МІНІСТЕРСТВО ОСВІТИ ТА НАУКИ УКРАЇНИ
ПВНЗ
«МІЖНАРОДНИЙ НАУКОВО-ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ АКАДЕМІКА Ю. БУГАЯ»

Кафедра інформаційних та комунікаційних технологій

Освітня компонента: «Інтелектуальний аналіз даних»

Дослідницька РГР – Візуальні компоненти

Виконав:
студент 4 курсу денної форми навчання
група І26
Бондаренко О.В.

Перевірила:
доц. Гриб'юк О.О.

Київ – 2025

Тема проєкту: Аналіз поведінкових даних веб-проєктів.

Прізвище та ім'я виконавців проєкту: Бондаренко Олег.

Концепція проєкту: посилання на github.

Мета проєкту: формування системи інтелектуального аналізу поведінкових даних веб-проєктів з метою моделювання сценаріїв розвитку бізнес-рішень, виявлення трендів та оптимізації цифрової взаємодії з користувачами.

Завдання проєкту:

1. Отримати вибіркові дані поведінкових метрик користувачів веб-проєкту з різних джерел (Google Analytics, Binotel, CRM) та провести їх первинну обробку.
2. Здійснити регресійний та survival-аналіз для виявлення впливу поведінкових ознак (наприклад, page.view, ad_click) на ймовірність покупки або конверсії.
3. Провести розкладання часових рядів поведінкових метрик на складові (тренд, сезонність, випадкова компонента) та оцінити наявність тенденцій за допомогою критерію серій.
4. Сформувати висновки та рекомендації щодо оптимізації бізнес-рішень на основі виявлених закономірностей, трендів та результатів моделювання.

ВІЗУАЛЬНІ КОМПОНЕНТИ

Вхідні данні. Вибірка містить такі змінні:

- page_view – кількість переглядів сторінок;
- ad_click – кількість кліків по рекламі;
- purchase – факт покупки (цільова змінна, де 1 – купив, 0 – не купив);
- duration відображає кількість хвилин, що минули з моменту показу реклами до покупки або завершення спостереження. Для purchase = 1 – тривалість варіюється від 100 до 5000 хв. Для purchase = 0 – тривалість становить від 5000 до 7200 хв, що моделює довше спостереження без покупки.

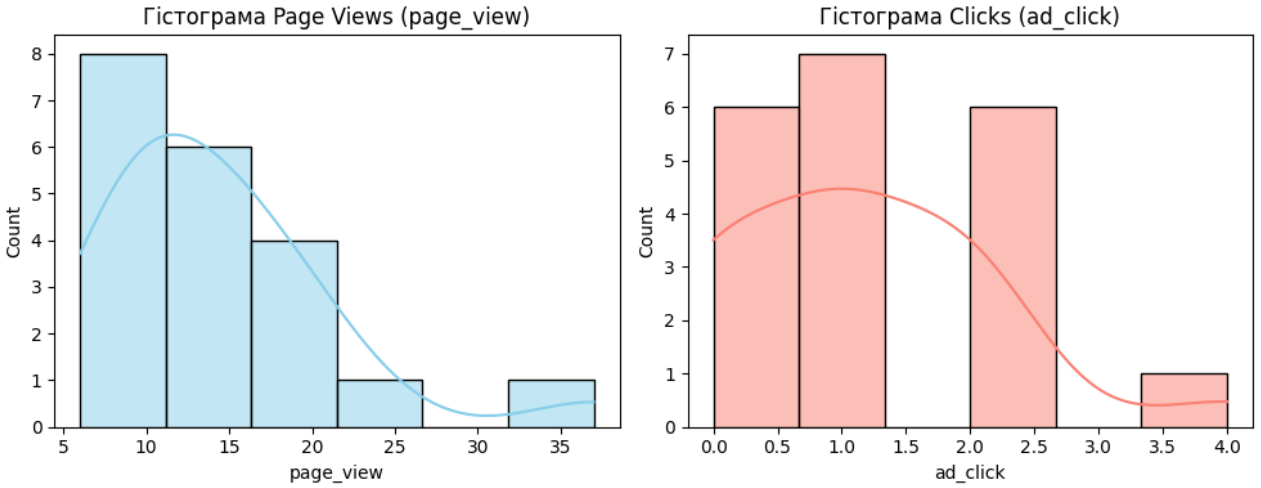
	page_view	ad_click	purchase	duration
Харків	37	4	1	4013
Київ	23	2	1	1764
Львів	20	1	0	6650
Одеса	8	0	0	7102
Дніпро	15	1	1	2345
Запоріжжя	12	0	1	2193
Полтава	18	2	1	3708
Черкаси	10	1	0	6272
Чернігів	9	0	0	5980
Суми	11	1	0	5821
Івано-Франківськ	14	2	1	1399
Тернопіль	13	1	0	6520
Ужгород	7	0	0	6655
Вінниця	16	2	1	3188
Житомир	10	1	0	7200
Кропивницький	6	0	0	6890
Миколаїв	19	2	1	1203
Хмельницький	17	2	1	2890
Рівне	12	1	1	495
Луцьк	9	0	0	7100

