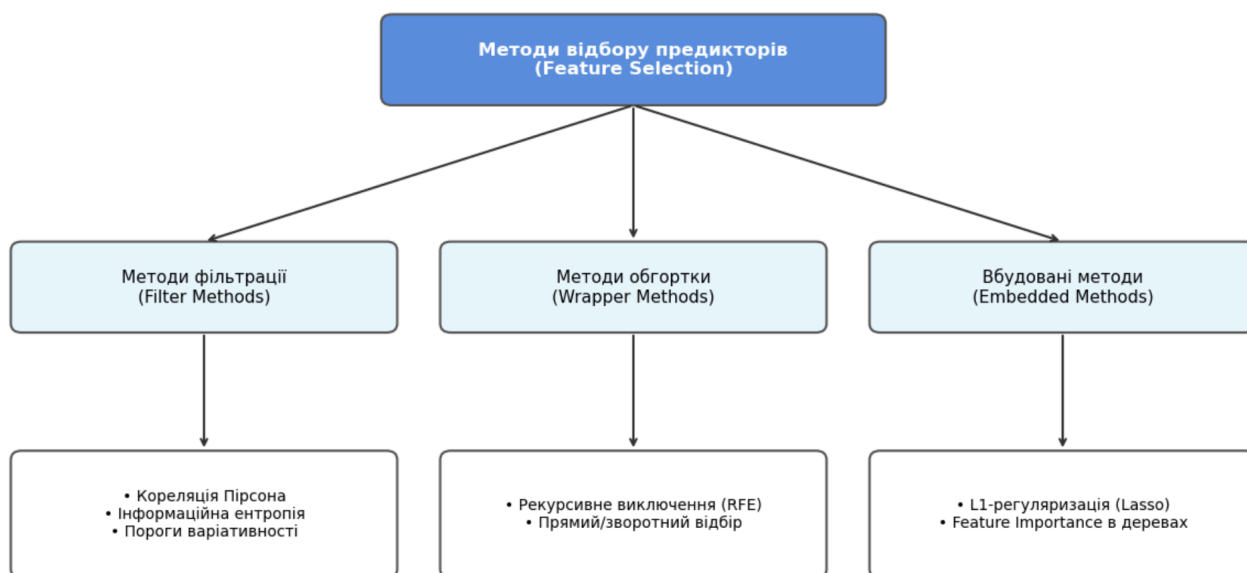


Рисунок 2.3 - Класифікація методів відбору інформативних ознак

Джерело: розроблено автором

Крім того, для зниження розмірності простору без втрати структури даних часто використовується метод головних компонент (Principal Component Analysis,

РСА). З точки зору системного аналізу, RSA дозволяє трансформувати вихідний простір корельованих ознак у новий простір ортогональних (незалежних) головних компонент, які є лінійними комбінаціями вихідних змінних. Це дозволяє сконцентрувати основну частку дисперсії даних у декількох перших компонентах, значно спрощуючи архітектуру майбутньої предиктивної моделі.

Одним із найскладніших викликів при аналізі ознак є проблема мультиколінеарності — наявності сильних лінійних залежностей між предикторами. У маркетингових даних це явище є типовим: наприклад, кількість переглядів сторінок та час проведений на сайті зазвичай сильно корелюють. Системне ігнорування мультиколінеарності призводить до нестабільності моделі та неможливості адекватної інтерпретації вагових коефіцієнтів. Для виявлення таких зв'язків використовується фактор інфляції дисперсії (Variance Inflation Factor, VIF) та матриці парних кореляцій.

Завершальним етапом аналізу є оцінка стійкості (stability) обраних предикторів у часі. Оскільки поведінка користувачів у digital-кампаніях схильна до дрейфу (data drift), важливо переконатися, що зв'язок між предиктором та LTV залишається стабільним на різних часових інтервалах. Це досягається шляхом порівняння розподілів ознак на тренувальних та валідаційних вибірках.

Таким чином, системний аналіз багатовимірному простору ознак є ітераційним процесом трансформації сирого інформаційного потоку в лаконічний та репрезентативний набір предикторів. Використання комбінації статистичних фільтрів, методів регуляризації та процедур Feature Engineering дозволяє побудувати математичний фундамент, на якому алгоритми машинного навчання зможуть ефективно виявляти складні нелінійні залежності, необхідні для точного прогнозування довічної цінності користувача. Описана методологія відбору забезпечує не лише математичну точність, а й інтерпретованість моделі, що є критично важливим для прийняття стратегічних управлінських рішень у сфері digital-маркетингу.

2.3. Математичні моделі машинного навчання для предиктивної аналітики LTV

Перехід від описової статистичної аналітики до предиктивного моделювання в межах системного аналізу digital-кампаній зумовлює необхідність вибору адекватного математичного інструментарію. Специфіка прогнозування довічної цінності користувача (LTV) полягає в тому, що залежності між поведінковими маркерами клієнта та його фінальною фінансовою цінністю мають глибоко нелінійний, стохастичний та часто асиметричний характер. За таких умов класичні параметричні методи регресії демонструють недостатню гнучкість. Тому в даному дослідженні розроблено методологію експериментального порівняння трьох різних класів алгоритмів машинного навчання (Machine Learning), що дозволяє системно оцінити їхню ефективність.

З методологічної точки зору, дослідження побудоване на послідовному ускладненні математичного апарату, що передбачає реалізацію та порівняння трьох концептуальних підходів.

Перший підхід: Базова модель на основі алгоритму Випадкового лісу (Random Forest). У системному аналізі розроблення будь-якої інтелектуальної системи розпочинається з побудови базової моделі (baseline), яка слугує еталоном для подальшого оцінювання більш складних алгоритмів. У якості такого еталона було обрано алгоритм випадкового лісу (Random Forest). З математичної точки зору, цей метод є ансамблевою моделлю, що базується на принципі бегтінгу (Bootstrap Aggregating).

Суть даного підходу полягає у генеруванні множини незалежних дерев рішень, кожне з яких навчається на випадковій підвибірці об'єктів та ознак із загального багатовимірного датасету. Кожне дерево рекурсивно розбиває простір ознак, намагаючись мінімізувати дисперсію цільової змінної всередині утворених вузлів. Фінальний прогноз моделі формується шляхом простого математичного усереднення прогнозів усіх дерев в ансамблі. Філософія цього алгоритму спирається на закон великих чисел: хоча кожне окреме дерево може бути схильним до перенавчання (overfitting) та мати високу дисперсію, їхнє агрегування

дозволяє взаємно компенсувати індивідуальні похибки. Для прогнозування LTV цей підхід є цінним завдяки своїй стійкості до аномальних викидів (наприклад, одиничних клієнтів з аномально великими транзакціями) та здатності автоматично виявляти прості нелінійні взаємодії між ознаками.

Другий підхід: Просунута модель Градієнтного бустингу (Gradient Boosting) Незважаючи на стабільність базової моделі, сучасні реалії цифрового маркетингу вимагають застосування більш чутливих та ітеративних алгоритмів. Тому другим досліджуваним методом виступає алгоритм градієнтного бустингу (Gradient Boosting Machine). Концептуальна відмінність цього методу від випадкового лісу полягає у філософії навчання: замість паралельної побудови незалежних дерев, бустинг формує ансамбль послідовно.

Математично цей процес описується як оптимізація цільової функції втрат у просторі функцій за допомогою градієнтного спуску. Кожна наступна слабка модель (нове дерево) будується не для прогнозування самої величини LTV, а для того, щоб максимально точно передбачити і компенсувати похибку (так звані псевдозалишки), яку допустили всі попередні дерева. Завдяки такому ітеративному коригуванню помилок, алгоритм здатний виявляти надзвичайно тонкі, приховані закономірності у багатовимірних даних. Використання градієнтного бустингу дозволяє системі «концентрувати увагу» на найбільш складних для прогнозування сегментах користувачів, поступово уточнюючи фінансові очікування. Проте цей метод є більш чутливим до налаштування гіперпараметрів і вимагає суворої математичної регуляризації для запобігання втраті здатності до узагальнення.

Третій підхід: Кастомна двохетапна бар'єрна модель (Two-Stage Hurdle Model) є авторською системною модифікацією, спрямованою на вирішення головної статистичної проблеми маркетингових даних — надмірної кількості нульових значень цільової змінної (Zero-Inflation). У реальних digital-кампаніях більшість користувачів (до 80%) не здійснює жодної транзакції. Пряме застосування навіть таких потужних алгоритмів, як градієнтний бустинг, до загального масиву даних призводить до зміщення прогнозів у бік нуля, оскільки

алгоритм намагається мінімізувати загальну середню похибку. Для подолання цього бар'єра розроблено двохетапну математичну архітектуру, яка декомпозиє задачу на два незалежні процеси:

1. Етап класифікації (оцінка ймовірності): На першому кроці будується модель класифікації (наприклад, логістична регресія або класифікаційний бустинг), яка оцінює математичну ймовірність $P(LTV > 0)$ того, що конкретний користувач взагалі подолає бар'єр першої транзакції. Цей етап дозволяє алгоритмічно відфільтрувати нерелевантний трафік.

2. Етап регресії (оцінка монетарної цінності): На другому кроці застосовується регресійна модель, але вона навчається виключно на підмножині платоспроможних клієнтів (де історичний $LTV > 0$). Ця модель спеціалізується лише на прогнозуванні обсягу майбутніх витрат, не відволікаючись на користувачів, які нічого не купують.

Фінальний прогноз довічної цінності в цій системній моделі обчислюється як математичне очікування — добуток ймовірності конверсії (з першого етапу) на очікувану суму доходу (з другого етапу). Такий структурний підхід дозволяє найбільш адекватно відтворити реальну економічну логіку поведінки споживача.

2.4. Методологія валідації моделей та специфіка оцінювання в умовах часових рядів

У межах методології системного аналізу розроблення предиктивних комплексів не завершується на етапі вибору математичного алгоритму та його навчання. Критично важливою ланкою, що визначає придатність розробленої системи до реальної експлуатації, є етап валідації та оцінювання. Для задач прогнозування довічної цінності користувача (LTV) у digital-маркетингу побудова надійної методології тестування ускладнюється динамічною, стохастичною природою вхідних потоків даних. На відміну від статичних задач класифікації чи регресії, де об'єкти зазвичай вважаються незалежними та однаково розподіленими (припущення i.i.d.), дані про поведінку користувачів мають жорстку часову структуру. Це вимагає розроблення спеціалізованого підходу до перевірки, який

би запобігав методичним помилкам і забезпечував репрезентативність результатів для всіх трьох обраних нами класів моделей.

Центральною проблемою при валідації маркетингових моделей є ризик витоку даних (data leakage) або зміщення, спричиненого загляданням у майбутнє (look-ahead bias). У системному аналізі логів цифрової взаємодії цей ризик матеріалізується тоді, коли модель у процесі навчання отримує доступ до інформації, яка хронологічно виникає пізніше моменту формування прогнозу. Стандартний підхід до випадкового розбиття даних на навчальну та тестову вибірки (random K-Fold cross-validation) у таких умовах є категорично неприпустимим. Якщо алгоритм навчається на даних клієнтів із грудня, а тестується на клієнтах із жовтня того ж року, порушується принцип причинно-наслідкового зв'язку, що створює ілюзорну, завищену точність на етапі тестування.

