ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ
ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА ІНЖЕНЕРІЇ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

КВАЛІФІКАЦІЙНА РОБОТА

на тему: « Підвищення результативності рекламної компанії на основі методів машинного навчання »

на здобуття освітнього ступеня магістра
зі спеціальності 121 Інженерія програмного забезпечення
(код, найменування спеціальності)
освітньо-професійної програми «Інженерія програмного забезпечення»
(назва)

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

_____ Юрій КУХАРЕНКО
(підпис)

Виконав: здобувач вищої освіти групи ПДМ-64
Юрій КУХАРЕНКО

Керівник: Володимир САДОВЕНКО
к.ф.-м.н., доцент

Рецензент:
науковий ступінь, вчене звання Ім’я, ПРІЗВИЩЕ

Київ 2024

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Навчально-науковий інститут інформаційних технологій

Кафедра Інженерії програмного забезпечення

Ступінь вищої освіти Магістр

Спеціальність 121 Інженерія програмного забезпечення

Освітньо-професійна програма «Інженерія програмного забезпечення»

ЗАТВЕРДЖУЮ
Завідувач кафедри
Інженерії програмного забезпечення
_____ Ірина ЗАМРІЙ
«_____» _____ 2023 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Кухаренку Юрію Дмитровичу

1. Тема кваліфікаційної роботи: «Підвищення результативності рекламної компанії на основі методів машинного навчання»

керівник кваліфікаційної роботи Володимир САДОВЕНКО к.ф.-м.н., доцент,

затверджені наказом Державного університету інформаційно-комунікаційних технологій від «19» жовтня 2023 р. №145

2. Строк подання кваліфікаційної роботи «29» грудня 2023р.

3. Вихідні дані до кваліфікаційної роботи: науково-технічна література з питань, пов'язаних з алгоритмами та методами прогнозування.

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

1. Огляд предметної області.
2. Опис методів прогнозування.
3. Розробка методу прогнозування

5. Перелік графічного матеріалу: *презентація*

1. Мета, об’єкт та предмет дослідження.
2. Актуальність теми.
3. Методи моделювання.
4. Математичін розрахунки та схема алгоритму градієнтний бустинг.
5. Алгоритм роботи методу.
6. Вхідні дані.
7. Практичний результат.
8. Результати тренування моделі.
9. Висновки.

6. Дата видачі завдання «19» жовтня 2023 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1	Підбір науково-технічної літератури	19.10-05.11.23	
2	Огляд предметної області	06.11-12.11.23	
3	Створення моделі для прогнозування погоди	13.11-19.11.23	
4	Аналіз методів прогнозування	20.11-26.11.23	
5	Застосування машинного навчання при обробці даних	27.11-03.12.23	
6	Розробка моделей та методів	04.12-10.12.23	
7	Оформлення роботи: вступ, висновки, реферат	11.12-20.12.23	
8	Розробка демонстраційних матеріалів	21.12-29.12.23	

Здобувач вищої освіти _____ Юрій КУХАРЕНКО
(підпис)

Керівник
кваліфікаційної роботи _____ Володимир САДОВЕНКО
(підпис)

РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття освітнього ступеня магістра: 70 стор., 34 рис., 35 джерел.

Мета роботи – збільшення доходності рекламної компанії на основі методів машинного навчання.

Об'єкт дослідження – процес прогнозування ефективності реклами.

Предмет дослідження – методи прогнозування машинного навчання для прогнозування ефективності реклами.

Короткий зміст роботи: У роботі проведено аналіз існуючих методів, типів та алгоритмів прогнозування ефективності. Проведено огляд сучасних систем автоматизації реклами, що використовуються для підвищення доходності реклами. Проаналізовано важливість використовування машинного навчання в сфері маркетингу та в інших професійних сферах.

Розроблено модель методу прогнозування ефективності реклами, що дозволяє здійснювати прогнозування того, наскільки прибутковою буде реклама, з використанням методу машинного навчання під назвою градієнтний бустинг, що проводить навчання на основі вибірки історичних даних та надає прогноз щодо майбутньої прибутковості реклами.

Було виконано аналіз моделей прогнозування за результатами, порівняно їх з існуючими методами прогнозування погоди. Визначено сильні та слабкі сторони цих моделей, встановлено ключові фактори, що значно впливають на точність прогнозу та виявлено залежність від вхідних даних.

КЛЮЧОВІ СЛОВА: МАШИННЕ НАВЧАННЯ, ДОХІДНІСТЬ РЕКЛАМИ, РЕКЛАМА, ІНОВАЦІЇ, МАРКЕТИНГ.

ABSTRACT

Text part of the master's qualification work: 70 pages, 34 pictures, 35 sources.

The purpose of the work - increasing the profitability of an advertising company based on machine learning methods.

Object of research – the process of predicting advertising effectiveness.

Subject of research – machine learning methods for predicting advertising effectiveness.

Summary of the work: The work analyzes existing methods, types, and algorithms for predicting effectiveness. A review of modern advertising automation systems used to enhance advertising profitability is conducted. The importance of using machine learning in marketing and other professional fields is analyzed.

A model of the advertising effectiveness prediction method is developed, allowing forecasting the profitability of advertising using the gradient boosting machine learning method. This method is trained based on a sample of historical data and provides forecasts regarding the future profitability of advertising.

An analysis of prediction models was carried out, comparing them with existing weather forecasting methods. The strengths and weaknesses of these models were identified, key factors significantly influencing prediction accuracy were determined, and dependence on input data was revealed.

KEYWORDS: MACHINE LEARNING, ADVERTISING PROFITABILITY, ADVERTISING, INNOVATION, MARKETING.

ЗМІСТ

ВСТУП

Прогнозування ефективності рекламних кампаній та автоматизація управління ними стають ключовими аспектами для продуктових компаній. Ці системи дозволяють ретельно контролювати множину рекламних ініціатив, точно оцінювати їх результативність і раціонально витрачати бюджетні ресурси.

Реклама – це ретельно науково спланована система заходів з метою просування товарів на ринок. Нині без належної маркетингової стратегії проекти майже не мають можливості існувати, адже головна мета маркетингу - ефективно розпоряджатися фінансовими ресурсами та привертати широке коло потенційних клієнтів.

У першому розділі цієї роботи проведено аналіз області маркетингу та докладно розглянуто актуальність розробки методів підвищення ефективності маркетингу. Проаналізовано основні проблеми, що виникають у вчених цього напрямку, а також формалізується постановка даної задачі.

У другому розділі розкрито сутність машинного навчання, різноманітні підходи до цієї техніки, математичні основи та методи побудови систем прогнозування. Тут детально розглядається теоретичний аспект машинного навчання. Третій розділ присвячено аналізу даних, які використовуються для тренування моделі, а також процесу побудови самої моделі. Тут надається інформація про вхідні дані, на яких буде тренуватися модель, та сам процес створення моделі і проведення аналізу отриманих результатів прогнозування.

1 ДОСЛІДЖЕННЯ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Актуальність проблеми

Продуктові компанії надто часто направляють величезні кошти на рекламу, чи то онлайн, чи в реальному світі, покладаючись на роботу маркетологів. Але тут постає проблема: через недостатню кількість даних маркетологам складно оцінити ефективність рекламних кампаній.

Є три ключові мотивації для вивчення рекламних кампаній. По-перше, це цікаво розібратися, як саме маркетологи впроваджують рекламні стратегії. Часто вони називають маркетинг мистецтвом і приймають рішення відчуттям.

По-друге, було б добре мати систему, яка виконує те ж саме, що й людина у маніпулюванні рекламними кампаніями. Це звільнило б маркетологів від рутинної роботи й дало б їм можливість більше творчо працювати та експериментувати.

Нарешті, система управління рекламними кампаніями може краще оцінювати ефективність реклами, ніж людина. Вона здатна аналізувати величезний обсяг історичних даних та виявляти закономірності, недоступні маркетологам, що дозволяє краще розпорошувати бюджет.

У сучасному світі для успішного продукту важливість не обмежується лише самих ідеєю та реалізацією, але й належить маркетинговим зусиллям. Неабияке значення у цьому контексті має роль маркетингового відділу в компаніях, які спеціалізуються на створенні продуктів для широкого кола споживачів.

Сучасний бізнес, особливо у сфері масового споживання, стикається з викликами, пов'язаними не лише з концепцією та виробництвом, але й з ефективністю й результативністю маркетингових стратегій. Споживачі, живучи в інформаційній епохи, стають більш вимогливими та освіченими, що підкреслює важливість правильного позиціонування продукту на ринку.

З іншого боку, чим більший проект, тим більше обсяг інформації, яку необхідно обробляти, і, відповідно, збільшується ризик виникнення помилок.

Щоб уникнути таких ситуацій і максимізувати ефективність маркетингових кампаній, сучасні компанії на шляху до успіху все частіше вдаються до використання автоматизованих систем маркетингу. Це дозволяє підвищити точність аналізу даних, забезпечити більш ефективну комунікацію з аудиторією та динамічно адаптувати стратегії відповідно до змін на ринку.

Такий підхід до маркетингових процесів відкриває перед компаніями можливості ефективнішого використання своїх ресурсів, сприяючи зниженню ризику помилок та значному підвищенню шансів на успіх у конкурентному середовищі. В сучасному бізнес-середовищі автоматизація маркетингу визнається не просто ефективним інструментом, але й однією з ключових стратегій для досягнення високого рівня конкурентоспроможності та забезпечення стабільного росту компанії. Цей підхід не лише сприяє оптимізації робочих процесів, але й дозволяє бізнесу більш гнучко та адаптивно реагувати на зміни в ринкових умовах, що є важливим аспектом у сучасній динамічній економіці.

Однією з головних проблем у прогнозуванні ефективності рекламних кампаній, що потребує негайного вирішення, є проблема з даними. Починаючи з 26 квітня 2021 року, компанія Apple випустила оновлення IOS 14.5, що вимагає дозволу від користувача перед збором їхніх даних зі сторонніх додатків та веб-сайтів для рекламних цілей [1]. Показання сповіщень для отримання цього дозволу можна побачити на рисунку 1.1. Ця нова політика обмежує доступ до повної інформації про основні метрики рекламної кампанії.