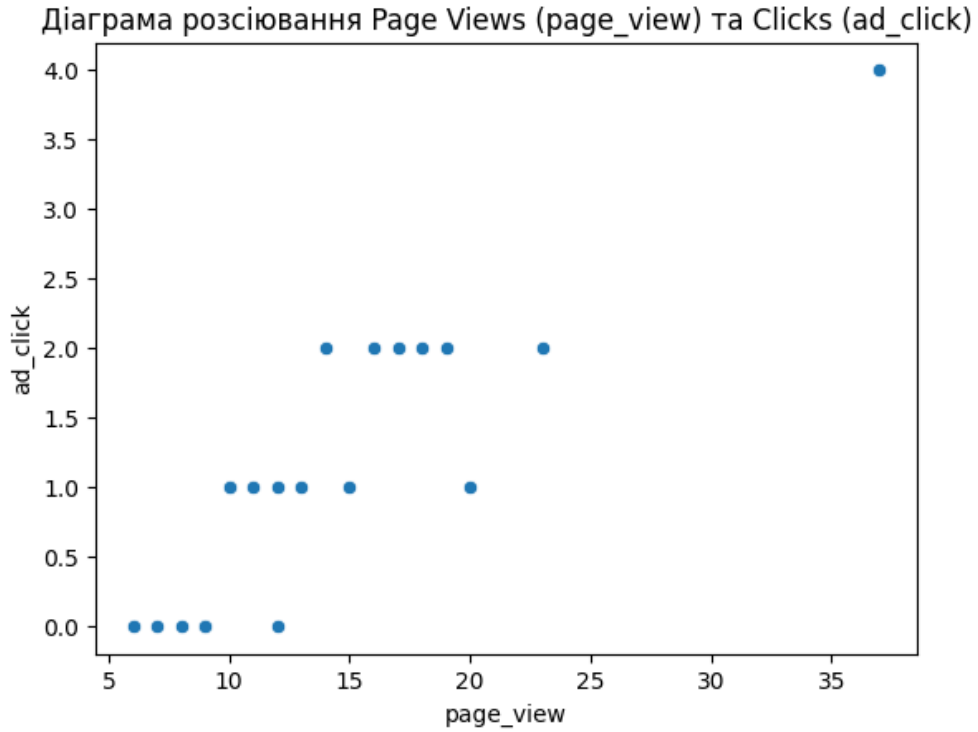
Вибірка охоплює 20 міст України з різними показниками переглядів, кліків та покупок. Змінна duration дозволяє моделювати часову динаміку покупки.

1. Метод кореляційного аналізу даних

Побудовано дві гістограми для змінних page_view та ad_click, а також діаграму розсіювання, яка демонструє наявність прямого зв’язку між ними:

- розподіл значень;
- графік кореляції.