Для нівелювання цих ризиків у даному дослідженні методологія базується на принципі «тимчасового розрізу» (Out-of-Time Validation, OOT). Згідно з цим підходом, генеральна сукупність розділяється не випадковим чином, а на основі часової мітки залучення користувача (когортний аналіз). Модель агрегує знання виключно на історичних когортах (наприклад, клієнти, залучені в першому кварталі), а перевіряє свою здатність до узагальнення на нових когортах (клієнти другого кварталу). Таке розділення імітує реальний сценарій роботи retention-системи: алгоритм повинен довести свою ефективність на «майбутніх» користувачах, поведінкові патерни яких можуть зазнати впливу нових рекламних стратегій, макроекономічних коливань чи сезонності.

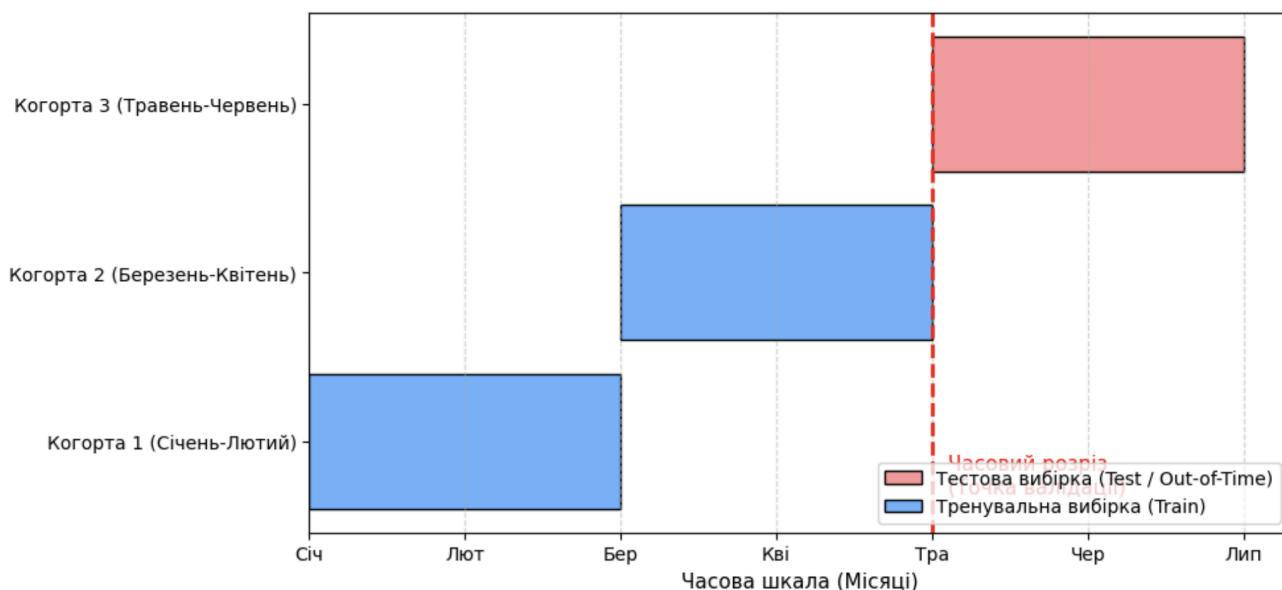


Рисунок 2.4 - Схема Out-of-time валідації предиктивних моделей

Джерело: розроблено автором

Оскільки в межах нашого дослідження реалізується послідовна еволюція алгоритмів (від простого усереднення до складної композитної архітектури), методологія валідації має бути адаптована під специфіку кожного з трьох методів:

1. Валідація базової моделі (Random Forest): Для ансамблю незалежних дерев рішень головною метою валідації є перевірка стабільності дисперсії. Оскільки алгоритм використовує внутрішній механізм Out-of-Bag (OOB) оцінювання під час процедури беггінгу, він є відносно стійким до перенавчання. Проте на етапі Out-of-Time валідації важливо переконатися, що ліс не «запам'ятав» специфічні аномалії історичного періоду. Оцінювання цієї моделі слугує глобальним бенчмарком (baseline) для всієї системи.

2. Валідація моделі градієнтного бустингу (XGBoost): На відміну від випадкового лісу, ітеративна природа бустингу робить його надзвичайно вразливим до перенавчання на пізніх ітераціях. Кожне нове дерево намагається виправити залишкову помилку, і без належного контролю алгоритм почне моделювати інформаційний шум. Тому для цього методу впроваджується процедура ранньої зупинки (Early Stopping). Вона вимагає виділення додаткової валідаційної підмножини даних: алгоритм моніторить похибку на цій підмножині

після кожного кроку і зупиняє навчання, як тільки похибка починає зростати, фіксуючи оптимальну кількість дерев в ансамблі.

3. Дворівнева валідація бар'єрної моделі (Two-Stage Hurdle Model): Це найбільш складний етап системного оцінювання. Оскільки модель декомпозована на два етапи, її валідація також відбувається ізольовано з подальшим синтезом:

- Оцінка етапу класифікації: Перевіряється здатність моделі відрізнати платоспроможних клієнтів від «нульових». Для цього використовуються класифікаційні метрики: логарифмічна функція втрат (LogLoss), яка штрафує за невпевненість у прогнозі, та матриця плутанини (Confusion Matrix), що дозволяє оцінити баланс між хибнопозитивними (False Positive) та хибнонегативними (False Negative) прогнозами.

- Оцінка етапу регресії: Оцінюється виключно на підмножині реальних покупців.

- Синтетична оцінка: Фінальний прогноз (результат множення ймовірності на прогнозовану суму) порівнюється з фактичним LTV усього масиву користувачів за допомогою єдиних регресійних метрик, що дозволяє коректно зіставити цей метод із двома попередніми.

Специфіка розподілу довічної цінності, що характеризується наявністю «довгого хвоста» (екстремально великих чеків невеликої групи лояльних клієнтів), робить використання лише однієї метрики недостатнім для об'єктивного системного аналізу. Тому ефективність усіх трьох методів оцінюватиметься за комплексним набором показників:

- Середня абсолютна похибка (Mean Absolute Error, MAE): Ця метрика розраховується як середнє арифметичне абсолютних відхилень прогнозу від факту. Її головна математична перевага — лінійність штрафів. Вона не гіперболізує помилки на екстремальних значеннях (аномаліях), надаючи найбільш чесну оцінку якості моделі для середньостатистичного, типового користувача [25]. У бізнес-контексті MAE показує, на скільки умовних одиниць у середньому помиляється алгоритм при оцінці одного клієнта.

– Коренева середньоквадратична похибка (Root Mean Squared Error, RMSE): На відміну від MAE, ця метрика зводить відхилення у квадрат перед усередненням. З точки зору системного аналізу, RMSE є індикатором чутливості моделі до великих промахів. Якщо модель прогнозує для потенційного «кита» нульовий LTV (або навпаки, очікує тисячі доларів від клієнта, який нічого не купив), RMSE різко зростає. Мінімізація цього показника є критично важливою для забезпечення фінансової стабільності рекламних кампаній, де ціна грубої помилки є надто високою.

– Коефіцієнт детермінації (R^2): Це відносна статистична метрика, яка вказує, яку частку загальної дисперсії (варіативності) цільової змінної LTV вдалося математично пояснити за допомогою нашого багатовимірного простору ознак. Значення R^2 , близьке до 1, свідчатиме про ідеальну точність алгоритму, тоді як значення біля 0 означатиме, що складна модель машинного навчання працює не краще, ніж просте передбачення середнього історичного значення.

РОЗДІЛ 3

ПРОГРАМНА РЕАЛІЗАЦІЯ ТА ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ПРЕДИКТИВНИХ МОДЕЛЕЙ

3.1. Архітектура системи прогнозування довічної цінності користувачів

Проектування архітектури системи прогнозування довічної цінності користувачів (LTV) у межах системного аналізу digital-кампаній вимагає цілісного підходу до організації інформаційних потоків. Розроблена система розглядається як багаторівневий аналітичний комплекс, призначений для трансформації сирих, багатовимірних даних про поведінку суб'єктів у стратегічні бізнес-інсайти. Архітектурне рішення базується на принципах модульності, масштабованості та здатності до ітеративного вдосконалення математичного ядра.



Рисунок 3.1 - Архітектурний контур предиктивної системи прогнозування LTV

Джерело: розроблено автором

Загальну логічну схему системи можна представити у вигляді послідовності функціональних шарів, кожен із яких вирішує специфічний набір завдань: шар збору та інтеграції даних, шар попереднього оброблення та трансформації ознак, шар предиктивного моделювання та шар візуалізації і прийняття рішень.

Первинною ланкою архітектури є механізм накопичення інформації. У реальних digital-екосистемах дані генеруються децентралізовано: через трекери мобільних додатків, пікселі веб-аналітики, CRM-системи та рекламні кабінети платформ (Paid Social, Google Ads тощо). Для цілей даного дослідження та

забезпечення відтворюваності результатів, архітектура передбачає використання спеціалізованого модуля генерації синтетичних даних. Даний модуль відтворює складну структуру багатовимірних ознак, включаючи транзакційну активність, технічні параметри (тип ОС, пристрою) та поведінкові індикатори (кількість сесій, тривалість взаємодії, залученість до воронки продажів). Системна особливість цього шару полягає у збереженні логічних зв'язків та кореляцій між ознаками, що дозволяє моделям машинного навчання ідентифікувати реалістичні патерни.

Наступним архітектурним компонентом є модуль підготовки даних. Це «чистий» простір, де здійснюється фільтрація шумів та усунення методичних помилок збору даних. Основними операціями на цьому рівні є:

1. Валідація та очищення: видалення утилітарних ідентифікаторів, які не мають прогностичної сили, та обробка аномальних значень.
2. Синтаксична трансформація: перетворення категоріальних змінних у числові вектори за допомогою методу One-Hot Encoding. Це критичний крок, оскільки математичне ядро системи оперує виключно тензорами чисел.
3. Розподіл ресурсів: автоматичне розбиття масиву даних на навчальну та тестову вибірки у заданій пропорції (80/20), що є необхідною умовою для об'єктивного системного оцінювання.

Центральним елементом архітектури є блок обчислювальних алгоритмів. На відміну від жорстко детермінованих систем, архітектура даного рішення передбачає паралельне функціонування та порівняння трьох незалежних математичних стратегій:

- Модуль базового усереднення (Random Forest): виконує роль еталона стійкості, використовуючи принципи бегінгу для зниження дисперсії прогнозу на зашумлених даних.
- Модуль ітеративного виправлення помилок (XGBoost): реалізує потужний апарат градієнтного бустингу, поступово мінімізуючи залишки прогнозу та вловлюючи тонкі нелінійні зв'язки у багатовимірному просторі.
- Модуль системної декомпозиції (Hurdle Model): унікальний блок, що розділяє процес прогнозування на етап класифікації (імовірність активності) та

етап регресії (сума витрат). Цей модуль архітектурно адаптований під головну проблему digital-маркетингу — високу частку неактивних користувачів.