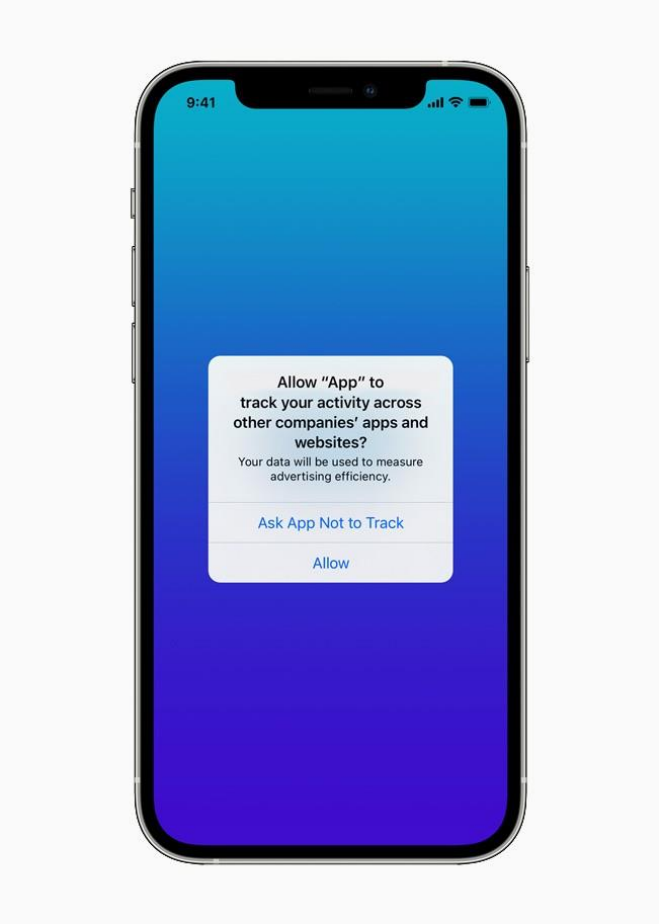


Рис. 1.1 Повідомлення для користувачів в IOS

Аналітика, належна Verizon Media, провела масштабний аналіз, охопивши понад мільйон мобільних додатків і збираючи загальну інформацію з 2 мільярдів мобільних пристроїв щомісяця. Прогнозується, що цей перехід призведе до значних змін у сфері персоналізованої реклами та атрибуції. Зокрема, його вплив значно відчутний буде в галузі мобільної реклами по всьому світу, що оцінюється на суму 189 мільярдів доларів [2].

За інформацією за два тижня (з 26.04.2023 по 08.05.2023) 87% користувачів не дозволяють передавати свою інформацію в додатки. Аналіз даних показано на графіках.

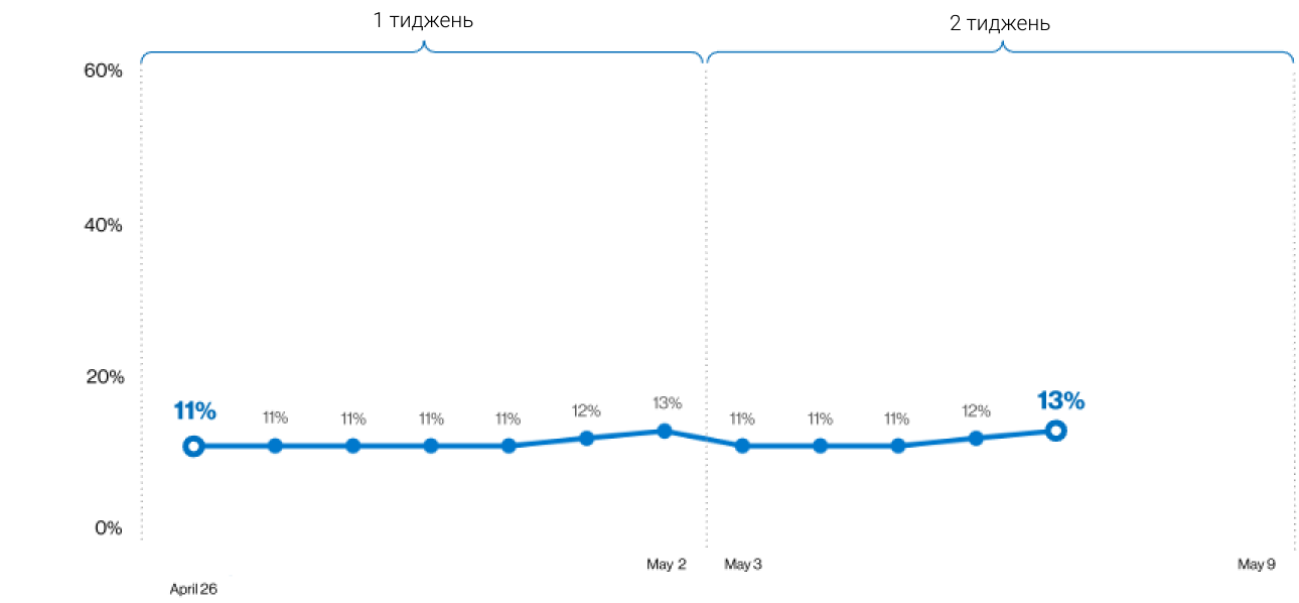


Рис. 1.2 Відсоток користувачів, що дозволили використання даних

Рис. 1.3 Доля користувачів, які дали дозвіл на використання їх даних в США

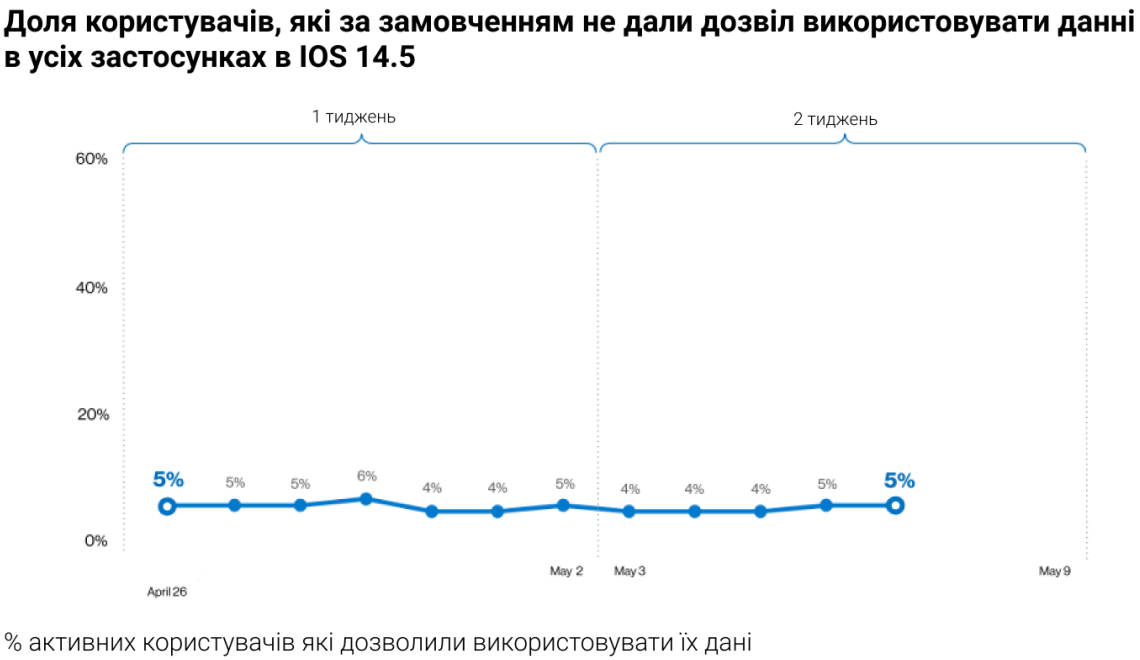


Рис. 1.4 Відсоток користувачів, які відмовились надавати доступ до даних

1.2 Наявні системи автоматизації керування рекламою

Компанія Netflix використовує систему автоматизації маркетингу для досягнення своїх цілей. Ось деякі з досягнень, які вони відзначають:

1. Оптимізація робочих процесів: Вони розробили технологію для організації робочих процесів з метою покращення ресурсів в маркетингу. Це дозволяє їхній команді маркетингу зосереджуватися на творчих аспектах, витрачаючи менше часу на рутинні завдання.

2. Створення єдиної платформи: Вони розробили внутрішню платформу для створення рекламних матеріалів та кампаній на різних рекламних каналах, включаючи Facebook, YouTube, Instagram, телебачення та інші медіа-платформи. Це дозволяє їм координувати та оптимізувати рекламу на різних каналах.

3. Технологія вимірювання та оптимізації ефективності: Netflix використовує технології для вимірювання та оптимізації ефективності своїх маркетингових кампаній. Це охоплює як онлайн-канали, так і офлайн-канали.

Наприклад, вони застосовують алгоритми на платформах, таких як Facebook та YouTube, для кращого розуміння впливу своїх кампаній та оптимізації витрат.

Ці підходи дозволяють Netflix не лише ефективно рекламувати свій контент, але й оптимізувати витрати та підвищувати ефективність маркетингових кампаній.

Netflix використовує розгалужену систему, побудовану на Java та Groovy, що базується на мікросервісах. Ці мікросервіси взаємодіють з різними сховищами NoSQL, такими як Cassandra та Elasticsearch, і використовують Kafka та Hermes для об'єднання компонентів, передачі даних та спрацювання подій, що активує запуск додатків у контейнерах на Titus.

Netflix інтенсивно використовує RxJava, а їх сервер оголошень, який обслуговує запити у реальному часі для відтворення рекламних дисплеїв та відео VAST, використовує RxNetty для своєї програмної реалізації. Цей підхід дозволяє налаштовувати операції з мінімальними можливостями та пов'язаними з ними накладними витратами. Для середнього рівня сервера оголошень Netflix використовує службу, яка базується на Tomcat / Jersey / Guice. Це дозволяє їм отримати більше можливостей та спрощує інтеграцію, зокрема для таких аспектів, як легка автентифікація та авторизація, а також повну підтримку хмарної екосистеми Netflix [4]. Схема системи показано на рисунку 1.5 та рисунку 1.6.