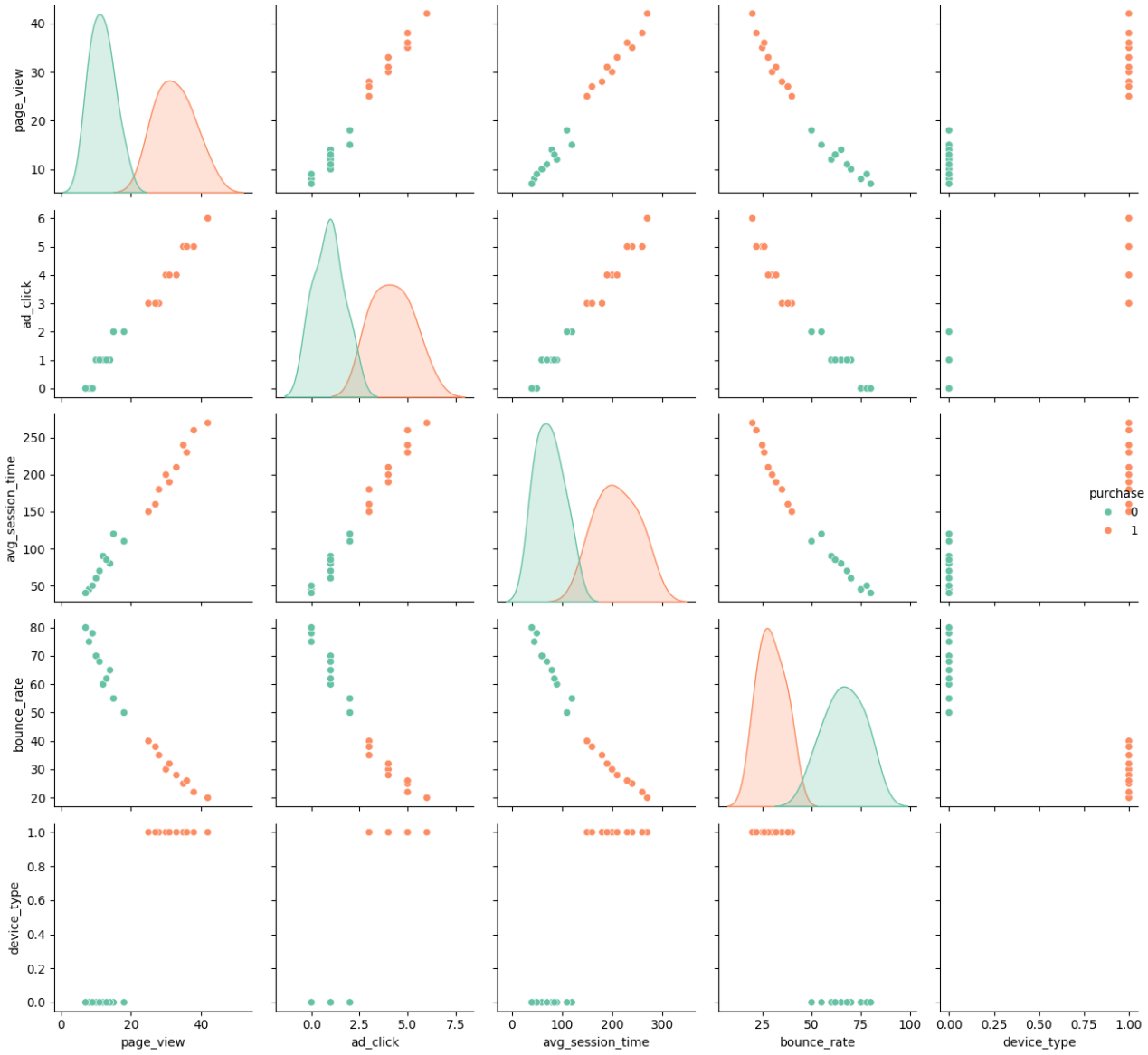


2. Лінійний регресійний аналіз + Інтелектуальна задача

Для дослідження взаємозв'язків між поведінковими метриками було побудовано pairplot, який показує:

- Розподіл кожної ознаки (на діагоналі – KDE-графіки)
- Взаємозв'язки між ознаками (scatterplots між парами)
- Класифікацію: кольором позначено, чи користувач здійснив покупку (purchase = 1) чи ні (purchase = 0)

Візуалізація дозволяє оцінити, які ознаки найкраще розділяють класи, і виявити потенційні кореляції між метриками.