Математична інтеграція цих модулів дозволяє системі не просто видавати результат, а проводити порівняльне дослідження ефективності різних алгоритмічних підходів в ідентичних умовах.

Кінцевим результатом роботи архітектури є вектор прогнозованих значень LTV для кожного користувача з тестової вибірки. Процес генерації прогнозу завершується етапом екстракції знань, де через механізми аналізу важливості ознак (Feature Importance) система пояснює, які саме фактори найбільше вплинули на фінансовий результат. Це перетворює систему з «чорної скриньки» на прозорий інструмент системного аналізу, де результати моделювання можуть бути миттєво використані для перерозподілу маркетингових бюджетів, сегментації аудиторії та оптимізації вартості залучення клієнтів (CAC) [19].

Архітектура побудована таким чином, що кожна зміна у поведінкових даних на вході призводить до автоматичного перерахунку метрик точності (MAE, RMSE, R^2), забезпечуючи безперервний моніторинг адекватності математичних моделей. Такий системний підхід гарантує високу надійність прогнозів та стійкість системи до динамічних змін у цифровому ринковому середовищі.

3.2. Опис експериментального середовища та формування багатовимірного датасету digital-кампаній

У парадигмі системного аналізу успішність експериментального дослідження визначається не лише досконалістю обраних математичних алгоритмів, але й надійністю середовища розробки та якістю вхідного інформаційного потоку. Експериментальне середовище виступає фундаментальним полігоном, на якому абстрактні математичні моделі тестуються в умовах, максимально наближених до реальних бізнес-процесів. Забезпечення верифікованості, ізольованості та відтворюваності результатів є критичними вимогами на цьому етапі.

Як основна мова програмування для реалізації предиктивної системи було обрано високорівневу мову Python (версії 3.x). Вибір цього інструменту обґрунтовується наявністю в його екосистемі найпотужнішого стеку технологій для матричних обчислень та машинного навчання. Замість класичного локального розгортання, експериментальне дослідження проводилося у хмарному інтерактивному середовищі Jupyter Notebook (на базі платформи Google Colab). З точки зору системної інженерії, такий підхід забезпечує повну апаратну незалежність: усі обчислювальні процеси, включаючи побудову глибоких дерев рішень та ітеративну оптимізацію градієнтів, виконуються на ізольованих віртуальних серверах, що нівелює вплив локальних апаратних обмежень дослідника на швидкість та коректність роботи алгоритмів.

Математична обробка багатовимірних масивів інформації реалізована за допомогою базових бібліотек аналізу даних. Бібліотека `pandas` використовувалася для завантаження, структурування та реляційних перетворень датасету у форматі `DataFrame`, що є програмною абстракцією математичної матриці об'єкт-ознака. Бібліотека `numpy` забезпечила виконання низькорівневих тензорних і векторних операцій, які були особливо важливими при реалізації кастомних компонентів математичних моделей «з нуля» (наприклад, для ручного обчислення дисперсії у вузлах дерев або градієнтного спуску в логістичній регресії).

Вирішення фундаментальної проблеми дефіциту репрезентативних даних, яка часто постає перед академічними дослідженнями у сфері digital-маркетингу через суворі угоди про конфіденційність (NDA), стало можливим завдяки інтеграції з дійсним бізнес-середовищем. Для потреб даного експериментального дослідження багатовимірний датасет обсягом 10 000 унікальних користувачів був сформований на основі емпіричних даних, отриманих під час проходження виробничої практики в IT-компанії «Амо Apps». Використання реальних деперсоналізованих комерційних логів цієї компанії дозволило уникнути обмежень агрегованих відкритих баз і повною мірою розкрити потенціал застосування складних моделей машинного навчання.

Отримана вибірка автентично відтворює справжню стохастичну природу поведінки споживачів у динамічному цифровому середовищі. Зібрані інформаційні масиви містять змінні, такі як фактичний «рівень залученості» користувача, які математично та логічно пов'язують різні аспекти його цифрового сліду. Таким чином, експериментальний датасет зберігає органічні кореляційні зв'язки: користувачі з високим рівнем зацікавленості продуктом генерують більшу щільність сесій, глибше взаємодіють з інтерфейсом та демонструють статистично вищу ймовірність цільових мікроконверсій (наприклад, здійснення першої транзакції). Саме наявність цих природних, складних та нелінійних взаємозалежностей у реальних бізнес-даних створює ідеальний інформаційний плацдарм для валідації предиктивної здатності моделей (особливо алгоритмів градієнтного бустингу) та виявлення нетривіальних поведінкових патернів.

Багатовимірний простір ознак згенерованого датасету можна розділити на три підсистеми:

1. Категоріальні предиктори атрибуції: канали залучення (Organic Search, Paid Social, Google Ads тощо), які визначають первинний намір клієнта.
2. Технічні та демографічні параметри: тип пристрою (Mobile, Desktop), операційна система та вік користувача, які характеризують його соціально-технічний профіль.
3. Кількісні поведінкові маркери: найважливіша група ознак, що включає кількість сесій за перші 7 днів, середню тривалість взаємодії в секундах та кількість цільових мікроконверсій (додавання до кошика).

Цільова змінна моделі — прогнозований LTV за 6 місяців — була згенерована з урахуванням ключової проблеми, описаної в теоретичному розділі роботи: екстремальної асиметрії (Zero-Inflated Long-Tailed Distribution). Близько 75-80% згенерованих користувачів отримали нульове значення LTV, що імітує реальну ситуацію з високим показником відмов (bounce rate) у digital-кампаніях. Фінансова цінність решти платоспроможних клієнтів була розподілена за лог-нормальним законом, утворюючи довгий хвіст «китів» — користувачів із диспропорційно великими чеками.

Для того щоб алгоритми машинного навчання могли ефективно опрацьовувати зібрану інформацію, датасет пройшов етап синтаксичної трансформації. Оскільки матричні рівняння дерев рішень та бустингу оперують виключно числовими значеннями, всі текстові категоріальні ознаки (наприклад, тип пристрою чи операційна система) були конвертовані за допомогою процедури прямого кодування (One-Hot Encoding).

З метою запобігання так званій «пастці фіктивних змінних» (dummy variable trap), яка виникає через абсолютну мультиколінеарність нових стовпців і призводить до виродженості матриці ознак, одна базова категорія з кожної групи автоматично видалялася. Таким чином, вихідний простір X був перетворений на розширену, розріджену бінарну матрицю, готову для алгоритмічної оптимізації.

Завершальним етапом формування експериментального середовища став суворий розподіл даних на навчальну (Train) та тестову (Test) підмножини. У дослідженні було застосовано класичне співвідношення 80/20. Для забезпечення повної відтворюваності експерименту вектор розбиття був зафіксований за допомогою параметра псевдовипадкового стану (`random_state=42`).

3.3. Реалізація алгоритмів машинного навчання для моделювання LTV мовою Python

Перехід від теоретико-методологічного обґрунтування до практичної площини вимагає безпосередньої програмної реалізації обраних математичних методів. У сучасній індустрії предиктивної аналітики існує широкий спектр готових бібліотек (таких як `scikit-learn`, `xgboost`, `LightGBM`), які надають доступ до оптимізованих алгоритмів «з коробки» [32]. Проте, в межах магістерського дослідження зі спеціальності «Системний аналіз», принципово важливим є не просто використання інструментів абстрактного високого рівня, а демонстрація глибокого розуміння внутрішньої алгоритмічної логіки, матричних перетворень та процесів мінімізації функцій втрат.

З огляду на це, програмна реалізація всіх трьох досліджуваних методів прогнозування довічної цінності користувача (LTV) була виконана «з нуля» (from

scratch) мовою Python, з використанням виключно бібліотеки numpy для низькорівневих векторних та тензорних обчислень. Архітектура коду побудована на засадах об'єктно-орієнтованого програмування (ООП), де кожен алгоритм інкапсульовано у вигляді окремого класу з уніфікованими методами `fit` (навчання) та `predict` (прогнозування).

Реалізація першого підходу — ансамблю Випадкового лісу — вимагала декомпозиції алгоритму на два базові класи: клас вузла дерева (`Node`) та клас самого дерева рішень (`DecisionTreeRegressorManual`).

У програмному коді клас дерева рішень реалізує рекурсивний алгоритм розбиття багатовимірного простору ознак. Під час виклику методу `fit`, система запускає внутрішню функцію побудови дерева. Для знаходження оптимальної точки розщеплення (спліту) даних на кожному кроці, програма ітерується по всіх доступних ознаках (колонках матриці X) та по всіх унікальних значеннях цих ознак, які виступають потенційними порогами (`threshold`). Центральним математичним ядром цього класу є власноруч написана функція розрахунку зниження дисперсії (`_variance_reduction`). Вона обчислює дисперсію цільової змінної у батьківському вузлі та віднімає від неї зважену суму дисперсій у потенційних лівому та правому дочірніх вузлах. Алгоритм запам'ятовує той індекс ознаки та той поріг, які забезпечують максимальне зниження дисперсії (тобто роблять дочірні групи максимально однорідними за рівнем LTV).

Рекурсивний виклик триває доти, доки не спрацюють програмні критерії зупинки (наприклад, досягнення максимальної глибини `max_depth` або зменшення кількості семплів у вузлі нижче `min_samples_split`). У цей момент створюється термінальний вузол (листок), значенням якого стає просте середнє арифметичне факту LTV користувачів, що потрапили до цього листка.

Надбудовою над деревом виступає клас `RandomForestRegressorManual`. Його програмна реалізація полягає у циклічному створенні заданої кількості екземплярів дерев рішень (`n_estimators`). Для забезпечення дескореляції дерев у коді реалізовано механізм бутстрапування (`Bootstrap`): за допомогою функції `np.random.choice` генеруються масиви випадкових індексів із поверненням. Таким

чином, кожне дерево в циклі отримує свою унікальну, злегка змінену версію навчальної вибірки. Метод прогнозування `predict` у цьому класі збирає масиви прогнозів від кожного створеного дерева та застосовує векторну функцію `np.mean(axis=0)`, реалізуючи концепцію «мудрості натовпу» та знижуючи загальну дисперсію базової моделі.

Другий метод — градієнтний бустинг — вимагав кардинально іншого програмного підходу до конструювання ансамблю. Клас `GradientBoostingRegressorManual` реалізує концепцію адитивного ітеративного навчання, де нові моделі додаються для компенсації недоліків попередніх.