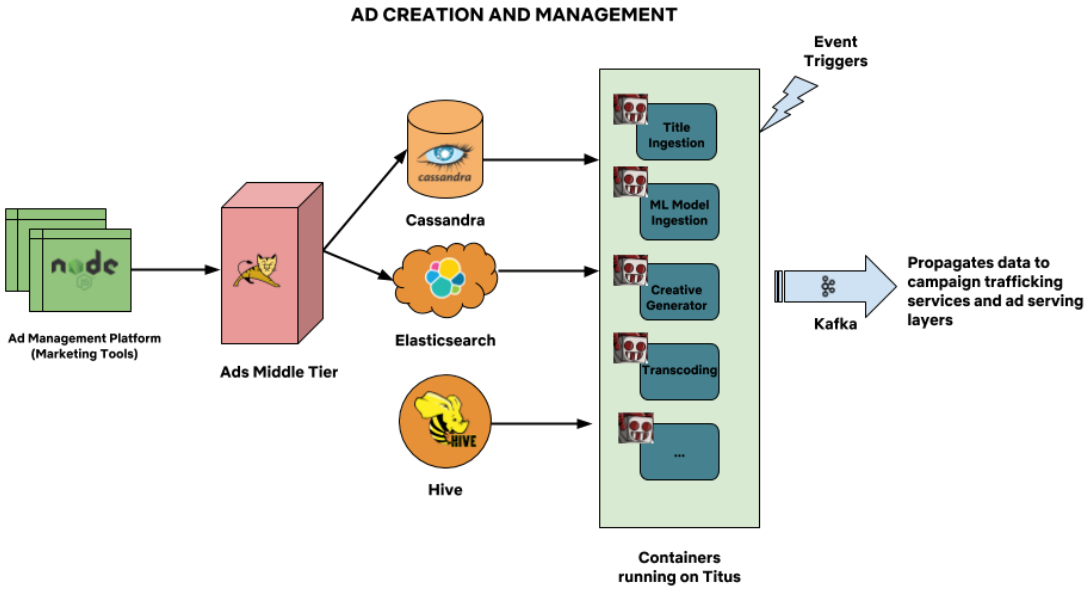


Рис. 1.5 Система автоматизації Netflix [4]

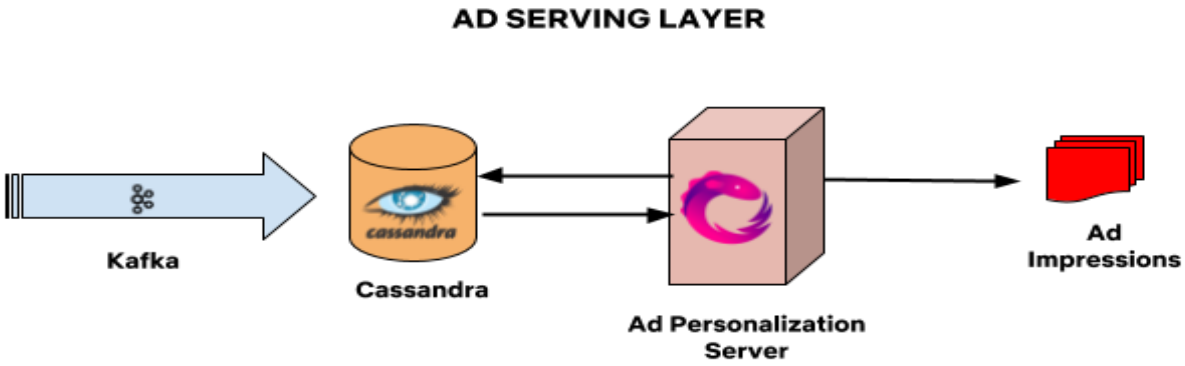


Рис. 1.6 Система автоматизації Netflix [4]

1.3 Інструменти поширювання реклами

Інструменти поширювання реклами є важливою складовою стратегії маркетингу, спрямованої на залучення уваги та збільшення свідомості про продукт чи послугу. Ці інструменти включають різноманітні канали та платформи, які дозволяють компаніям досягати своєї аудиторії. Ось деякі з популярних інструментів поширювання реклами:

1. Соціальні мережі: Платформи, такі як Facebook, Instagram, Twitter та LinkedIn, дозволяють розміщувати рекламу перед широкою аудиторією. Реклама може бути таргетованою на конкретних користувачів за їхніми інтересами, демографічними характеристиками та поведінковими патернами.

2. Пошуковий маркетинг: Реклама в пошукових системах, таких як Google Ads, Bing Ads, спрямована на тих, хто активно шукає конкретні продукти чи послуги. Рекламодавці можуть використовувати ключові слова для підвищення видимості свого контенту.

3. Електронна пошта: Розсилання рекламних пропозицій та інформаційних бюлетенів через електронну пошту залишається ефективним методом зв'язку з цільовою аудиторією. Важливо створювати

персоналізовані та цікаві листи.

4. Відеореклама: Рекламні ролики на платформах, таких як YouTube, або відеореклама в мережах додатків та сайтів можуть привертати увагу через візуальність та звуковий ефект.

5. Партнерський маркетинг: Залучення партнерів та використання їхнього впливу на аудиторію для розповсюдження рекламних повідомлень.

6. Банерна реклама: Розміщення графічних банерів на веб-сайтах та мобільних додатках для залучення уваги.

7. Нативна реклама: Інтеграція рекламних повідомлень у контент таким чином, щоб вони виглядали природно та не виходили за межі формату платформи.

8. Реклама в месенджерах: Використання платформ, таких як WhatsApp, Facebook Messenger, для комунікації та розміщення реклами.

Комбінування різних інструментів дозволяє рекламодавцям створювати комплексні кампанії та максимально ефективно взаємодіяти з аудиторією.

Facebook Business Manager - це потужний інструмент для керування всіма аспектами маркетингу та реклами на платформі Facebook. Цей універсальний інструмент дозволяє створювати, аналізувати та керувати рекламою з різних акаунтів у централізованому інтерфейсі [5].

Основні можливості цієї платформи включають:

- Управління кількома об'єктами: Вона надає можливість керувати декількома рекламними акаунтами одночасно. Це означає, що ви можете вести рекламні кампанії для різних об'єктів та відстежувати їх результативність в одному місці.
- AI алгоритми для просування реклами: Facebook Business Manager використовує штучний інтелект для підвищення ефективності рекламних кампаній. Це означає використання алгоритмів машинного навчання та AI для оптимізації та автоматизації процесів реклами, що допомагає досягти кращих результатів у відповідь на різноманітні маркетингові цілі.

Рис. 1.7 Кількість користувачів Facebook [5]

1.4 Прибутковість рекламної компанії

Кожну годину в світі запускається безліч рекламних кампаній, кожна з яких прослідковує свою унікальну мету та стратегію. Одні з основних цілей, які вивчаються в рамках дипломного дослідження, стосуються реклами, орієнтованої на генерацію доходу.

Це важливий аспект в сучасному бізнес-середовищі, де конкуренція за увагу та гроші споживачів надзвичайно велика. Рекламні кампанії, спрямовані на забезпечення фінансового успіху, можуть приймати різні форми та використовувати різноманітні стратегії:

Основні показники для вимірювання ефективності контекстної реклами, якими користуються маркетологи:

1. **CTR (відношення переходів до показів)**: Цей показник вказує на відсоток користувачів, які перейшли за посиланням або виконали дію (наприклад, клікнули на рекламу) серед загальної кількості показів реклами. Високий CTR може свідчити про ефективність реклами в привертанні уваги аудиторії.

CTR - це первинна метрика, що демонструє співвідношення показів рекламного оголошення і переходів на нього. CTR допомагає зрозуміти, на яких майданчиках оголошення залучають більше переходів користувачів [14].Формула CTR показана на рисунку 1.8.

$$CTR = \frac{\text{кількість кліків}}{\text{кількість показів}} \times 100\%$$

Рис. 1.8 Формула CTR

Розглянемо ситуацію з рекламними показниками на двох різних ресурсах: на першому оголошення було показане 10 тисяч раз і отримало 100 переходів, а на другому - також 10 тисяч показів, але отримало 1000 переходів.

Хоча кількість переходів (CTR) на другому ресурсі вища (1000 переходів), це ще не гарантує високої конверсії або реального прибутку від реклами. Важливо враховувати, що високий обсяг переходів може не завжди перекладатися у реальні покупки чи конверсії.

Маркетологам також важливо аналізувати конверсійні показники, які вказують на реальний вплив реклами на покупки або бажані дії користувачів. Наприклад, кількість продажів або конверсія сайту може бути додатковими метриками для визначення успішності рекламної кампанії.

Отже, хоча показники переходів важливі для оцінки привертання уваги аудиторії, вони не завжди відображають реальний прибуток від реклами. Комбінування цих даних з конверсійними показниками може дати більш повний образ ефективності рекламної кампанії.

2. **CPA (ціна за дію)**: Це вартість, яку ви сплачуєте за певну дію, таку як конверсія (покупка товару, заповнення форми тощо). Низький CPA означає більшу ефективність рекламної кампанії, оскільки ви витрачаєте менше грошей на кожну конкретну дію.

3. **CPO (вартість замовлення)**: Цей показник визначає, скільки коштує замовлення або транзакція, здійснена через рекламну кампанію. Зниження CPO може свідчити про ефективніші та оптимізовані процеси покупок.

Показник CPO допомагає зрозуміти у скільки клієнту обходиться кожне замовлення. Формула CPO показана на рисунку 1.9. Наприклад, на контекст було витрачено 100 тисяч доларів, і в результатірекламної кампанії інтернет-магазин зробив тисячу продажів. Таким чином, вартість одного замовлення склала 100 доларів. Якщо ця вартість нижче ціни кожного проданого товару, то кампанія визнається ефективною [14]. При цьому, як правило, загальний показник CPO встановлюється для всіх замовлень з сайту.

$$CPO = \frac{\text{сума витрат на рекламу}}{\text{кількість підтверджених замовлень}} X\ 100\%$$

Рис. 1.9 Формула CPO

4. **ROAS (Return on Ad Spend)**: Це співвідношення між витратами на рекламу та доходом, отриманим в результаті цих витрат. Високий ROAS означає, що ви отримуєте більше прибутку від кожного витраченого долара на рекламу.

Це найголовніший показник в цій роботі, бо для нього і буде будуватися прогноз доходу від реклами для розуміння ефективності рекламних кампаній. ROAS, або повернення на рекламний витрати, це показник, що вимірює, наскільки кожен витрачений долар на рекламу приносить прибуток. Цей показник допомагає в оцінці ефективності рекламних кампаній: якщо ROAS більше 100%, це означає, що ви отримуєте більше прибутку, ніж витрачаєте на рекламу.

Наприклад, якщо ви порівняєте ROAS для кількох кампаній, ви можете визначити їхню успішність: якщо ROAS > 100%, кампанія прибуткова, якщо ROAS < 100%, вона не виправдала витрат. Це дає можливість розуміти, які маркетингові підходи працюють краще та які варто налаштувати чи відмінити.