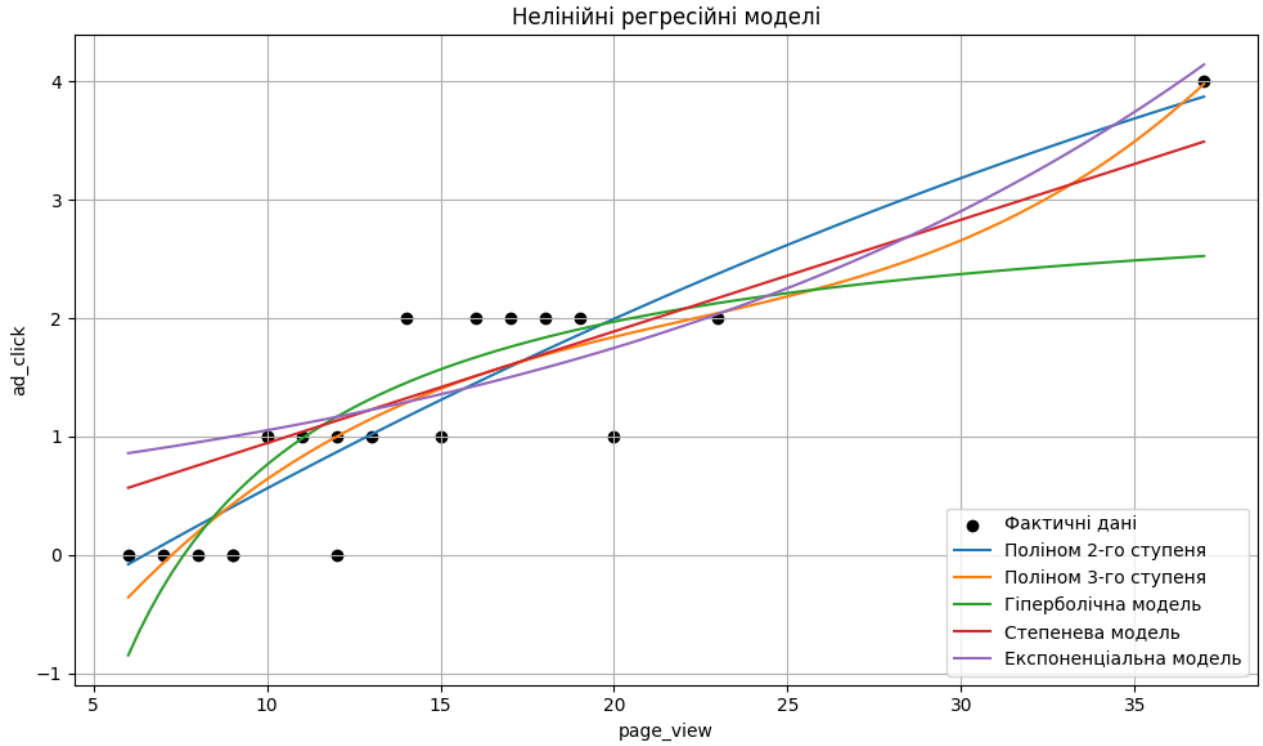


3. Нелінійний регресійний аналіз

На графіку нижче зображено фактичні дані (чорні точки) та криві п’яти побудованих моделей:

- Поліном 2-го ступеня – синя лінія.
- Поліном 3-го ступеня – помаранчева лінія.
- Гіперболічна модель – зелена лінія.
- Степенева модель – червона лінія.
- Експоненціальна модель – фіолетова лінія.

Графік дозволяє візуально порівняти, як кожна модель апроксимує залежність між page_view та ad_click. Найкраще узгодження з фактичними даними демонструє поліном 3-го ступеня, який плавно проходить через основні кластери точок. Степенева модель також показує хорошу відповідність, особливо в середньому діапазоні значень.



Побудоване дерево рішень дозволяє формалізувати поведінку користувача на основі двох ознак: page_view та ad_click. Нижче наведено розбір кожного вузла дерева, включаючи умови розгалуження, метрики чистоти вузла та прогнозовані класи.

gini – ступінь неоднорідності вузла, індекс Джині (Gini impurity). Показує, наскільки вузол «змішаний» за класами. Значення від 0,0 (чистий вузол, всі об'єкти одного класу) до 0,5 (максимальна неоднорідність при 2 класах).

samples – кількість об'єктів у вузлі. Скільки рядків даних потрапили в цей вузол.

value – розподіл об'єктів по класах. Список: [кількість класу 0 (не купили), кількість класу 1 (купили)].

class – прогнозований клас вузла (Yes/No). Це модельне рішення: який клас буде присвоєно новому об'єкту, якщо він потрапить у цей вузол. Визначається як той, що має більшість у value.

Вузол 0 – Корінь дерева. Розділяє всі дані на дві гілки: ліворуч – мало переглядів (≤11,5) та праворуч – більше переглядів (>11,5). gini = 0.15 – майже чистий вузол, домінує клас «Yes», value = [3, 17] – 3 не купили, 17 купили.

Ліва гілка (True):

Вузол 1. Усі 8 об'єктів – купили. Абсолютно чистий вузол. Модель впевнено прогнозує «Yes».

Права гілка (False):

Вузол 2. Додаткове розгалуження: якщо кліків ≤1,5, value = [2, 10] – 2 не купили, 10 купили, gini = 0,28 – помірна неоднорідність.

Вузол 3 – Ліва гілка від ad_click <= 1,5 (True). Слабка домінанта класу «Yes», gini = 0,48 – вузол неоднорідний, додаткове розгалуження за page_view.

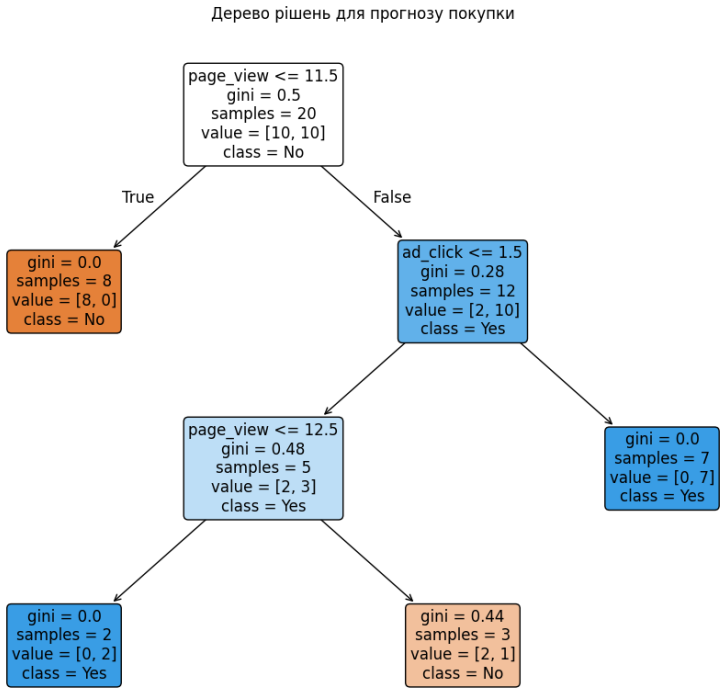
Вузол 4 – Ліва гілка від page_view <= 12,5 (True). Обидва об'єкти – купили, чистий вузол, прогноз «Yes».