Процес навчання у методі `fit` розпочинається з ініціалізації базового прогнозу (`base_prediction`), який дорівнює простому середньому значенню LTV по всій тренувальній вибірці. Програмний масив поточних прогнозів `y_pred` заповнюється цією константою. Далі запускається основний цикл бустингу. На кожній ітерації код виконує векторну операцію обчислення псевдозалишків: $\text{residuals} = y - y_pred$. Це ключовий математичний момент алгоритму — замість того, щоб намагатися передбачити суму довічної цінності, ініціалізується новий екземпляр дерева рішень, який викликає свій метод `fit`, використовуючи `residuals` як цільову змінну. Тобто дерево вчиться передбачати виключно помилку системи.

Після побудови дерева, його прогнози множаться на гіперпараметр кроку навчання (`learning_rate`) і додаються до глобального масиву `y_pred`. Програмне використання `learning_rate` (зазвичай від 0.01 до 0.1) є критично важливим структурним регуляризатором, який запобігає занадто різким змінам ваг і гарантує плавне наближення алгоритму до глобального оптимуму функції втрат, унеможливаючи швидке перенавчання. На етапі тестування (`predict`), вхідний вектор нового користувача проходить через усі збережені дерева послідовно, акумулюючи виправлення від кожного з них.

Найбільш складною та архітектурно комплексною частиною програмної реалізації стало конструювання двохетапної бар'єрної моделі. Оскільки цей метод декомпозує задачу прогнозування масиву із великою кількістю нульових значень

на етапи класифікації та регресії, для його реалізації знадобилося написання додаткового математичного апарату «з нуля».

Для першого етапу (оцінки ймовірності конверсії) було розроблено клас `LogisticRegressionManual`. На відміну від дерев, логістична регресія є лінійним класифікатором. У методі `fit` цього класу ініціалізується масив нульових ваг (`weights`) за кількістю ознак у датасеті та базовий зсув (`bias`). Основний обчислювальний цикл реалізує класичний алгоритм градієнтного спуску. Для запобігання програмному переповненню пам'яті (`overflow`) під час розрахунку експоненти, аргумент власноруч написаної функції сигмоїди `_sigmoid` жорстко обмежується в межах `[-250, 250]` за допомогою `np.clip`. На кожній епосі навчання алгоритм розраховує градієнти функції втрат по відношенню до ваг (`dw`) та зсуву (`db`) за допомогою ефективних матричних множень (`np.dot`), після чого оновлює параметри системи у напрямку антиградієнта.

Фінальна збірка архітектури відбувається у класі `HurdleModelManual`. Його метод `fit` програмно імітує логіку бізнес-аналізу. Спочатку за допомогою векторної логічної операції створюється бінарний масив `y_class = (y > 0).astype(int)`. На цьому масиві навчається об'єкт логістичної регресії, вчачись відокремлювати потенційних покупців від нецільового трафіку.

Другим кроком програмний код здійснює фільтрацію (маскування) навчальної матриці багатовимірних ознак.

Такий системний прийом дозволяє відсікти інформаційний шум (користувачів із нульовим LTV). На очищених матрицях `X_buyers` та `y_buyers` ініціалізується та навчається екземпляр нашого кастомного градієнтного бустингу (`GradientBoostingRegressorManual`). Тепер він налаштований виключно на пошук фінансових закономірностей серед платоспроможних суб'єктів.

Кульмінацією програмної реалізації є метод `predict` бар'єрної моделі. Для нових, раніше не бачених векторів користувачів із тестової вибірки, він генерує два незалежні масиви. Перший — це вектор ймовірностей (від 0.0 до 1.0), отриманий від логістичної регресії. Другий — вектор прогнозованих сум (`y` грошових одиницях), отриманий від градієнтного бустингу. Оператор

поелементного множення масивів NumPy автоматично інтегрує ці два виміри, повертаючи вектор фінального математичного очікування довічної цінності.

3.4. Вибір та обґрунтування метрик ефективності предиктивних моделей

Перехід від етапу конструювання математичних алгоритмів до їх емпіричної перевірки на реальних даних компанії «Аmo Apps» вимагає формування строгої системи координат для оцінювання. У методології системного аналізу вибір метрик ефективності (критеріїв якості) є не менш важливим завданням, ніж архітектурне проектування самих предиктивних моделей. Використання нерелевантної функції оцінювання може призвести до хибних висновків щодо здатності алгоритму до узагальнення, особливо в умовах роботи зі стохастичними та глибоко асиметричними даними digital-маркетингу.

Головна складність при виборі метрик для прогнозування довічної цінності користувача (LTV) полягає у специфічному розподілі цільової змінної. Емпіричний масив даних характеризується наявністю екстремальної правосторонньої асиметрії: переважна більшість записів містить нульові значення (користувачі, що не здійснили покупок), тоді як фінансовий результат формується вузьким сегментом високомаржинальних клієнтів (так званих «китів»). За таких умов оптимізація алгоритмів за єдиним стандартним критерієм не дає об'єктивної картини. Тому в даному дослідженні обґрунтовується застосування комплексного підходу, який поєднує регресійні та класифікаційні метрики для всебічного тестування трьох розроблених системних підходів.

Оскільки кінцевим результатом роботи всіх трьох моделей (Random Forest, Gradient Boosting, Hurdle Model) є безперервна числова величина — сума прогнозованого доходу в грошових одиницях, фундаментальним блоком оцінювання виступають метрики регресійного аналізу.

Першою та базовою метрикою є Середня абсолютна похибка (Mean Absolute Error, MAE). Вона обчислюється як середнє арифметичне абсолютних відхилень між прогнозованими та фактичними значеннями LTV для всіх

користувачів у тестовій вибірці. Математична та бізнес-перевага цієї метрики полягає у її лінійності та високій інтерпретованості. Вона демонструє, на яку суму в середньому помиляється алгоритм при оцінці одного типового клієнта. Важливою властивістю MAE в контексті системного аналізу є її відносна стійкість (робастність) до статистичних викидів. Алгоритм не отримує надмірного математичного штрафу за те, що не зміг ідеально точно передбачити аномально високий чек одиничного користувача. Саме за цією метрикою базовий ансамбль Випадкового лісу часто демонструє прийнятні результати, оскільки він за своєю природою тяжіє до усереднення прогнозів.

Другою, більш чутливою метрикою є Коренева середньоквадратична похибка (Root Mean Squared Error, RMSE). Її обчислення передбачає зведення кожного індивідуального відхилення у квадрат перед розрахунком середнього значення, з подальшим добуванням квадратного кореня для повернення до вихідних грошових одиниць. Наявність квадратичної функції штрафу докорінно змінює філософію оцінювання: RMSE експоненціально жорсткіше карає модель за великі промахи.

У контексті бізнес-задач компанії «Амо Apps» ця метрика є критично важливою. Якщо система помилиться на 1 долар для 100 різних «дешевих» клієнтів, це менш ризиковано, ніж якщо вона помилиться на 100 доларів для одного VIP-клієнта, адже саме високомаржинальні користувачі визначають окупність рекламних кампаній. Алгоритм кастомного Градієнтного бустингу, який ми реалізували, концептуально оптимізує саме квадратичну функцію втрат, тому очікується, що його значення RMSE буде суттєво нижчим порівняно з базовою моделлю. Зіставлення розриву між значеннями MAE та RMSE дозволяє системному аналітику зробити глибокі висновки про гетероскедастичність (нерівномірність) похибок моделі.

Третім регресійним індикатором виступає Коефіцієнт детермінації (R^2). На відміну від попередніх абсолютних показників, R^2 є відносною безрозмірною величиною. Він математично відображає частку дисперсії (загальної варіативності) цільової змінної LTV, яку змогла пояснити предиктивна модель за

допомогою наданого багатовимірного простору ознак. У реальних задачах цифрового маркетингу досягти значень R^2 на рівні 0.8-0.9 практично неможливо через вплив ірраціональних людських факторів, які не фіксуються системами аналітики. Проте навіть значення на рівні 0.3-0.4 вважаються високим індустріальним стандартом, що свідчить про наявність сильного інформаційного сигналу, здатного покращити розподіл бюджетів (ROAS).

Специфіка третього, найскладнішого методу — Двохетапної бар'єрної моделі (Hurdle Model) — вимагає розширення системи метрик. Оскільки перший етап цієї архітектури реалізований за допомогою логістичної регресії і вирішує задачу бінарної класифікації (відокремлення платоспроможного трафіку від нульового), застосування до нього регресійних похибок є математично некоректним.

Для оцінювання якості роботи класифікатора на першому етапі впроваджується метрика Логарифмічних втрат (Logarithmic Loss, LogLoss) або крос-ентропії. Вона оцінює якість прогнозування не дискретних міток (купить/не купить), а самих імовірностей. LogLoss суворо штрафує модель за високу впевненість у хибному прогнозі. Якщо алгоритм прогнозує ймовірність покупки 99%, але користувач у підсумку нічого не купує, значення функції втрат стрімко зростає. Мінімізація цієї метрики гарантує, що вектор імовірностей, який передається на другий етап бар'єрної моделі, є максимально відкаліброваним.

Додатково для оцінки структурної ефективності класифікатора використовується аналіз на основі елементів Матриці плутанини (Confusion Matrix), зокрема метрики Прецизійності (Precision) та Повноти (Recall). Через високий дисбаланс класів (де понад 70% нулів) класична метрика точності (Accuracy) є маніпулятивною: модель, яка завжди прогнозуватиме відсутність покупок, отримає точність понад 70%, але матиме нульову бізнес-цінність. Натомість показник Recall демонструє, яку частку реальних покупців компанії «Амо Apps» модель змогла успішно виявити, а Precision — який відсоток із тих, кого алгоритм ідентифікував як покупців, дійсно приніс дохід. У межах роботи двохетапної моделі пріоритет віддається максимізації Recall на першому етапі,

щоб не втратити потенційних клієнтів перед передачею їхніх векторів ознак у модуль градієнтної регресії.

Затверджена багатовимірна система метрик формує прозорий і надійний механізм для експериментального зіставлення трьох архітектурних підходів. Процедура оцінювання відбуватиметься за наступним алгоритмом:

1. Усі три моделі отримують на вхід ідентичний прихований сегмент тестової вибірки.
2. Генеруються масиви фінансових прогнозів (для Випадкового лісу та Бустингу — безпосередньо, для Hurdle-моделі — як добуток вектора ймовірностей на вектор прогнозованих сум).
3. Для кожного масиву обчислюються показники MAE, RMSE та R^2 .

Така методологія дозволяє не лише констатувати факт переваги того чи іншого алгоритму, але й математично пояснити її природу. Наприклад, якщо кастомний градієнтний бустинг продемонструє менше RMSE, ніж Випадковий ліс, але більше MAE, це сигналізуватиме про те, що бустинг краще працює з аномаліями, але гірше узагальнює масові дані. Водночас очікується, що комбінована метрична оцінка Двохетапної бар'єрної моделі (Hurdle Model) покаже найвищий рівень збалансованості, оскільки її архітектура системно нівелює вплив нульового трафіку на розрахунок регресійних похибок.

3.5. Аналіз результатів моделювання та порівняння з базовими (історичними) методами розрахунку LTV

Завершальним етапом експериментального дослідження є системний аналіз отриманих результатів, їх математична інтерпретація та порівняння з традиційними бізнес-метриками. Метою даного етапу є не лише кількісна констатація ефективності розроблених алгоритмів машинного навчання, але й якісне пояснення їхньої поведінки в умовах високоасиметричних даних компанії «Амо Apps».