Клікі, покази та конверсії корисні для аналізу рекламних кампаній, але вони не відображають економічну ефективність. Розглянути лише ці параметри може призвести до помилкових висновків, тому що вони не враховують прибутковість реклами. Наприклад, скорочення витрат на кампанію, яка приводить найменше трафіку, може призвести до великого зменшення продажів.

ROAS допомагає визначити, яка реклама приносить прибуток, і розуміти, які стратегії та інструменти працюють найкраще для вашої компанії. Ви можете перерозподілити бюджет, щоб збільшити прибуток без збільшення витрат.

Наприклад, якщо порівняти три кампанії з однаковим бюджетом, ROAS може показати, яка з них є більш прибутковою, що допомагає визначити, які кампанії доцільно розвивати або припинити.

Кампанії	Покази	Кліки	Витрати
Кампанія 1	5000	1000	$100
Кампанія 2	2500	100	$100
Кампанія 3	500	10	$100

Рис. 1.10 Аналіз кампаній без ROAS

Якщо робити висновки, спираючись тільки на «кількісні» показники, можна вирішити, що краще спрацювала Кампанія 1 - вона привела найбільшетрафіку. Але варто нам додати в таблицю дохід з ROAS - і картина зміниться.Таблиця з ROAS показана на рисунку 1.11.

Кампанії	Покази	Кліки	Витрати	ROAS
Кампанія 1	5000	1000	$100	100%
Кампанія 2	2500	100	$100	250%
Кампанія 3	500	10	$100	500%

Рис. 1.11 Аналіз кампаній з ROAS

Бачимо, що ROAS вище у Кампанії 3, значить, вона приносить більше доходу. Знаючи цю інформацію, ви можете внести корегування в свої кампанії, щоб швидше досягти поставлених цілей. Наприклад, якщо ви хочете підвищити пізнаваність бренду, більше інвестуйте в першу кампанію, а якщо хочете підвищити прибуток - виберіть третю кампанію.

$$ROAS = \frac{\text{дохід від реклами}}{\text{витрати на рекламу}} X\ 100\%$$

Рис. 1.12 Формула ROAS

1.5 Машинне навчання

Машинне навчання є захоплюючим полем, яке поєднує науку та мистецтво програмування комп'ютерів для того, щоб вони здатні були самостійно навчатися на основі накопичених даних. Одним із ключових технічних визначень цієї галузі було запропоноване Томом Мітчеллом у 1997 році.

Згідно з його визначенням, машинне навчання можна розглядати як процес, під час якого комп'ютерна програма постійно вдосконалює свою продуктивність у виконанні завдань, ґрунтуючись на здобутих досвідчених даних. Основною ідеєю є те, що програма вчиться і поліпшує свої навички без явного програмування з боку розробників.

Процес машинного навчання можна уявити як аналогію до того, як людина навчається на власних помилках та досвіді. Програма взаємодіє з великою кількістю даних, аналізує їх та вивчає корисні залежності та закономірності, які дозволяють їй ефективно вирішувати конкретні завдання. Цей підхід особливо корисний у випадках, коли завдання складне або важко формалізується традиційним програмуванням. Машинне навчання дозволяє створювати моделі, які адаптуються до змін у вхідних даних та навіть можуть прогнозувати майбутні події на основі раніше здобутого досвіду.

На рисунку 1.13 зображена загальна схема того, як відбувається машинне навчання

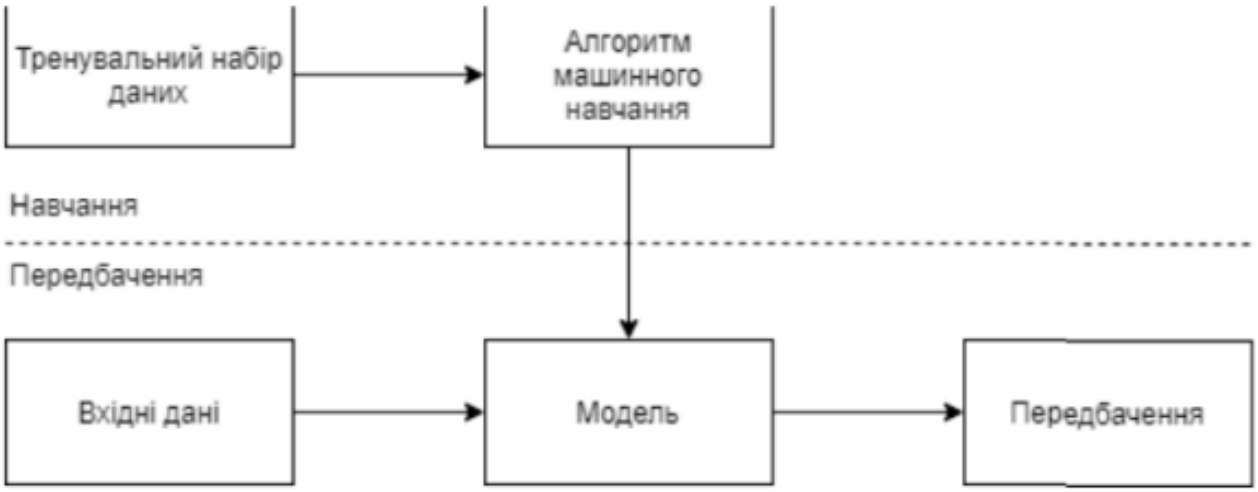


Рис. 1.13 Схема процесу машинного навчання

Таким чином, машинне навчання визначається не лише як технічна наука, але і як вид мистецтва, де програми розвивають свою інтелігенцію, аналізуючи та використовуючи величезні обсяги інформації для вирішення завдань ефективніше та точніше.

У розмаїтті сучасного машинного навчання можна виділити різноманітні варіації, які групуються за різними ключовими критеріями, що визначають їх функціональність та способи використання:

1. Тип навчання:

- Навчання з вчителем (Supervised Learning):

Моделі отримують дані, які мають явні мітки чи класифікації, і вони вчаться передбачати чи класифікувати нові, раніше невідомі дані.

- Навчання без вчителя (Unsupervised Learning):

Моделі вивчають без наявності міток чи класифікацій, шукаючи внутрішні патерни та взаємозв'язки в даних.

- Часткове навчання (Semi-supervised Learning):

Комбінує елементи обох, використовуючи як мітковані, так і немітковані дані для навчання моделі.

- Навчання з підкріпленням (Reinforcement Learning):

Моделі взаємодіють з середовищем, отримуючи відзнаки або штрафи відповідно до їхніх дій, і вчаться максимізувати кумулятивний виграш часом.

2. Режим навчання:

- Динамічне навчання (Online Learning):

Моделі можуть навчатися поступово, долучаючи нові дані по мірі їх надходження.

- Пакетне навчання (Batch Learning):

Вимагає пакетної обробки всіх даних перед навчанням, що може викликати великі обчислювальні витрати.

3. Підхід до вивчення даних:

- На основі зразків (Instance-based Learning):

Моделі навчаються на конкретних прикладах, зберігаючи інформацію про окремі екземпляри даних.

- На основі моделей (Model-based Learning):

Моделі намагаються знайти патерни або правила в даних, які можна використовувати для побудови прогностичних моделей чи порівняння нових даних з вже відомими прикладами.

Кожен з вказаних методів є унікальним і використовується в конкретних умовах, залежно від конкретної ситуації. Процес машинного навчання за допомогою вчителя включає в себе використання наперед визначеного алгоритму, який, протягом навчання, буде періодично сумніватися в своїх виборах і аналізувати лише правильні результати та ті, що схожі на них.

Для алгоритму самоорганізуючого навчання необхідно задати більше початкових параметрів відразу, щоб запустити процес прогнозування. Як тільки алгоритм зрозуміє поставлену ціль, він розпочне навчання, яке може тривати трошки довше, але буде більш ефективним. Однак у цього алгоритму є суттєвий недолік: після початку його роботи буде неможливо втрутитися ззовні, і при отриманні неправильних результатів з самого початку навчання їх не можна буде виправити. Це може призвести до накопичення помилок з самого початку роботи алгоритму, що призведе до невірних результатів.

Машинне навчання, або глибоке навчання, є підгалуззю штучного інтелекту, проте глибоке навчання виступає підтипом машинного навчання. Відмінність між

глибоким та машинним навчанням полягає у способі, яким алгоритми вчаться.

У глибокому навчанні значна частина процесу автоматизована для ефективного видалення властивостей, і для його реалізації не потрібно використовувати конкретні людські процеси. Цей метод дозволяє використовувати великі обсяги даних [6].

Глибоке навчання може ефективно обробляти величезну кількість неструктурованих даних, на обробку яких людині потрібно багато років власної праці. Такий підхід дає можливість працювати з нелінійними процесами прийняття рішень [6].

Рис. 1.14 Класифікація методів машинного навчання

Ця різноманітність підходів і типів машинного навчання дозволяє вибирати оптимальний метод для конкретної задачі та оточення, враховуючи особливості доступних даних і поставлених завдань.

Навчання з вчителем

В роботі розглядається цей тип машинного навчання. При навчанні з вчителем навчальні дані, що поставляються алгоритму, включають бажані

рішення, названі мітками (label).

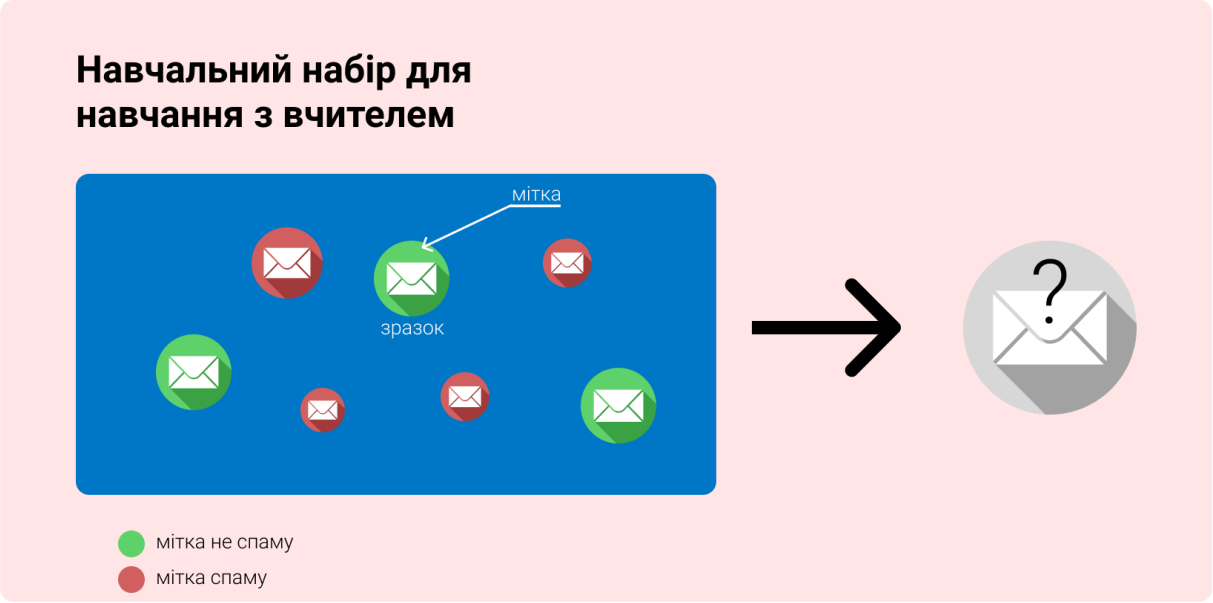


Рис. 1.15 Позначений навчальний набір для навчання з вчителем

Типові завдання навчання з вчителем включають класифікацію та регресію.