Вузол 5 – Права гілка від page_view <= 12,5 (False). Більшість – не купили, gini = 0,44 – неоднорідність, прогноз – «No».

Вузол 6 – Права гілка від ad_click <= 1,5 (False). Усі 7 об'єктів – купили, чистий вузол, прогноз «Yes».

Загальна логіка дерева (логіка прийняття рішень на основі дерева):

Умова	Прогноз
page_view ≤ 11,5	Не купив
page_view > 11,5 ∧ ad_click > 1,5	Купив
page_view > 11,5 ∧ ad_click ≤ 1,5 ∧ page_view ≤ 12,5 ∧ value = [0,2]	Купив
page_view > 11,5 ∧ ad_click ≤ 1,5 ∧ page_view > 12,5 ∧ value = [2,1]	Не купив



На графіку нижче зображено межу класифікації, яку сформувала нейронна мережа (архітектура 3-1) на основі ознак page_view та ad_click.

Фон поділений на дві зони:

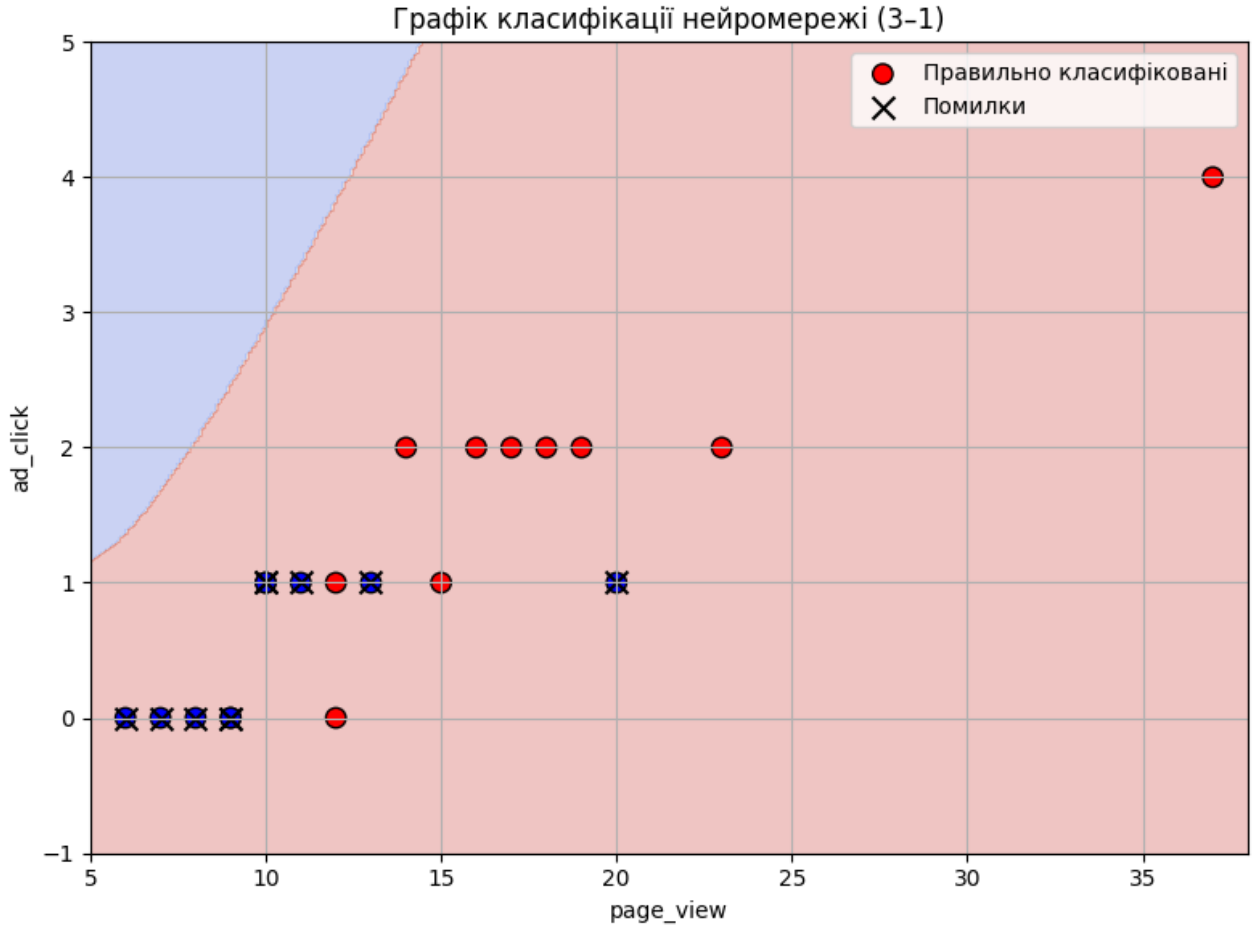
- Червона зона – модель класифікує як клас «Купив» (1).
- Синя зона – модель класифікує як клас «Не купив» (0).