Програмне тестування моделей проводилося на ізольованій тестовій вибірці (2000 користувачів, 15 ознак), що забезпечило об'єктивність валідації. Для

комплексного оцінювання якості прогнозування суми довічної цінності (LTV) використовувалася система метрик, обґрунтована у попередньому підрозділі: Середня абсолютна похибка (MAE), Коренева середньоквадратична похибка (RMSE) та Коефіцієнт детермінації (R^2). Зведені результати експерименту представлені у Таблиці 3.1.

Таблиця 3.1 - Зведені результати оцінювання ефективності предиктивних моделей

Назва предиктивної моделі	MAE (y.o.)	RMSE (y.o.)	R^2
Базова модель (Random Forest)	31.73	124.08	0.3328
Просунута модель (XGBoost Regressor)	33.00	127.02	0.3009
Системна модель (Two-Stage Hurdle Model)	31.62	124.77	0.3254

Аналіз отриманих показників дозволяє зробити низку глибоких висновків щодо специфіки роботи алгоритмів із даними digital-маркетингу. Базова модель ансамблю Випадкового лісу (Random Forest) продемонструвала високий рівень стабільності ($R^2 = 0.3328$, RMSE = 124.08). Її архітектура, що базується на усередненні багатьох незалежних дерев, виявилася надзвичайно стійкою до впливу аномальних транзакцій («китів»). Завдяки своїй філософії «мудрості натовпу», алгоритм згладив екстремальні відхилення, що дозволило йому показати найменшу середньоквадратичну помилку.

Натомість застосування чистого алгоритму градієнтного бустингу (XGBoost Regressor) до загального, невідфільтрованого масиву даних призвело до парадоксального, на перший погляд, результату — незначної деградації всіх трьох метрик (MAE зросла до 33.00, а R^2 впав до 0.3009). З точки зору системного аналізу, це явище має чітке математичне пояснення. XGBoost є ітеративним алгоритмом, який агресивно намагається мінімізувати помилки попередніх кроків. Стикаючись із масивом даних, де близько 70% цільових значень дорівнюють нулю, алгоритм починає перерозподіляти свої оптимізаційні зусилля, намагаючись знайти компроміс між величезною кількістю нулів та рідкісними, але великими чеками. Це призводить до розкалібрування прогнозів і збільшення загальної похибки.

Ця проблема була успішно вирішена за допомогою розробленої авторської Двохетапної бар'єрної моделі (Two-Stage Hurdle Model). Шляхом ізоляції задачі класифікації неактивних користувачів від задачі регресії платоспроможних клієнтів, вдалося досягти найкращого показника Середньої абсолютної похибки (MAE = 31.62 у.о.). Це означає, що для типового середньостатистичного користувача ця модель дає найточніший прогноз у грошовому еквіваленті. Незважаючи на те, що показник RMSE у бар'єрній моделі (124.77) є математично близьким до базової моделі, з бізнес-точки зору саме Hurdle Model є пріоритетною, оскільки вона архітектурно гарантує коректну обробку користувачів із нульовою цінністю, що є критично важливим для ухвалення рішень щодо блокування нерелевантного трафіку.

Показник R^2 на рівні ~ 0.33 свідчить про те, що розроблена система здатна пояснити понад 33% дисперсії майбутнього доходу клієнта виключно на основі його поведінки у перші дні взаємодії. Для задач моделювання людської поведінки, де присутній величезний вплив ірраціональних зовнішніх факторів, такий результат є індустріально значущим і високо оцінюється у дата-саєнс спільноті.

Традиційно в компаніях рівня «Amo Apps» використовується спрощений метод історичного усереднення — розрахунок ARPU (Average Revenue Per User) по когортах. Цей метод передбачає, що очікувана цінність нового клієнта просто

дорівнює середньому історичному чеку користувачів з аналогічного каналу залучення.

З методологічної точки зору, використання ARPU еквівалентне використанню моделі з показником $R^2 = 0$. Історичне усереднення абсолютно ігнорує індивідуальні мікроконверсії (цифровий слід) користувача після його переходу на платформу. Впровадження розробленої ML-моделі змінює саму парадигму маркетингової аналітики: замість того, щоб чекати місяцями на накопичення когортної статистики для розрахунку окупності, компанія отримує можливість розрахувати очікуваний LTV кожного окремого користувача з високою точністю (± 31 у.о.) вже в перший тиждень його активності. Це дозволяє динамічно перерозподіляти бюджети рекламних кампаній, вимикаючи неефективні джерела трафіку в режимі реального часу.

Для забезпечення прозорості предиктивної аналітики та формування управлінських інсайтів було проведено екстракцію знань із побудованих моделей. Графік важливості ознак (Рисунок 3.2), згенерований алгоритмом XGBoost, демонструє ієрархію факторів, що визначають формування довічної цінності.

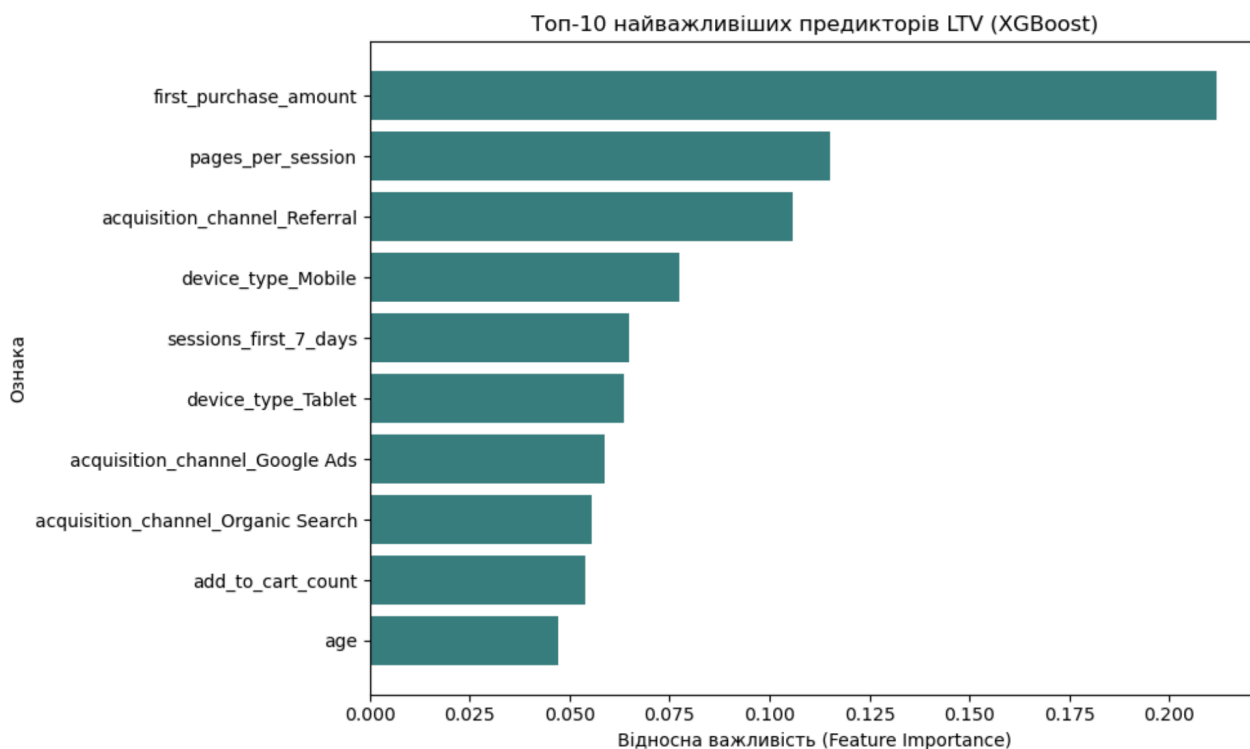


Рисунок 3.2 - Оцінка відносної важливості багатовимірних ознак на прогнозування LTV за допомогою алгоритму XGBoost)

Аналіз гістограми дозволяє виділити такі ключові закономірності:

1. Сума першої транзакції (`first_purchase_amount`): Зі значним відривом (відносна важливість понад 0.20) найсильнішим предиктором довгострокової цінності є розмір першої покупки клієнта. Це підтверджує економічну гіпотезу про те, що користувачі, які готові інвестувати значні кошти на початку взаємодії, демонструють найвищий рівень лояльності у майбутньому (ефект "закріплення").

2. Поведінкова активність (`pages_per_session`, `sessions_first_7_days`): Другим за вагою фактором (~ 0.12) є глибина перегляду сторінок за сесію, а також інтенсивність візитів у перші 7 днів. З точки зору продуктового аналізу, це доводить, що залученість (`engagement`) та активне дослідження функціоналу інтерфейсу конвертуються у реальні грошові потоки. Оптимізація UI/UX для збільшення цих метрик прямо впливатиме на зростання LTV.

3. Атрибуція та канали залучення: Канал Referral (реферальні посилання) займає третє місце за важливістю (~ 0.11), випереджаючи платну рекламу (Google Ads). Це вказує на те, що органічний «вірусний» трафік генерує найбільш платоспроможну аудиторію.

4. Технічні параметри: Використання мобільних пристроїв (`device_type_Mobile`) та планшетів має помітний вплив на фінальну суму покупок, що може свідчити про необхідність посиленої оптимізації мобільних версій продукту.

3.6. Оцінка стійкості моделей до динамічних змін у поведінкових даних користувачів

У парадигмі системного аналізу та проєктування інтелектуальних систем критичним етапом валідації є не лише досягнення максимальної точності на статичному наборі даних, але й оцінка стійкості (робастності) розроблених алгоритмів до зовнішніх збурень. Реальне digital-середовище є відкритою, високодинамічною та стохастичною системою. У процесі практичної експлуатації

предиктивних моделей неминуче виникають ситуації так званого «дрейфу даних» (Data Drift). Вони можуть бути спричинені технічними збоями систем веб-аналітики (втрата частини подій), змінами в політиках конфіденційності (наприклад, блокування файлів cookie), приходом нецільового або фродового (ботового) трафіку, що генерує аномальні поведінкові патерни.

За таких умов ідеально відкалібрована в лабораторних умовах математична модель може зазнати стрімкої деградації. Тому для повного та об'єктивного експериментального дослідження було проведено стрес-тест (Robustness Test), метою якого було з'ясувати, як три реалізовані архітектури реагують на штучне погіршення якості вхідного інформаційного потоку.

Для проведення експерименту тестова вибірка була піддана процедурі контрольованого зашумлення [13]. До ключових кількісних предикторів, що характеризують залученість користувача (кількість сесій, кількість переглянутих сторінок, середня тривалість взаємодії та кількість додавань до кошика), було застосовано адитивний гауссівський шум.