Класифікація є завданням машинного навчання, де модель намагається вивчити патерни та закономірності, щоб розподіляти дані на конкретні класи або категорії. Цей підхід заснований на використанні прикладів даних, які мають відомі мітки чи класифікації. Одним із найпоширеніших прикладів класифікаційних завдань є фільтр спаму, який ефективно відокремлює повідомлення на "спам" та "не спам".

У цьому контексті, фільтр спаму використовується для автоматизованого визначення того, чи є конкретне повідомлення небажаним спамом чи ж поваги варте. Процес навчання моделі полягає в аналізі великої кількості попередньо класифікованих повідомлень – спаму та звичайних. Модель вивчає характеристики, які вказують на те, до якого класу належить кожне повідомлення, і використовує ці знання для того, щоб автоматично класифікувати нові повідомлення в майбутньому.

Такий підхід дозволяє фільтру спаму адаптуватися до змінних форм та характеристик спам-повідомлень, що допомагає підтримувати високу точність в

роботі фільтра. Крім того, важливо зауважити, що класифікація в контексті фільтра спаму є лише одним із безлічі застосувань цього машинного навчання, яке успішно використовується у різних областях для автоматизації визначення класів чи категорій у невідомих даних.

Іншим стандартним завданням є регресія, яка має на меті передбачення числових значень на основі набору характеристик. Наприклад, у випадку прогнозування доходів від рекламних кампаній, використовуючи параметри, такі як бюджет та кількість переходів, застосовується метод регресії. Щоб навчити модель, необхідно подати набір прикладів рекламних кампаній, разом із заздалегідь відомими доходами, щоб система могла вивчити залежності та здатна була передбачати нові значення [6].

У цьому випадку, модель навчається аналізувати взаємозв'язки між різними параметрами рекламних кампаній та їхнім фінансовим результатом. Зазначені характеристики, такі як бюджет та кількість переходів, служать визначальними факторами для передбачення доходу. Цей підхід дозволяє системі аналізувати велику кількість даних і вивчати закономірності, що дозволяє ефективно прогнозувати доходи рекламних кампаній на основі нових вхідних даних.

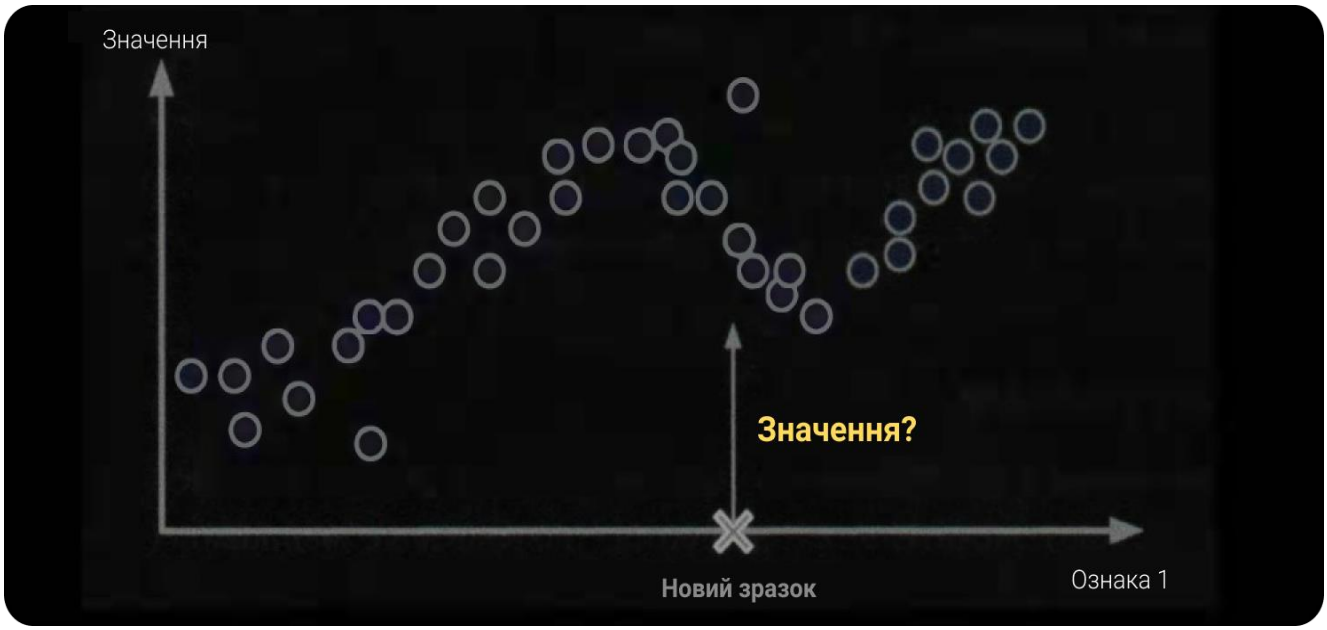


Рис. 1.16 Процес регресії

Виділимо найбільш важливі алгоритми навчання з вчителем [6]:

- k найближчих сусідів (k-nearest neighbors);
- лінійна регресія (linear regression);
- логістична регресія (logistic regression);
- метод опорних векторів (Support Vector Machine - SVM);
- дерева прийняття рішень (decision tree) і випадкові лісу (randomforest);
- нейронні мережі (neural network).

1.6 Основні проблеми при розробці програмного продукту, що використовує машинне навчання

Недостатній розмір навчальних даних

Недостатній розмір навчальних даних є серйозною проблемою при розробці програмного продукту, який використовує машинне навчання (МН). Ця проблема виникає, коли обсяг доступних даних для тренування моделі є обмеженим або недостатнім для того, щоб алгоритм мав достатню кількість інформації для ефективного вивчення та генералізації.

Основні проблеми, пов'язані з недостатнім розміром навчальних даних, включають:

- Перенавчання;
- Низька точність та стабільність;
- Неможливість вирішення складних завдань;
- Чутливість до вхідних змін;
- Важкість у вдосконаленні.

Нерепрезентативні навчальні дані

Проблема зміщення вибірки виявляється надзвичайно важливою в контексті навчання алгоритмів, особливо коли наявні навчальні дані недостатньо

відображають всю різноманітність нових прикладів, на яких ми хочемо застосувати модель. Це стає ключовим фактором як для навчання на основі зразків, так і для модельного навчання.

Один із визначених випадків зміщення вибірки відбувся під час президентських виборів у США в 1936 році, коли Literary Digest проводив опитування. Зазначений журнал відправив листи приблизно 10 мільйонам осіб, але отримав лише 2.4 мільйона відповідей. На підставі цих даних передбачили перемогу Лендона з 57% голосів, однак фактично переміг Рузвельт із 62%.

Ця помилка була обумовлена методом вибірки, використаним Literary Digest:

• По-перше, для отримання адрес використовувалися телефонні довідники, журнальні переліки та інші джерела, що включали в себе більш заможний клас людей, що спричинило перевагу голосів за Лендона.

• По-друге, менше 25% отримувачів листів відповіли на опитування, що виключило певні групи та спотворило результати через відсутність відповідей.

Ці приклади наочно демонструють, що навчальні дані повинні бути репрезентативними та відображати різноманітність вхідних даних для ефективного навчання моделі та точних передбачень на нових прикладах. Потрібно враховувати це зміщення вибірки та вживати заходів для забезпечення адекватного відображення різноманітності у навчальних даних.

Дані поганої якості

Виправлення помилок та обробка навчальних даних є ключовим етапом у підготовці для навчання моделей машинного навчання. Коли дані мають помилки, це може негативно позначитися на роботі моделі, тому важливо вирішити ці питання перед навчанням.

Ось кілька можливих заходів:

1. **Видалення викидів**: Якщо деякі приклади мають аномальні значення або помилки, вони можуть спотворити роботу моделі. Прибрати ці викиди або виправити помилки вручну.

2. **Робота з відсутніми ознаками**:
3. **Ігнорування ознаки**: Іноді можна просто не використовувати ознаку, яка має багато відсутніх значень.
4. **Видалення прикладів з пустими ознаками**: Якщо обробка пустих значень не можлива, можна видалити приклади з такими ознаками.
5. **Заповнення пропусків**: Можна використовувати різні стратегії для заповнення пропущених значень, наприклад, середнє або медіанне значення, значення на основі сусідніх даних і т. д.
6. **Навчання моделей з різними підходами до ознак**: Можна спробувати тренувати моделі, виключаючи та включаючи певні ознаки, щоб оцінити вплив відсутніх даних на результати.

Ці заходи можуть допомогти у покращенні якості даних та забезпечити більш точне навчання моделей.

Несуттєві ознаки

Процес вибору та створення ознак - це ключовий етап підготовки даних для машинного навчання, оскільки від цього залежить якість та ефективність моделі. Ось деякі аспекти цього процесу:

1. **Вибір ознак**: Це включає в себе відбір найважливіших ознак для навчання моделі. Це може стосуватися зменшення кількості ознак до тих, що найбільше впливають на результат, або відбір тільки тих ознак, які мають значущий внесок.

2. **Виділення ознак**: Цей етап полягає у створенні нових ознак на основі існуючих шляхом їх комбінації або перетворення. Це може включати в себе створення нових характеристик на основі відомих, таких як відношення, різниця, сума, або агрегація інформації з кількох ознак для отримання більш інформативних.