На графіку також відображено реальні об'єкти:

- Червоні точки – користувачі, які реально здійснили покупку.
- Сині точки – користувачі, які не купили.
- Чорні хрестики – помилки класифікації, тобто об'єкти, які модель класифікувала неправильно.

Межа класифікації має вигнуту форму, що свідчить про нелінійний характер моделі. Незважаючи на баланс класів у вибірці (10 купив / 10 не купив), модель демонструє зміщення в сторону класу «Купив», що видно з великої червоної зони та кількості помилок у синій частині.

Ця візуалізація дозволяє оцінити якість класифікації, виявити граничні випадки та сформулювати рекомендації щодо покращення моделі (наприклад, додавання нових ознак або зміна архітектури).



4. Логістична регресія

Для наочного представлення роботи моделі логістичної регресії побудовано три графіки.

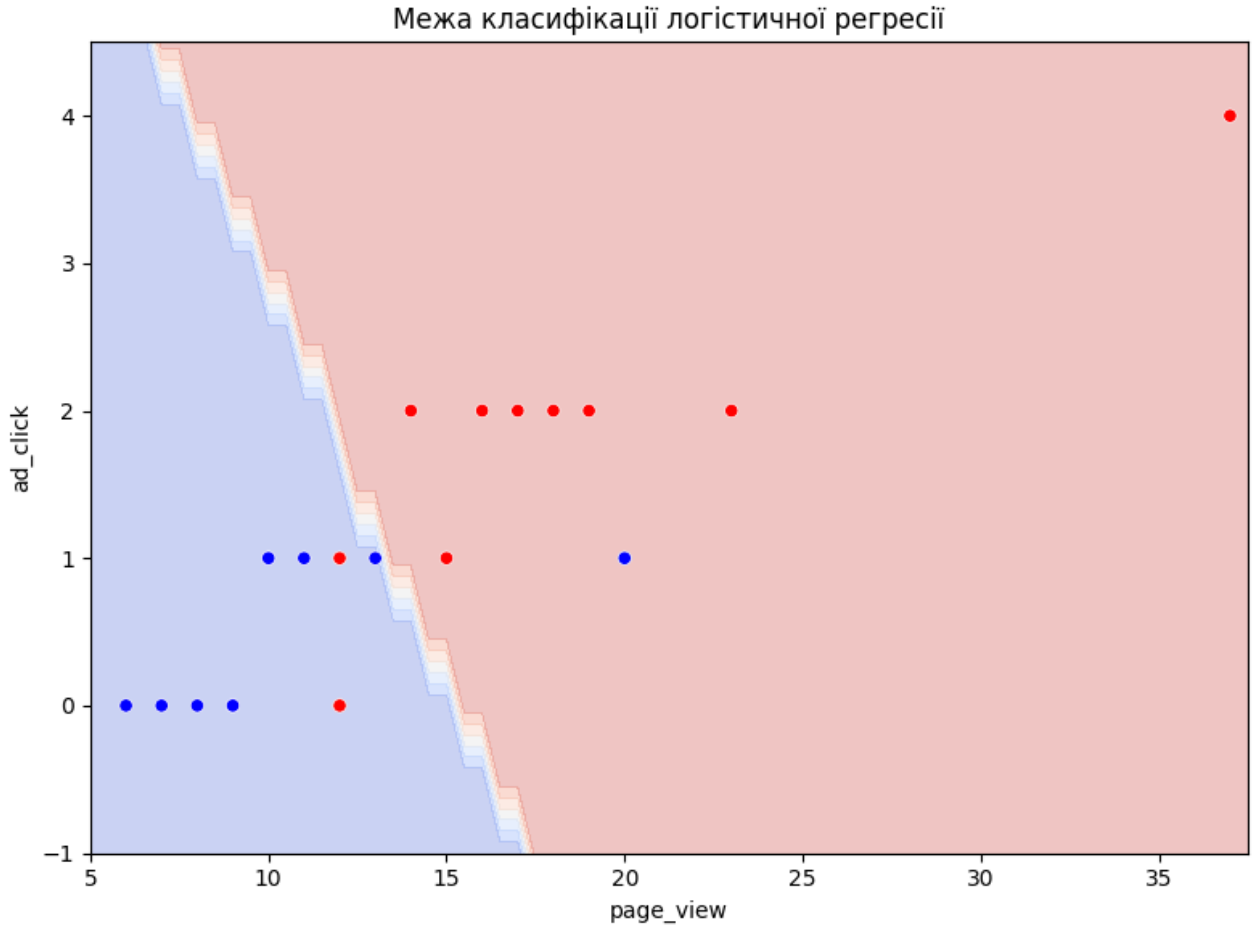
1. Межа класифікації логістичної регресії. На графіку зображено, як модель розділяє простір ознак page_view та ad_click.