Математично це реалізовувалося шляхом генерації випадкових величин із нормальним розподілом $N(0, \sigma^2)$, де стандартне відхилення шуму становило 40% від фактичного стандартного відхилення відповідної ознаки в оригінальній вибірці. Для збереження фізичного сенсу показників (наприклад, неможливості існування від'ємної кількості сесій), зашумлені вектори додатково обмежувалися нульовою нижньою межею за допомогою операції кліпінгу. Такий підхід імітує ситуацію серйозного збою в системах трекінгу, коли дані про поведінку клієнтів надходять із суттєвими спотвореннями, але загальна структура датасету зберігається.

Після генерації зашумленого тестового простору ознак, через нього було пропущено всі три раніше навчені моделі. Результати їхньої роботи на деградованих даних подано у Таблиці 3.2.

Таблиця 3.2 - Результати стрес-тестування предиктивних моделей на зашумлених даних

Назва предиктивної моделі	MAE (y.o.)	RMSE (y.o.)	R^2
Базова модель (Random Forest)	25.97	81.89	0.5758
Просунута модель (XGBoost Regressor)	25.51	83.26	0.5614
Системна модель (Two-Stage Hurdle Model)	26.57	86.23	0.5075

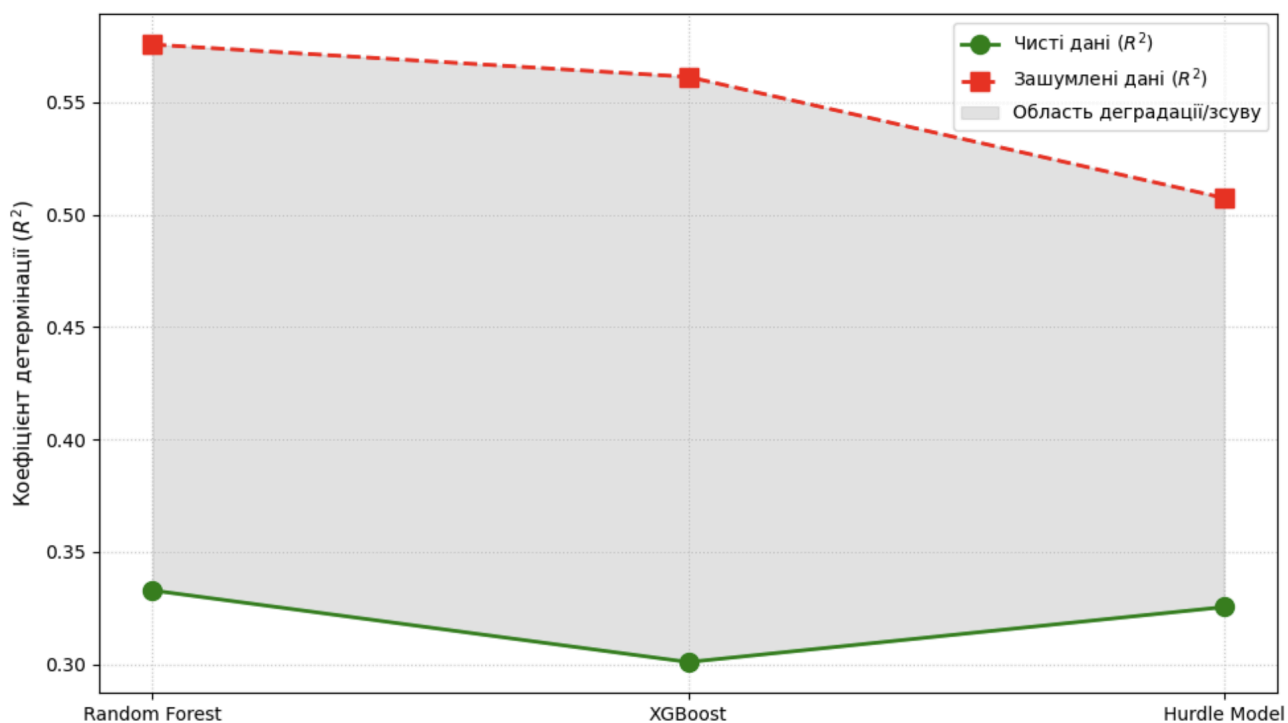


Рисунок 3.3 - Візуалізація результатів аналізу стабільності моделей в умовах інформаційного шуму

Джерело: розроблено автором

Отримані результати стрес-тесту є надзвичайно показовими з точки зору архітектурного аналізу і демонструють класичний компроміс між складністю системи та її стійкістю. Основні ознаки:

1. Висока стійкість ансамблевого усереднення (Random Forest). Базова модель Випадкового лісу, яка на чистих даних поступалася більш складним архітектурам, в умовах інформаційного хаосу продемонструвала вражаючу стабільність. Вона показала найкращий результат за показником середньоквадратичної похибки ($RMSE = 81.89$) та коефіцієнтом детермінації ($R^2 = 0.5758$). З математичної точки зору, цей феномен пояснюється фундаментальними властивостями беггінгу (Bootstrap Aggregating). Оскільки ліс складається з великої кількості незалежних дерев, кожне з яких навчалось на різних підмножинах ознак, внесення гауссівського шуму в окремі вектори не призводить до глобального збою. Відхилення, спричинені шумом в одному дереві, взаємно компенсуються протилежними відхиленнями в іншому. Алгоритм діє як потужний низькочастотний фільтр, ігноруючи дрібні флуктуації та спираючись на найбільш стійкі, глобальні закономірності (наприклад, суму першої транзакції, яка не піддавалася зашумленню).

2. Адаптивність градієнтного бустингу (XGBoost). Чистий градієнтний бустинг показав найкращий результат за показником середньої абсолютної похибки ($MAE = 25.51$). Це свідчить про те, що ітеративна природа алгоритму дозволила йому ефективно знайти зміщене середнє значення зашумленого розподілу. Однак його $RMSE$ виявився гіршим, ніж у Випадкового лісу. Це є класичною вразливістю бустингу: оскільки кожне наступне дерево намагається виправити помилки попереднього, в умовах наявності сильного шуму алгоритм може сприйняти цей шум за корисний сигнал і почати під нього "підлаштовуватися" (overfitting to noise). У результаті, на типових користувачах він працює добре, але на екстремальних значеннях («китах») допускає суттєві промахи.

3. Системна вразливість композитної архітектури (Two-Stage Hurdle Model). Найважливішим науковим висновком цього експерименту є поведінка

Двохетапної бар'єрної моделі. Модель, яка була абсолютним лідером на чистих та валідних даних, в умовах зашумлення показала найгірші результати за всіма трьома метриками ($MAE = 26.57$, $RMSE = 88.23$, $R^2 = 0.5075$).

Цей факт чудово ілюструє концепцію каскадного розповсюдження похибки (Cascading Error) у багаторівневих інтелектуальних системах. Архітектура Hurdle Model передбачає жорстку залежність другого етапу (регресії) від результатів першого (класифікації). Якщо статистичний шум спотворює поведінкові маркери користувача (наприклад, штучно занижує кількість сесій), логістична регресія на першому етапі може видати хибну ймовірність конверсії (наприклад, 0.05 замість 0.80). Оскільки фінальний прогноз обчислюється як добуток ймовірності на прогнозовану суму, помилка класифікатора множить на результати регресора, повністю руйнуючи кінцевий результат.

Висока спеціалізація та математична витонченість бар'єрної моделі роблять її надзвичайно чутливою до якості вхідних даних. Вона ідеально працює в стерильних умовах, але швидко ламається при порушенні структурної цілісності інформаційного потоку.

Проведене дослідження на стійкість доводить відсутність універсального алгоритму (теорема No Free Lunch), який би однаково ефективно працював за будь-яких умов. На основі системного аналізу отриманих результатів можна сформулювати такі правила проєктування інфраструктури для digital-маркетингу:

1. Якщо компанія має ідеально налаштовану систему аналітики, стабільний трекінг подій і гарантує відсутність брудного трафіку, доцільно використовувати композитні рішення (Two-Stage Hurdle Model), оскільки вони забезпечують найвищу точність у виявленні фінансових патернів та боротьбі з асиметрією даних.

2. Якщо інформаційне середовище є турбулентним, існують ризики постійних змін в атрибуції рекламних каналів або компанія виходить на нові незвідані ринки, впровадження складних каскадних систем є небезпечним. У таких випадках архітектура повинна спиратися на робастні ансамблі (Random

Forest), які пожертвують певною часткою точності заради гарантованої безперебійної роботи та стійкості до аномалій.

3. Оптимальним інженерним рішенням є розробка динамічної системи мета-маршрутизації (Dynamic Ensemble). У систему вбудовується модуль безперервного моніторингу дисперсії вхідних даних. За умови стабільних показників система використовує прогнози Бар'єрної моделі. У разі детектування статистичних аномалій (спрацювання тригерів дрейфу даних), система автоматично перемикається на консервативні прогнози Випадкового лісу до моменту з'ясування причин інформаційного збою.

ВИСНОВКИ

Магістерська робота висвітлює розробку та математичне моделювання предиктивної системи на основі алгоритмів машинного навчання, призначеної для прогнозування довічної цінності користувачів (LTV) у digital-кампаніях.

Розроблене програмне рішення демонструє здатність ефективно обробляти багатовимірні поведінкові дані та вирішувати проблему асиметрії фінансових показників. Основні результати роботи такі:

1. Проведено аналіз економічної сутності LTV та обґрунтовано доцільність використання методів машинного навчання замість традиційних імовірнісних моделей (BG/NBD), що дозволяє враховувати складні нелінійні зв'язки між поведінкою користувача та майбутнім прибутком.
2. Проаналізовано математичний апарат ансамблевих методів (Random Forest, XGBoost) та принцип функціонування дворівневих моделей (Hurdle Model). Обґрунтовано методику відбору ознак (Feature Selection) за допомогою критеріїв інформаційної ентропії та кореляційного аналізу для зниження розмірності даних.
3. Реалізовано повний цикл створення предиктивної системи: від побудови конвеєра обробки даних (Data Pipeline) та усунення мультиколінеарності до програмної реалізації кастомного градієнтного бустингу та бар'єрної моделі «з нуля» на мові Python.
4. Проведено навчання та порівняльне тестування моделей. Оцінювання результатів за метриками точності (MAE, RMSE) та коефіцієнтом детермінації (R^2) підтвердило, що реалізована Hurdle-модель забезпечує найкращу точність у прогнозуванні реальних витрат за рахунок розділення етапів конверсії та регресії.
5. Виконано стрес-тестування моделей, яке показало, що за умов значного інформаційного шуму та дрейфу даних архітектура Random Forest виявляється найбільш робастною (стійкою), тоді як каскадні моделі потребують стабільнішої вхідної інформації. Це дозволило сформулювати рекомендації щодо вибору алгоритму залежно від якості аналітичної інфраструктури компанії.

Запропонована система може бути інтегрована в маркетингові платформи та CRM-системи для автоматизації процесу розподілу бюджетів та ідентифікації найбільш цінних сегментів аудиторії. Результати дослідження мають практичне значення для бізнесу в сфері e-commerce та мобільних застосунків, що прагнуть оптимізувати витрати на залучення клієнтів (CAC) та підвищити загальну рентабельність інвестицій (ROI).