3. **Створення нових ознак зі збору даних**: Це може означати збір нової інформації або ознак, які допоможуть покращити передбачення моделі. Це може бути додатковими даними з нових джерел або збір інформації з нових джерел для покращення якості моделі.

Цей процес допомагає підготувати дані для моделей машинного навчання, роблячи їх більш інформативними та готовими до ефективного навчання.

1.7 Великі дані, їх види та обробка

Великі дані (Big Data) представляють собою обсяги інформації величезних розмірів, які перевищують можливості традиційних методів зберігання, обробки та аналізу даних[11].

Об'єм даних великих даних означає, що маємо справу з великою кількістю інформації, яка може варіюватися від кількох терабайт до зеттабайт та більше. Це включає в себе дані з різних джерел, такі як соціальні мережі, датчики, транзакції та інші.

Швидкість стосується темпу, з яким генеруються та обробляються дані. Великі дані часто надходять в реальному часі, вимагаючи швидкої обробки та аналізу для прийняття рішень. Різноманітність вказує на різноманіття типів даних, які входять в великі дані. Це можуть бути тексти, зображення, відео, аудіо, графи, структуровані та неструктуровані дані.

Обробка великих даних вимагає використання спеціалізованих технологій, таких як розподілені системи зберігання, техніки масштабованої обробки даних, аналітичні алгоритми та інструменти для візуалізації результатів. Великі дані використовуються в різних галузях, включаючи бізнес, медицину, науку, технології та інші, для отримання нових інсайтів, прогнозування тенденцій та підтримки прийняття рішень.

Поняття Big Data характеризується трьома основними вимірюваннями, що входять в "3V" - об'єм, швидкість та різноманітність. Об’єм «Volume» - це велика кількість даних, які описані за їх розміром. Різноманітність «Variety» визначає певні типи, за якими поділяються дані, тобто чи наявні у виборці певні форми даних. Швидкість «Velocity» швидкість з якою проходить генерація даних та аналіз даних. Достовірність «Veracity» відповідає за вибір надійних даних. При погіршенні записаних даних, може погіршуватись точність аналізу. Змінливість

«Variability», якщо буде виникати неузгоджена інформація, то вона буде ускладнювати, а іноді й заплутувати процеси обробки та управління даними.

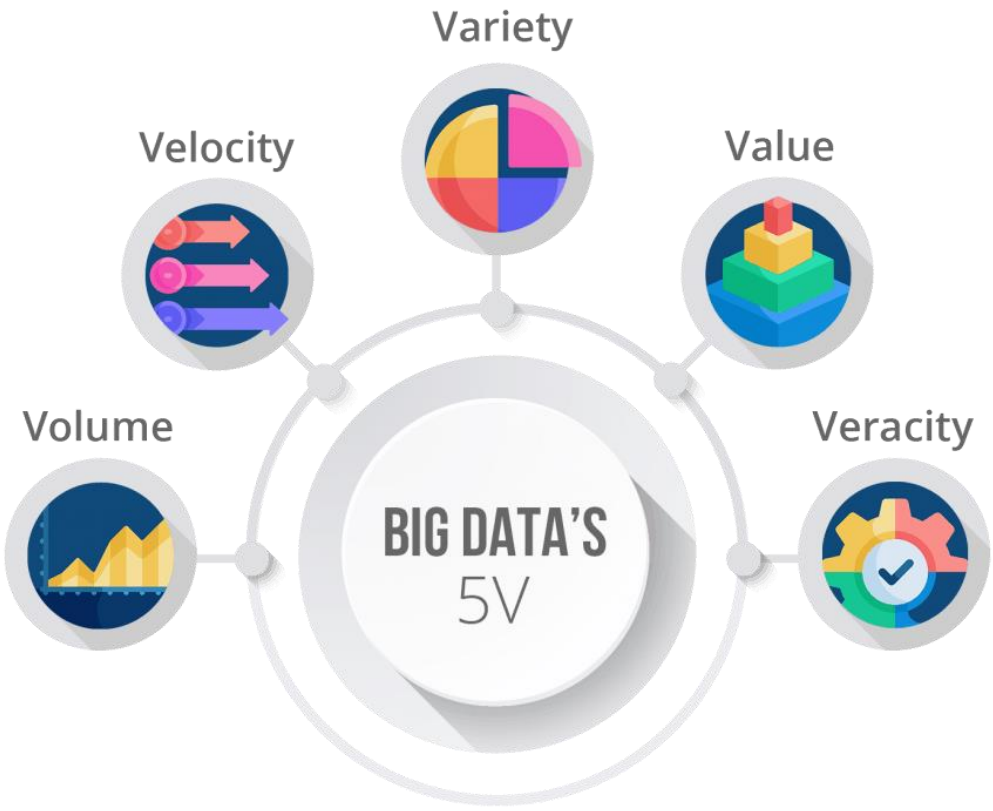


Рис. 1.17 Виміри Big Data

Використання сучасних технологій Big Data на сьогоднішній день є широко поширеним та важливим напрямом розвитку. За допомогою різноманітних видів інформації, користувач отримує можливість вивчати інтереси своїх клієнтів, аналізувати ефективність маркетингових стратегій та проводити комплексний аналіз ризиків у проекті.

Big Data відкриває безліч можливостей для підприємств у різних сферах. Важливо визначити, як використовуючи ці технології, можна вдосконалити бізнес-процеси та приймати обґрунтовані рішення.

Зокрема, збір та аналіз великих обсягів даних дозволяє відстежувати поведінку клієнтів і надавати персоналізовані послуги. За допомогою цих

технологій можна не лише виявити потреби споживачів, але й адаптувати продукти чи послуги під їхні вимоги.

Маркетингові стратегії стають більш ефективними завдяки здатності аналізувати великі масиви даних щодо реакцій споживачів на конкретні заходи. Це дозволяє компаніям покращувати свої комунікаційні стратегії, точно визначати цільову аудиторію та забезпечувати успішні результати в сфері маркетингу.

Узагальнюючи, використання технологій Big Data перетворює спосіб, яким компанії взаємодіють з інформацією та вирішують стратегічні завдання. Це забезпечує конкурентну перевагу та допомагає досягати нових висот у бізнесі.

2. МОДЕЛЮВАННЯ МЕТОДУ ПРОГНОЗУВАННЯ ЕФЕКТИВНОСТІ РЕКЛАМИ

2.1 Різновид математичних моделей

Математичні моделі в програмному забезпеченні використовуються для опису поведінки системи або алгоритму через математичні вирази, що можуть бути обчислені комп'ютером. Ось декілька прикладів:

1. **Моделі баз даних**: Математичні моделі даних визначають, як дані організовані і зберігаються в базі даних. Вони можуть включати схеми, таблиці, відношення між даними і способи доступу до них.
2. **Алгоритмічні моделі**: Це моделі, які використовують математичні концепції для опису роботи алгоритмів. Наприклад, моделі сортування, пошуку або оптимізації, які описують, як алгоритми працюють на рівні кроків чи процесів.
3. **Моделі програмних інтерфейсів (API)**: Інколи API описуються за допомогою математичних моделей, щоб уточнити, як системи або компоненти взаємодіють один з одним.
4. **Математичне моделювання алгоритмів та структур даних**: Це включає в себе математичні моделі, які описують ефективність алгоритмів та структур даних у різних умовах, таких як швидкість виконання, складність або використання ресурсів.

Математичні моделі та методи програмування можна класифікувати за різними ознаками.

За кількістю цільових функцій моделі поділяються на однокритеріальні та багатокритеріальні, які мають дві чи більше цільових функцій.

Залежно від особливостей явища, моделі можуть бути детермінованими, стохастичними чи містити елементи невизначеності.

Детерміновані моделі передбачають жорсткі функціональні зв'язки між змінними та параметрами. В таких моделях інформація вважається однозначною та

достовірною.

Стохастичні моделі - це моделі, де змінні представлені випадковими величинами з відомими розподілами та статистичними характеристиками.

До стохастичних моделей належать моделі стохастичного програмування, теорії випадкових процесів та теорії масового обслуговування.

Для моделювання ситуацій, де фактори невідомі або неможливо зібрати статистичні дані, використовуються моделі з елементами невизначеності: теорія ігор та імітаційні моделі.

Даний матеріал охоплює лише детерміновані моделі, де не розглядається вплив випадкових подій на досліджувані показники. Незважаючи на здавалося б простоту таких моделей, до них можна зведувати багато практичних завдань, включаючи більшість економічних. Залежно від виду цільової функції Z та функцій i(x), що входять в систему обмежень, моделі математичного програмування можуть належати до лінійного або нелінійного програмування.

Лінійне програмування - це розділ математичного програмування, який застосовується для розробки методів пошуку екстремуму (максимуму чи мінімуму) лінійних функцій багатьох змінних за лінійних обмежень. За типом задач його методи можна розділити на універсальні та спеціальні. Універсальні методи (наприклад, симплекс-метод) можуть вирішувати будь-які задачі лінійного програмування. Спеціальні методи враховують особливості цільової функції та системи обмежень. Методи лінійного програмування широко використовуються в промисловості для оптимізації виробничої програми, планування грузопотоків, розрізання матеріалів, вибору технологій тощо.

Якщо хоча б одна з функцій Z (цільова функція) чи (i) (функції, що входять у систему обмежень) є нелінійною за керованими параметрами, то маємо задачу нелінійного програмування. Якщо цільова функція такої задачі є опуклою, а область допустимих рішень також є опуклим множиною, то говорять про задачу опуклого програмування. Методи опуклого програмування використовуються при розв'язанні завдань розрахунку оптимальної партії випуску деталей, управління складними поставками та запасами, розподілі обмежених ресурсів і т.д.

Нелінійне програмування - це галузь математичного програмування, де цільова функція або обмеження містять нелінійні складові. Вони можуть бути використані для вирішення широкого спектру завдань, від оптимізації складних виробничих процесів до управління ресурсами та розв'язання фінансових задач.

Однією з ключових галузей нелінійного програмування є методи оптимізації, які ставлять за мету знаходження мінімуму або максимуму нелінійних цільових функцій при врахуванні різноманітних обмежень. Ці методи можуть бути використані для оптимізації процесів, що мають складні функціональні залежності або коли варіанти впливу на цільову функцію є нелінійними.

Зокрема, галузь оптимізації може застосовуватися в економіці для максимізації прибутку чи мінімізації витрат у складних економічних системах, у виробництві для планування оптимальних виробничих програм, в інженерії для проектування оптимальних конструкцій, а також в багатьох інших галузях.