Фон зафарбовано відповідно до передбаченого класу:

🔴 червоний – модель прогнозує, що користувач не здійснить покупку

🔵 синій – модель прогнозує, що користувач здійснить покупку

Реальні точки (міста) нанесено поверх фону, що дозволяє оцінити точність класифікації. Це свідчить про те, що модель має високу чутливість до ознак, але в прикордонних зонах ймовірність помилки зростає – що є типовим для логістичної регресії.



2. Логістична крива: purchase ~ page_view

Цей графік демонструє, як змінюється ймовірність покупки залежно від кількості переглядів сторінок (page_view).

- Чорна крива – логістична функція
- Сірий фон – 95% довірчий інтервал
- Точки – реальні об'єкти (🔵 не купив, 🔴 купив)

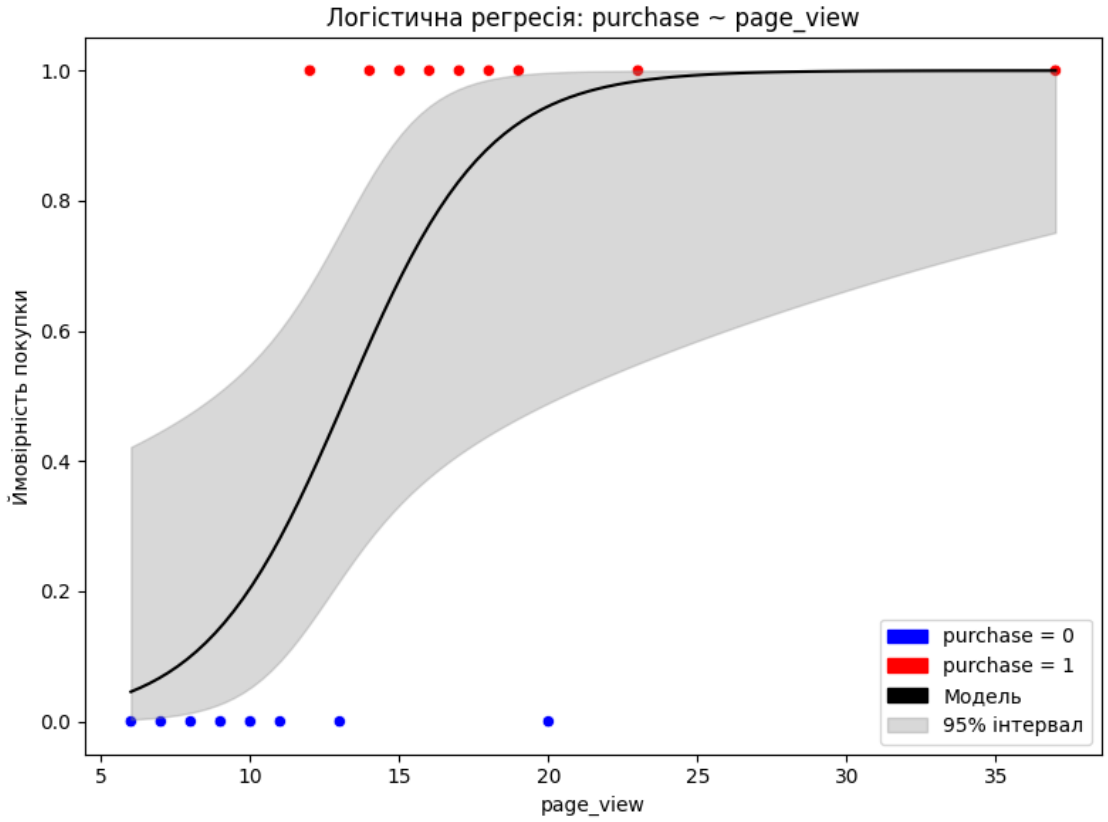
Це узгоджується з гіпотезою про те, що активність користувача (page_view) є індикатором зацікавленості, що підвищує шанси на конверсію.