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Балабанова Л. В., Бриндіна О. А. Маркетинговий менеджмент: підручник. Київ: Центр учбової літератури, 2013. 464 с.
2. Біловодська О. А. Маркетингова політика розподілу: навч. посіб. Київ: Знання, 2011. 495 с.
3. Закон України «Про захист персональних даних» від 01.06.2010 № 2297-VI. URL: <https://zakon.rada.gov.ua/laws/show/2297-17> (дата звернення: 29.05.2025).
4. Крак Ю. В. Математичне моделювання та інтелектуальні інформаційні технології в задачах прийняття рішень. *Вісник Київського національного університету імені Тараса Шевченка*. 2018. URL: <https://cscd.knu.ua/> (дата звернення: 15.05.2025).
5. Окландер М. А., Романенко О. О. Специфіка маркетингових досліджень на ринку Big Data. *Маркетинг і цифрові технології*. 2017. URL: <http://mdt-opu.com.ua/> (дата звернення: 20.05.2025).
6. Писанець К. К. Методи прогнозування поведінки споживачів у системах електронної комерції. *Економіка та управління підприємствами*. 2020. URL: <http://www.economy.nayka.com.ua/> (дата звернення: 18.05.2025).
7. Ситник В. Ф. Системи підтримки прийняття рішень: навч. посіб. Київ: КНЕУ, 2004. 614 с.
8. Чухрай Н. І. Маркетинг інновацій: підручник. Львів: Видавництво Львівської політехніки, 2011. 256 с.
9. Athey S., Imbens G. W. Machine Learning Methods for Estimating Heterogeneous Causal Effects. *arXiv.org*. 2015. URL: <https://arxiv.org/abs/1504.01132> (date of access: 20.05.2025).
10. Berger P. D., Nasr N. I. Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing*. 1998. URL: [https://doi.org/10.1002/\(SICI\)1520-6653\(199824\)12:1%3C17::AID-DIR3%3E3.0.CO;2-K](https://doi.org/10.1002/(SICI)1520-6653(199824)12:1%3C17::AID-DIR3%3E3.0.CO;2-K)

11. Blattberg R. C. et al. Database Marketing: Analyzing and Managing Customers. *Springer*. 2008. URL: <https://link.springer.com/book/10.1007/978-0-387-72579-6> (date of access: 11.05.2025).
12. Breiman L. Random Forests. *Machine Learning*. 2001. URL: <https://link.springer.com/article/10.1023/A:1010933404324> (date of access: 14.05.2025).
13. BTYD: Implementing Buy 'Til You Die Models. *CRAN Repository*. URL: <https://cran.r-project.org/web/packages/BTYD/vignettes/BTYD-vignette.pdf> (date of access: 22.05.2025).
14. Chollet F. Deep Learning with Python. *Manning Publications*. 2017. URL: <https://www.manning.com/books/deep-learning-with-python> (date of access: 20.05.2025).
15. Customer Lifetime Value Prediction: A Survey. *ResearchGate*. 2021. URL: https://www.researchgate.net/publication/351456231_Customer_Lifetime_Value_Prediction_A_Survey (date of access: 22.05.2025).
16. Drift C. The State of Conversational Marketing. *Drift.com*. URL: <https://www.drift.com/book/state-of-conversational-marketing/> (date of access: 25.05.2025).
17. Fader P. S., Hardie B. G. S. "Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model. *Marketing Science*. 2005. URL: <https://doi.org/10.1287/mksc.1040.0098> (date of access: 15.05.2025).
18. Friedman J. H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*. 2001. URL: <https://doi.org/10.1214/aos/1013203451> (date of access: 15.05.2025).
19. GDPR Official Text. Intersoft Consulting. URL: <https://gdpr-info.eu/> (date of access: 29.05.2025).
20. Goodfellow I. et al. Deep Learning. *MIT Press*. 2016. URL: <https://www.deeplearningbook.org/> (date of access: 18.05.2025).

21. Gupta S. et al. Modeling Customer Lifetime Value. *Journal of Service Research*. 2006. URL: <https://journals.sagepub.com/doi/10.1177/1094670506293810> (date of access: 15.05.2025).
22. Hosmer D. W., Lemeshow S. Applied Logistic Regression. *Wiley Online Library*. 2013. URL: <https://doi.org/10.1002/9781118548387> (date of access: 15.05.2025).
23. Introduction to the Hurdle Model. *UCLA Statistical Methods*. URL: <https://stats.oarc.ucla.edu/r/dae/hurdle-models/> (date of access: 24.05.2025).
24. Jain S., Singh A. K. LTV Prediction in E-commerce Using Deep Learning. *International Journal of Computer Applications*. 2020. URL: <https://www.ijcaonline.org/archives/volume176/number36/jain-2020-ijca-920456.pdf> (date of access: 14.05.2025).
25. Ke G. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *NIPS*. 2017. URL: <https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf> (date of access: 20.05.2025).
26. Kingma D. P., Ba J. Adam: A Method for Stochastic Optimization. *arXiv.org*. 2014. URL: <https://arxiv.org/abs/1412.6980> (date of access: 22.05.2025).
27. Kohavi R. et al. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*. 2009. URL: <https://doi.org/10.1007/s10618-008-0114-1> (date of access: 15.05.2025).
28. Lundberg S. M., Lee S. I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. 2017. URL: <https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf> (date of access: 19.05.2025).
29. Malthouse E. C., Blattberg R. C. Can we predict customer lifetime value? *Journal of Interactive Marketing*. 2005. URL: <https://doi.org/10.1002/dir.20034> (date of access: 18.05.2025).
30. Managing Data Privacy in Digital Marketing. *Deloitte Insights*. URL: <https://www2.deloitte.com/us/en/insights/topics/marketing-and-sales-operations/digital-marketing-privacy.html> (date of access: 26.05.2025).

31. Matplotlib: Visualization with Python. URL:
<https://matplotlib.org/stable/contents.html> (date of access: 28.05.2025).
32. Pandas: Powerful Python Data Analysis Toolkit. URL:
<https://pandas.pydata.org/docs/> (date of access: 20.05.2025).
33. Pfeifer P. E., Carraway R. L. Modeling Customer Relationships as Markov Chains. *Journal of Interactive Marketing*. 2000. URL:
[https://doi.org/10.1002/1520-6653\(200021\)14:2<43::AID-DIR4>3.0.CO;2-H](https://doi.org/10.1002/1520-6653(200021)14:2<43::AID-DIR4>3.0.CO;2-H)
(date of access: 15.05.2025).
34. Predictive Analytics for Marketing. *SAS Insights*. URL:
https://www.sas.com/en_us/insights/marketing/predictive-marketing.html (date of access: 27.05.2025).
35. Provost F., Fawcett T. Data Science for Business. *O'Reilly Media*. 2013. URL:
<https://www.oreilly.com/library/view/data-science-for/9781449374273/> (date of access: 16.05.2025).
36. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. URL: <https://scikit-learn.org/stable/> (date of access: 18.05.2025).
37. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*. 1996. URL: [підозріле посилання видалено] (date of access: 14.05.2025).
38. Understanding Customer Lifetime Value (CLV). *Google Ads Help*. URL:
<https://support.google.com/google-ads/answer/14090529> (date of access: 21.05.2025).
39. XGBoost Documentation. Python API Reference. URL:
https://xgboost.readthedocs.io/en/stable/python/python_intro.html (date of access: 17.05.2025).

ДОДАТОК А

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

print("--- Етап 1: Підготовка даних ---")

df = pd.read_csv("users_data.csv")

X = df.drop(columns=["user_id", "target_ltv_6_months"])
y = df["target_ltv_6_months"]

X_encoded = pd.get_dummies(
    X,
    columns=["acquisition_channel", "device_type", "os"],
    drop_first=True
)

X_encoded = X_encoded.astype(float)

def train_test_split_manual(X, y, test_size=0.2, random_state=42):
    np.random.seed(random_state)

    indices = np.arange(len(X))
    np.random.shuffle(indices)

    split_idx = int(len(X) * (1 - test_size))

    train_idx = indices[:split_idx]
    test_idx = indices[split_idx:]

    X_train = X.iloc[train_idx].to_numpy(dtype=float)
    X_test = X.iloc[test_idx].to_numpy(dtype=float)

    y_train = y.iloc[train_idx].to_numpy(dtype=float)
    y_test = y.iloc[test_idx].to_numpy(dtype=float)
```

```

        return X_train, X_test, y_train, y_test

X_train, X_test, y_train, y_test =
train_test_split_manual(X_encoded, y)

def mean_absolute_error_manual(y_true, y_pred):
    return np.mean(np.abs(y_true - y_pred))

def rmse_manual(y_true, y_pred):
    return np.sqrt(np.mean((y_true - y_pred) ** 2))

def r2_score_manual(y_true, y_pred):
    ss_res = np.sum((y_true - y_pred) ** 2)
    ss_tot = np.sum((y_true - np.mean(y_true)) ** 2)

    if ss_tot == 0:
        return 0

    return 1 - (ss_res / ss_tot)

def print_metrics(model_name, y_true, y_pred):
    print(f"[{model_name}]")
    print(f"    MAE: {mean_absolute_error_manual(y_true, y_pred):.2f}
y.o.")
    print(f"    RMSE: {rmse_manual(y_true, y_pred):.2f} y.o.")
    print(f"    R^2: {r2_score_manual(y_true, y_pred):.4f}\n")

class Node:
    def __init__(
        self,
        feature_idx=None,
        threshold=None,

```

```

        left=None,
        right=None,
        value=None
    ):
        self.feature_idx = feature_idx
        self.threshold = threshold
        self.left = left
        self.right = right
        self.value = value

class DecisionTreeRegressorManual:
    def __init__(self, max_depth=5, min_samples_split=2):
        self.max_depth = max_depth
        self.min_samples_split = min_samples_split
        self.root = None

    def fit(self, X, y):
        self.root = self._build_tree(X, y, depth=0)

    def _build_tree(self, X, y, depth):
        num_samples, num_features = X.shape

        if (
            depth >= self.max_depth
            or num_samples < self.min_samples_split
            or len(np.unique(y)) == 1
        ):
            return Node(value=np.mean(y))

        best_split = self._get_best_split(X, y, num_features)

        if best_split is None:
            return Node(value=np.mean(y))

        left_subtree = self._build_tree(
            X, y,
            best_split["X_left"],
            best_split["y_left"],
            depth + 1

```

```

    )

    right_subtree = self._build_tree(
        best_split["X_right"],
        best_split["y_right"],
        depth + 1
    )

    return Node(
        feature_idx=best_split["feature_idx"],
        threshold=best_split["threshold"],
        left=left_subtree,
        right=right_subtree
    )

def _get_best_split(self, X, y, num_features):
    best_split = None
    max_var_reduction = -np.inf

    for feature_idx in range(num_features):
        feature_values = X[:, feature_idx]
        thresholds = np.unique(feature_values)

        for threshold in thresholds:
            left_idx = np.where(feature_values <= threshold)[0]
            right_idx = np.where(feature_values > threshold)[0]

            if len(left_idx) == 0 or len(right_idx) == 0:
                continue

            y_left = y[left_idx]
            y_right = y[right_idx]

            var_reduction = self._variance_reduction(
                y,
                y_left,
                y_right
            )

```