Крім того, методи нелінійного програмування застосовуються в аналізі ризиків та прийнятті рішень. Вони дозволяють враховувати складні взаємозв'язки та неоднорідність у вирішенні проблем, що допомагає у прийнятті кращих рішень в умовах невизначеності.

Ці методи знаходять своє застосування в багатьох сферах, де неоднорідність, нелінійність і складність є ключовими аспектами у прийнятті оптимальних рішень чи у вирішенні завдань оптимізації.

Математичні моделі у методах машинного навчання - це формалізовані математичні структури, які використовуються для опису та передбачення залежностей у навчальних даних. Вони є основою алгоритмів машинного навчання та дозволяють програмам самостійно вчитися з даних, розпізнавати патерни та робити прогнози для нових даних.

У методах машинного навчання використовуються різні типи математичних моделей, такі як:

1. **Лінійні моделі**: Вони базуються на лінійних залежностях між вхідними факторами та цільовою змінною. Це можуть бути моделі регресії або класифікації, які працюють на основі лінійних комбінацій вхідних ознак.

2. **Дерева рішень та ансамблі**: Вони представляють собою структури, які розгалужуються в залежності від вхідних ознак і приймають рішення на основі розділень даних.

3. **Нейронні мережі**: Це моделі, інспіровані біологічною нейронною системою. Вони складаються зі штучних нейронів, що обробляють інформацію та передають сигнали через шари для вирішення завдань класифікації або регресії.

4. **Методи глибинного навчання**: Це підклас нейронних мереж, які мають багато шарів та вміють автоматично вивчати корисні представлення даних для рішення складних завдань.

5. **Методи групування та кластеризації**: Вони використовуються для групування подібних об'єктів у класи або кластери.

Ці моделі базуються на математичних концепціях, таких як оптимізація, ймовірність, алгебра, статистика та інші, і вони застосовуються для розв'язання різноманітних задач, від прогнозування до розпізнавання образів та прийняття рішень.

2.2 Математична модель прогнозування

Математичний підхід до прогнозування є фундаментальною основою для розробки математичних моделей та алгоритмів. Для аналізу можливих варіантів прогнозування необхідно вміти вирішувати це завдання, використовуючи математичний підхід до процесу. Використання математики у прогнозуванні визначається як ключовий елемент при створенні відповідних систем. Для більшості завдань, пов'язаних з прогнозуванням, важливо усвідомити основну математичну модель, від якої відбувається розгалуження і модифікація всіх інших моделей.

Основну модель прогнозування можна виразити наступним чином:

$$y = mx + n \tag{2.1}$$

де «у» це кінцевий результат при прогнозуванні.При прогнозуванні на великій вибіркі даних, таких кінцевих результатів буде багато, тому що для аналізу береться велика кількість параметрів. Тому для початку потрібно розділити кожен параметр, а тільки після такої обробки можна перейти до їх об’єднування[15]. Змінна «х» відповідає за певний проміжок часу. При цьому отримуються чіткі параметри для алгоритму прогнозування. Параметр «m» відповідає за нахил прогнозу.

Загалом, існує чотири основні методи математичного прогнозування, які включають в себе:

- Пряму лінію;
- Ковзке середнє;
- Просту лінійну регресію;
- Множинну лінійну регресію.

Рис. 2.1 Різновид методів прогнозування за ступенем формалізації

Метод ковзкої середньої передбачає використання зваженого підходу до

визначення значень за обраними періодами. Аналіз проводиться на основі показників, які відомі з минулого. У контексті прогнозування погоди важливо враховувати постійні факти, такі як негативна температура повітря взимку та позитивна влітку. В цьому методі кожна наступна реляційна модель має подібність до попередньої.

Прямолінійний метод прогнозування визначається як найбільш простий підхід, що ґрунтується на лінійній залежності від вхідних параметрів. Цей метод застосовується тільки у випадках, коли очікувані значення мають тенденцію накопичення, тобто постійно зростають або спадають. Використання даного методу до такої сфери як реклама, де дані є мінливими є не нерезультативним.

Множинна лінійна регресія для прогнозування погоди ефективно використовує дві чи більше незалежних змінних для створення прогнозу. В цьому методі взаємодія зі статистичними параметрами стає важливим елементом, а для їхнього прорахунку найчастіше застосовується метод середнього квадрата:

$$\sigma = \sqrt{\frac{\sum_{j-1}^{n}(x_j - x)^2}{n}} \tag{2.2}$$

Цей підхід дозволяє враховувати взаємодію різних змінних та їх вплив на прогнозування погодних умов. Користуючись множинною лінійною регресією, можна отримати більш точні та надійні прогнози, що враховують різноманітні аспекти, які можуть впливати на погодні явища.

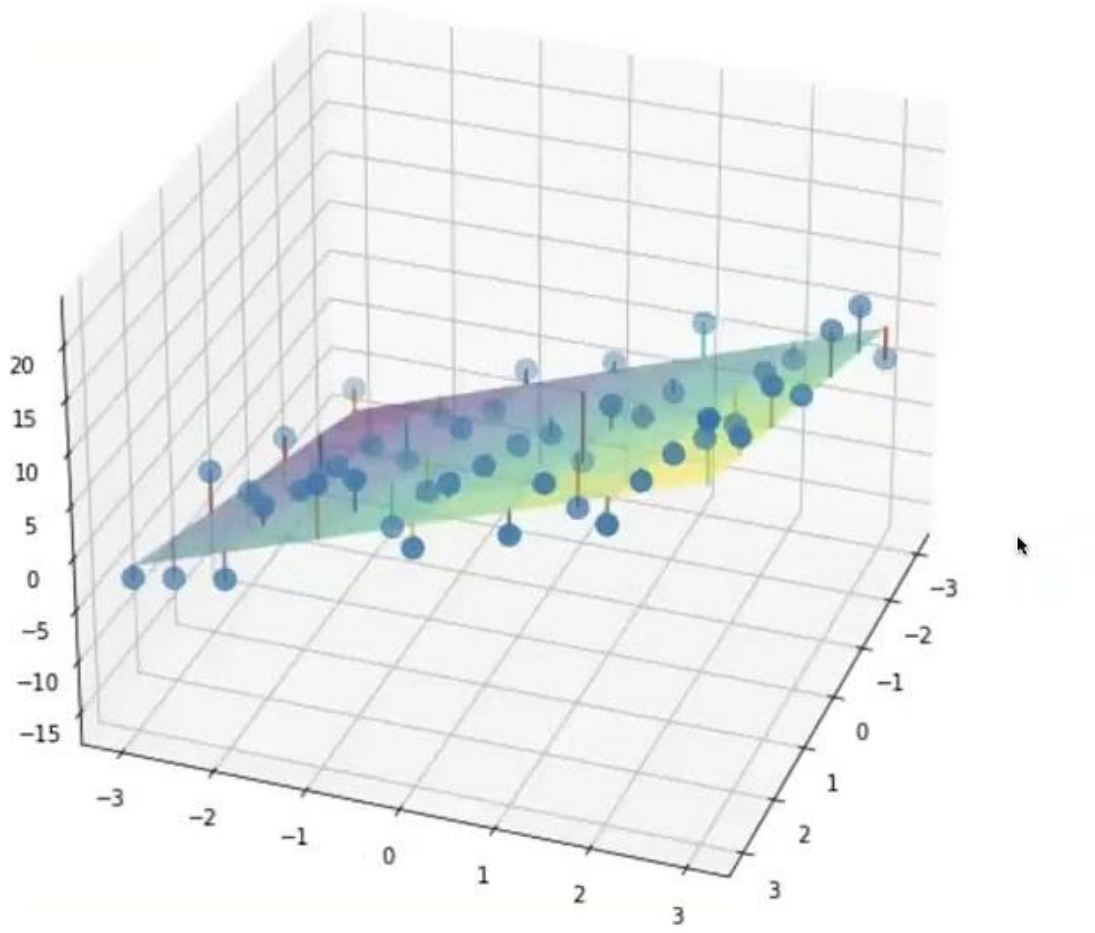


Рис. 2.2 Множинна лінійна регресія. Візуальне відображення

Існує велика різноманітність методів прогнозування, проте їхня основна схожість полягає в тому, що вони спрямовані на одну ціль - створення прогнозу. Таким чином, методи прогнозування поділяються на два основних типи: інтуїтивні та формалізовані.

Для отримання надійного прогнозу в умовах змінних, які постійно змінюються, критично використовувати побудову графіків з кривими лініями. Вони більш точно відображають середнє значення в конкретний період часу. Це особливо важливо в контексті прогнозування на даних, що мають мінливу природу. Лінійний прогноз не може забезпечити точний результат, якщо йому надаються лише отримані значення у вигляді прямої лінії зростання або спадання.

Наприклад, навіть невелика зміна в декількох високих значеннях може значно вплинути на графік, якщо вони розташовані поруч, що призводить до значних коливань. Таким чином, для зональної регресії використовується наступна

формула:

$$D = \sum_i (f1(i) - yo_i)^2 \quad (2.3)$$

Для більш детального аналізу необхідно враховувати середню вагу. У випадку симетричної частоти градацій важливо враховувати значущі ваги. Для цього слід визначити максимальні та мінімальні значення для кожного випадку прогнозування та розрахувати середню похибку, орієнтуючись на всі отримані результати.

$$I = \frac{x_{max} - x_{min}}{n} \quad (2.4)$$

де n – кількість отриманих результатів.

Моделювання вплинуло на маркетинг як ніколи раніше, перетворюючи стратегії та підходи до реклами та споживчої поведінки. Завдяки сучасним алгоритмам машинного навчання, маркетологи можуть аналізувати величезні обсяги даних щодо покупців, їхніх переваг та поведінкових тенденцій. Це дозволяє точно налаштовувати рекламні кампанії та персоналізувати комунікацію, забезпечуючи споживачам продукти та послуги, які відповідають їхнім індивідуальним потребам.

При використанні моделювання у маркетингових стратегіях збільшується точність прогнозів, що сприяє ефективній взаємодії з аудиторією та покращенню стратегій впливу. Крім того, моделі дозволяють виявляти нові можливості для росту та інновацій, створюючи простір для кращого розуміння ринкових динамік і вдосконалення рішень у маркетинговому середовищі.