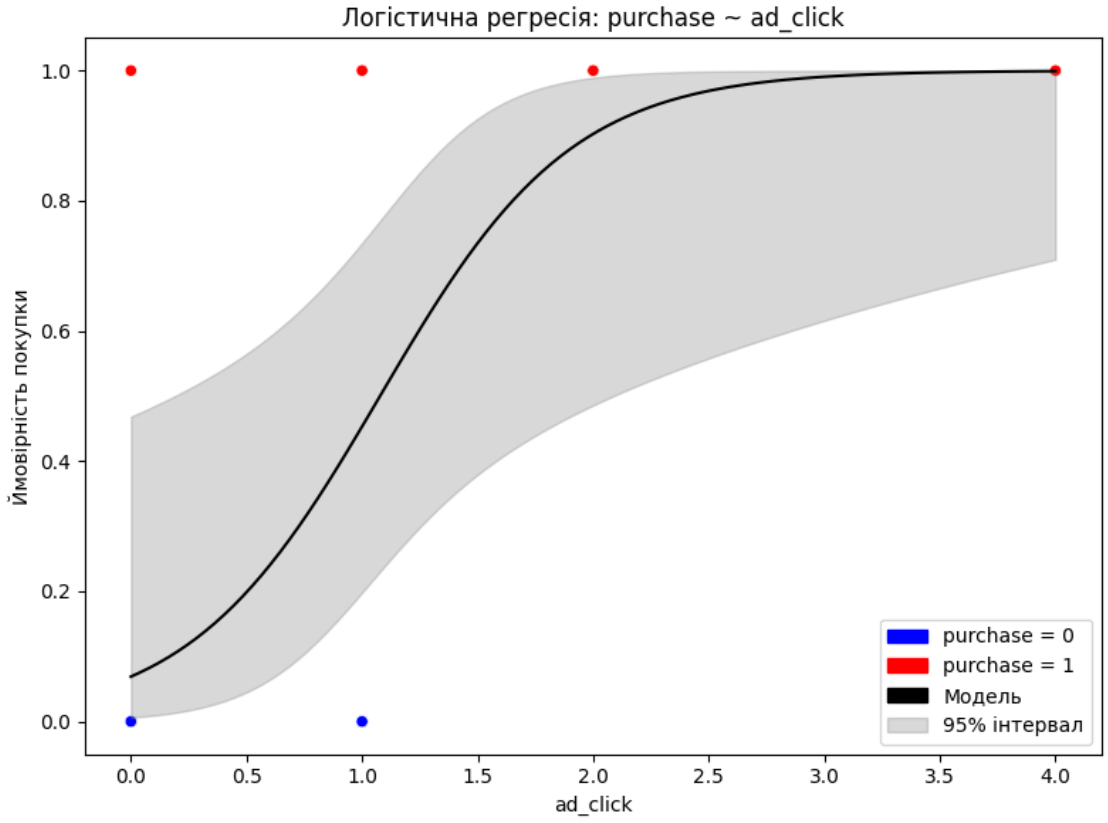
3. Логістична крива: purchase ~ ad_click

Аналогічно, цей графік показує залежність ймовірності покупки від кількості кліків по рекламі (ad_click).

Модель демонструє чітку позитивну залежність: чим більше кліків – тим вища ймовірність покупки.

Таким чином, ознака ad_click може бути використана як ключовий тригер для таргетованих маркетингових рішень.





5. Регресія Кокса

Для наочного представлення результатів Cox-регресії побудовано графік коефіцієнтів моделі у вигляді логарифмів hazard ratio (log(HR)) з 95% довірчими інтервалами.

На графіку зображено дві ознаки:

- ad_click – кількість кліків по рекламі;
- page_view – кількість переглядів сторінок.

Кожна ознака представлена точкою (оцінка коефіцієнта) та горизонтальною лінією (довірчий інтервал). Вертикальна пунктирна лінія на рівні 0 відповідає нейтральному ефекту (HR = 1).

Інтерпретація:

- ad_click має позитивний коефіцієнт (log(HR) > 0), що свідчить про зростання ризику покупки з кожним додатковим кліком. Проте довірчий інтервал перетинає нуль, тому ефект не є статистично значущим.
- page_view має негативний коефіцієнт (log(HR) < 0), що вказує на зниження ризику покупки при збільшенні переглядів. Довірчий інтервал не перетинає нуль, тому ефект можна вважати статистично значущим.