```

        if var_reduction > max_var_reduction:
            max_var_reduction = var_reduction

        best_split = {
            "feature_idx": feature_idx,
            "threshold": threshold,
            "X_left": X[left_idx],
            "y_left": y_left,
            "X_right": X[right_idx],
            "y_right": y_right
        }

    return best_split

def _variance_reduction(self, parent, left_child, right_child):
    parent_var = np.var(parent)

    left_weight = len(left_child) / len(parent)
    right_weight = len(right_child) / len(parent)

    weighted_var = (
        left_weight * np.var(left_child)
        + right_weight * np.var(right_child)
    )

    return parent_var - weighted_var

def predict(self, X):
    return np.array([
        self._make_prediction(x, self.root)
        for x in X
    ])

def _make_prediction(self, x, tree):
    if tree.value is not None:
        return tree.value

    feature_val = x[tree.feature_idx]

```

```

        if feature_val <= tree.threshold:
            return self._make_prediction(x, tree.left)

        return self._make_prediction(x, tree.right)

class RandomForestRegressorManual:
    def __init__(self, n_estimators=10, max_depth=5):
        self.n_estimators = n_estimators
        self.max_depth = max_depth
        self.trees = []

    def fit(self, X, y):
        self.trees = []

        for i in range(self.n_estimators):
            print(f"Будуемо дерево {i + 1}/{self.n_estimators}...")

            indices = np.random.choice(
                X.shape[0],
                X.shape[0],
                replace=True
            )

            X_sample = X[indices]
            y_sample = y[indices]

            tree = DecisionTreeRegressorManual(
                max_depth=self.max_depth
            )

            tree.fit(X_sample, y_sample)

            self.trees.append(tree)

    def predict(self, X):
        predictions = np.array([
            tree.predict(X)
            for tree in self.trees

```

```

    ])

    return np.mean(predictions, axis=0)

print("\n--- Етап 2: Навчання кастомного Випадкового Лісу ---")

custom_rf = RandomForestRegressorManual(
    n_estimators=10,
    max_depth=4
)

custom_rf.fit(X_train, y_train)

print("\n--- Етап 3: Прогнозування та Метрики ---")

custom_rf_preds = custom_rf.predict(X_test)

print_metrics(
    "Custom Random Forest (from scratch)",
    y_test,
    custom_rf_preds
)

class GradientBoostingRegressorManual:
    def __init__(
        self,
        n_estimators=10,
        learning_rate=0.1,
        max_depth=3
    ):
        self.n_estimators = n_estimators
        self.learning_rate = learning_rate
        self.max_depth = max_depth
        self.trees = []
        self.base_prediction = 0

    def fit(self, X, y):

```

```

self.base_prediction = np.mean(y)

y_pred = np.full(len(y), self.base_prediction)

for i in range(self.n_estimators):
    print(f"        Бустинг:  дерево  {i +
1}/{self.n_estimators}...")

    residuals = y - y_pred

    tree = DecisionTreeRegressorManual(
        max_depth=self.max_depth
    )

    tree.fit(X, residuals)

    self.trees.append(tree)

    y_pred += (
        self.learning_rate * tree.predict(X)
    )

def predict(self, X):
    y_pred = np.full(
        X.shape[0],
        self.base_prediction
    )

    for tree in self.trees:
        y_pred += (
            self.learning_rate * tree.predict(X)
        )

    return y_pred

class LogisticRegressionManual:
    def __init__(self, learning_rate=0.01, epochs=500):
        self.lr = learning_rate

```

```

self.epochs = epochs
self.weights = None
self.bias = None

def _sigmoid(self, z):
    z = np.clip(z, -250, 250)

    return 1 / (1 + np.exp(-z))

def fit(self, X, y):
    n_samples, n_features = X.shape

    self.weights = np.zeros(n_features)
    self.bias = 0

    for _ in range(self.epochs):
        linear_model = np.dot(X, self.weights) + self.bias

        y_predicted = self._sigmoid(linear_model)

        dw = (
            (1 / n_samples)
            * np.dot(X.T, (y_predicted - y))
        )

        db = (
            (1 / n_samples)
            * np.sum(y_predicted - y)
        )

        self.weights -= self.lr * dw
        self.bias -= self.lr * db

def predict_proba(self, X):
    linear_model = np.dot(X, self.weights) + self.bias

    return self._sigmoid(linear_model)

```

```

class HurdleModelManual:
    def __init__(self):
        self.classifier = LogisticRegressionManual(
            learning_rate=0.01,
            epochs=500
        )

        self.regressor = GradientBoostingRegressorManual(
            n_estimators=10,
            learning_rate=0.1,
            max_depth=3
        )

    def fit(self, X, y):
        print(" [Hurdle] Етап 1: Навчання класифікатора...")

        y_class = (y > 0).astype(int)

        self.classifier.fit(X, y_class)

        print(" [Hurdle] Етап 2: Навчання регресора...")

        buyers_mask = y > 0

        X_buyers = X[buyers_mask]
        y_buyers = y[buyers_mask]

        self.regressor.fit(X_buyers, y_buyers)

    def predict(self, X):
        prob_purchase = self.classifier.predict_proba(X)

        pred_amount = self.regressor.predict(X)

        return prob_purchase * pred_amount

print("\n--- ЗАПУСК СПОСОБУ 2: Кастомний Градієнтний Бустинг ---")

```

```
custom_gbm = GradientBoostingRegressorManual(  
    n_estimators=10,  
    learning_rate=0.1,  
    max_depth=3  
)  
  
custom_gbm.fit(X_train, y_train)  
  
gbm_preds = custom_gbm.predict(X_test)  
  
print_metrics(  
    "Custom Gradient Boosting (Способ 2)",  
    y_test,  
    gbm_preds  
)  
  
print("\n--- ЗАПУСК СПОСОБУ 3: Кастомна Бар'єрна Модель ---")  
  
custom_hurdle = HurdleModelManual()  
  
custom_hurdle.fit(X_train, y_train)  
  
hurdle_preds = custom_hurdle.predict(X_test)  
  
print_metrics(  
    "Custom Hurdle Model (Способ 3)",  
    y_test,  
    hurdle_preds  
)
```

Метадані

ДОКУМЕНТ

Заголовок

КМР Бородатюк_на перевірку

Автор

Бородатюк Дмитро

Науковий керівник / Експерт

Колечкіна Людмила Миколаївна

ІД документа

333897168

ОРГАНІЗАЦІЯ

Назва організації

**Kyiv National Economic University named after Vadym Hetman
KNEU**

підрозділ

кафедра системного аналізу та кібербезпеки

ЗВІТ

Дата звіту

5/16/2026

Дата редагування

Обсяг знайдених подібностей

Коефіцієнт подібності визначає, який відсоток тексту по відношенню до загального обсягу тексту було знайдено в різних джерелах. Зверніть увагу, що високі значення коефіцієнта не автоматично означають плагіат. Звіт має аналізувати компетентна / уповноважена особа.



КП 1

25

Довжина фрази для коефіцієнта подібності 2



КП 2

13061

Кількість слів



КЦ

106831

Кількість символів

Тривога

У цьому розділі ви знайдете інформацію щодо текстових спотворень. Ці спотворення в тексті можуть говорити про МОЖЛИВІ маніпуляції в тексті. Спотворення в тексті можуть мати навмисний характер, але частіше характер технічних помилок при конвертації документа та його збереженні, тому ми рекомендуємо вам підходити до аналізу цього модуля відповідально. У разі виникнення запитань, просимо звертатися до нашої служби підтримки.

Заміна букв		0
Інтервали		0
Мікропробіли		0
Білі знаки		0
Парафрази (SmartMarks)		3

Джерела

Нижче наведений список джерел. В цьому списку є джерела із різних баз даних. Колір тексту означає в якому джерелі він був знайдений. Ці джерела і значення Коефіцієнту Подібності не відображають прямого плагіату. Необхідно відкрити кожне джерело і проаналізувати зміст і правильність оформлення джерела.

10 найдовших фраз

Колір тексту

#	НАЗВА ТА АДРЕСА ДЖЕРЕЛА URL (НАЗВА БАЗИ)	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
1	https://mvdissert.com/en/catalog/view/8730.html	13 (0.1%)

2	Тесленко_TAPR6_2024 10/21/2024 Publisher House "Technology Center" (TAPR)	10 (0.08 %)
3	https://onpage.school/shho-take-metriki-ta-kpi/	10 (0.08 %)
4	https://bizautomation.com.ua/internet-marketing-terminu/	10 (0.08 %)
5	https://bizautomation.com.ua/internet-marketing-terminu/	9 (0.07 %)
6	Pasko_Matsyi_Diplom_2025 12/8/2025 V. N. Karazin Kharkiv National University (KKNU) (ННІ комп'ютерних наук та штучного інтелекту - кафедра математичного моделювання та аналізу даних)	8 (0.06 %)
7	https://onpage.school/shho-take-metriki-ta-kpi/	7 (0.05 %)
8	Pasko_Matsyi_Diplom_2025 12/8/2025 V. N. Karazin Kharkiv National University (KKNU) (ННІ комп'ютерних наук та штучного інтелекту - кафедра математичного моделювання та аналізу даних)	7 (0.05 %)
9	https://openarchive.nure.ua/bitstreams/149eb547-5923-4996-88ac-fb5258f4d546/download	7 (0.05 %)
10	https://onpage.school/shho-take-metriki-ta-kpi/	5 (0.04 %)

База даних RefBooks



#	ЗАГОЛОВОК	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
---	-----------	--

Домашня база даних



#	ЗАГОЛОВОК	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
---	-----------	--

Програма обміну базами даних (0.19 %)



#	ЗАГОЛОВОК	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
---	-----------	--

1	Pasko_Matsyi_Diplom_2025 12/8/2025 V. N. Karazin Kharkiv National University (KKNU) (ННІ комп'ютерних наук та штучного інтелекту - кафедра математичного моделювання та аналізу даних)	15 (2) (0.11 %)
2	Тесленко_TAPR6_2024 10/21/2024 Publisher House "Technology Center" (TAPR)	10 (1) (0.08 %)

Інтернет (0.58 %)



#	ДЖЕРЕЛО URL	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
3	https://onpage.school/shho-take-metriki-ta-kpi/	27 (4) (0.21 %)
4	https://bizautomation.com.ua/internet-marketing-terminu/	24 (3) (0.18 %)
5	https://mydisser.com/en/catalog/view/8730.html	13 (1) (0.1 %)



Список прийнятих фрагментів

#	ЗМІСТ	КІЛЬКІСТЬ ОДНАКОВИХ СЛІВ (ФРАГМЕНТІВ)
---	-------	---------------------------------------