2.3 Методи прогнозування в машинному навчанні

Методи прогнозування машинного навчання є необхідним інструментарієм для аналізу та передбачення подій у різних сферах людської діяльності. Вони здатні

виявляти складні залежності в масивах даних, враховуючи різноманіття та динаміку інформації. Ці методи використовуються для вирішення ряду завдань та надають вагомий внесок в процеси прийняття рішень.

У сучасному економічному середовищі прогнозування має ключове значення для фірм, організацій і держав. Методи машинного навчання використовуються для аналізу фінансових ринків, прогнозування попиту на товари та послуги, а також для стратегічного планування. Вони дозволяють оцінити ефективність бізнес-процесів, а також побудувати оптимальні стратегії розвитку, враховуючи динаміку змін в економіці.

В медицині методи прогнозування машинного навчання використовуються для діагностики захворювань та прогнозування рівня ризику для пацієнтів. Аналіз клінічних даних та виявлення взаємозв'язків дозволяють розробляти персоналізовані підходи до лікування, збільшуючи ефективність медичних послуг та покращуючи результати лікування.

У сфері транспорту та логістики методи прогнозування машинного навчання допомагають вирішувати завдання, пов'язані з управлінням транспортними потоками та оптимізацією маршрутів. Це сприяє підвищенню ефективності доставки товарів, зменшенню витрат та поліпшенню якості обслуговування.

В сільському господарстві методи прогнозування застосовуються для прогнозування врожаїв та управління сівозмінами. Вони дозволяють аграріям ефективно реагувати на зміни у природних умовах та максимізувати виробництво сільськогосподарської продукції. В інших сферах, таких як наука, освіта, технології та багато інших, методи прогнозування машинного навчання відкривають нові можливості для дослідження, оптимізації процесів та прийняття обгрунтованих рішень. Завдяки їхній потужності аналізу та передбачення, вони стають необхідним інструментом в епоху великих обсягів даних та штучного інтелекту.

Рис.2.3 Схема проведення навчання моделі

Машинне навчання включає в себе ряд математичних методів для створення прогностичних моделей, спрямованих на вирішення різноманітних завдань у різних галузях.

Основні методи та їх характеристики:

- **Лінійна Регресія** - це метод, який моделює лінійну залежність між вхідними змінними і вихідною змінною.
- **Випадковий Ліс (Random Forest)**: Випадковий ліс є ансамблем рішаючих дерев. Кожне дерево вирішує частину проблеми, і результати об'єднуються для отримання більш точного прогнозу. Математично, випадковий ліс базується на багатьох різних деревах рішень, які голосують за результат.
- **Нейронні Мережі**: мережі моделюють взаємозв'язки, навчаючись вагам між штучними нейронами. Вони складаються з входів, ваг, функцій активації та виходів. Тренування полягає у налаштуванні ваг для мінімізації функції втрат.
- **Градієнтний Бустінг**: Градієнтний бустінг поєднує послідовні слабкі моделі

для створення сильної. Він минулого додає до нового, коригуючи помилки. Математично, в цьому методі використовується градієнт функції втрати для покращення моделі.

Кожен з цих методів має свої сильні та слабкі сторони, і їх вибір залежить від конкретного завдання та характеристик даних. Машинне навчання відкриває широкі можливості для прогнозування, дозволяючи вирішувати завдання в реальному часі та визначати закономірності у великих обсягах даних.

2.4 Лінійна регресія

Лінійна регресія представляє собою метод у машинному навчанні, що застосовується для визначення зв'язку між двома різними ознаками або змінними. У контексті лінійної регресії розглядаються два основних типи змінних: залежна та незалежна. Незалежна змінна є самостійною та не залежить від інших змінних. Зміни у незалежній змінній призводять до коливань у значеннях залежної змінної. Залежна змінна є тією, яка є об'єктом дослідження, і яку модель лінійної регресії намагається передбачити.

Для кращого розуміння лінійної регресії давайте розглянемо конкретний приклад та ознайомимося з формулою, яку вона використовує. Припустимо, у нас є набір даних, що включає розміри жорстких дисків та їхню вартість. В цьому контексті дві змінні - обсяг пам'яті та вартість - розглядаються як незалежна та залежна змінні відповідно. Наприклад, можемо припустити, що збільшення обсягу пам'яті призводить до зростання вартості. Графічно це може бути представлено точковою діаграмою.

Рис. 2.4

Лінійна регресія використовує математичний підхід для побудови лінії (регресійної лінії), яка найкращим чином відображає відносини між незалежною та залежною змінною. Ця лінія може служити інструментом прогнозування вартості (залежної змінної) на основі значень обсягу пам'яті (незалежної змінної).

Точне відношення між обсягом пам'яті та вартістю може варіюватися в залежності від виробника та моделі жорсткого диска. Однак загальна тенденція даних може бути описана як рух від нижнього лівого кута (де жорсткі диски є дешевшими та мають менший обсяг) до верхнього правого кута (де накопичувачі дорожчі та мають більший обсяг).

Якщо ми представимо обсяг пам'яті на вісі X, а вартість на вісі Y, то лінія, що відображає зв'язок між цими змінними, стартує з нижнього лівого кута і просувається вгору та праворуч. Це вказує на те, що із зростанням обсягу пам'яті вартість також збільшується. Такий рух лінії відповідає загальній тенденції у даному контексті, де обсяг пам'яті та вартість взаємопов'язані.

Рис. 2.5

Функція регресійної моделі спрямована на визначення лінійного взаємозв'язку між змінними X і Y, що найкраще відображає залежність між ними. У випадку лінійної регресії передбачається, що значення Y можна виразити як певну комбінацію вхідних змінних X. Взаємозв'язок між вхідними змінними (X) та цільовими змінними (Y) може бути визначений лінією, яка проходить через точки на графіку. Ця лінія представляє функцію, що найкраще описує зв'язок між X і Y (наприклад, при збільшенні X на 3, Y збільшується на 2). Основна мета полягає в пошуку оптимальної "лінії регресії" або функції, яка найкраще підходить до наведених даних.

Лінії регресії часто виражаються рівнянням:

$$Y = m * X + b \tag{2.5}$$

Тут X вказує на незалежну змінну, Y - на залежну. Параметр m визначає нахил лінії, a b - величину "підйому" над "пробігом". У контексті машинного навчання використовується інше рівняння для визначення нахилу, яке виглядає наступним чином: y(x) = w0 + w1 * x. Тут y - цільова змінна, w - параметри моделі, а x - вхідні дані. Рівняння означає: "Функція, яка визначає Y в залежності від X, дорівнює

параметрам моделі, помноженим на вхідні дані". Під час навчання параметри моделі коригуються для досягнення оптимальної лінії регресії.

2.5 Випадковий ліс

Випадковий ліс представляє собою сукупність дерев рішень, які об'єднуються для отримання єдиного агрегованого результату. Цей метод має широкий спектр використання в області класифікації та регресійного аналізу даних і часто розглядається як один із найточніших алгоритмів навчання. Він застосовується в різних галузях, включаючи банківську сферу, маркетинг та продажі, для прогнозування цін і інших параметрів.

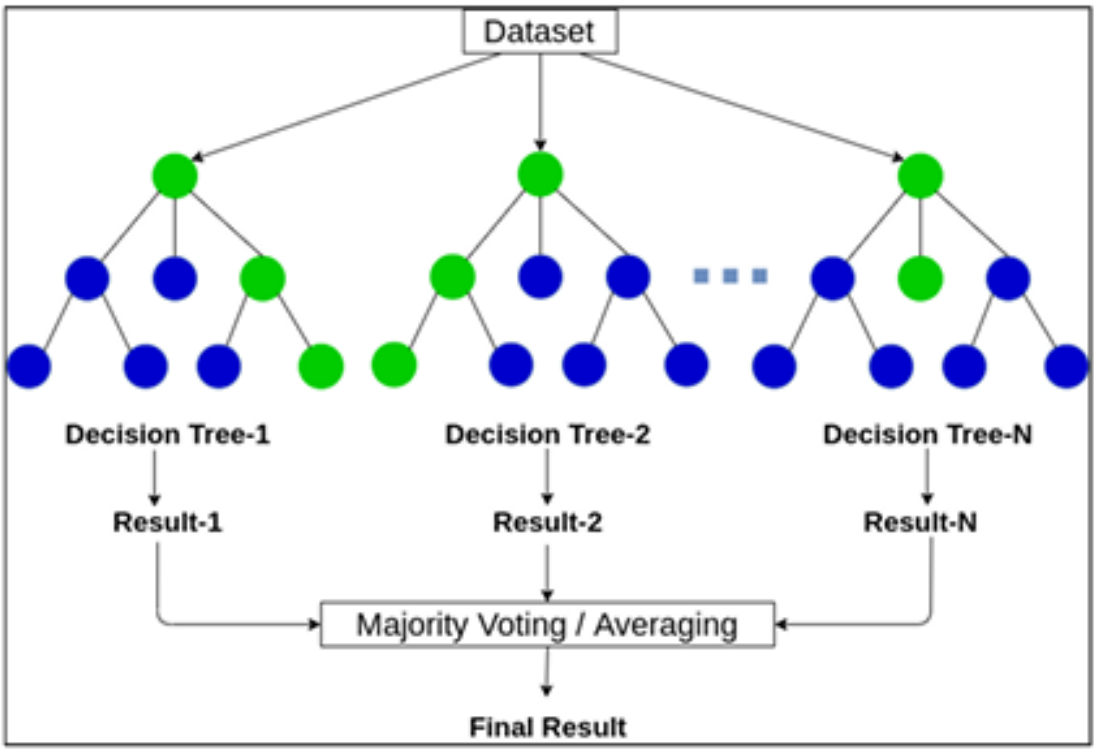


Рис. 2.6 Схема випадкового лісу

Алгоритм випадкового лісу використовує техніку ансамблевого навчання, що об'єднує багато класифікаторів для рішення різноманітних завдань. Його основою є набір дерев рішень, які навчаються через пакетування або агрегацію. Для підвищення точності використовується ансамблевий метаалгоритм, такий як

Bagging.

Алгоритм визначає результат на основі передбачень дерев рішень, використовуючи середнє значення отриманих результатів. Збільшення кількості дерев сприяє вищій точності результату, але може призвести до повільного процесу. Випадкові ліси зменшують дисперсію у деревах рішень за рахунок різноманітності прикладів, випадкового вибору ознак та комбінування невеликих (shallow) дерев.

Дерева рішень є основною складовою (блоками) алгоритму випадкового лісу. Дерево рішень - це техніка прийняття рішень, яка формує деревоподібну структуру. Дерево рішень складається з трьох компонентів: вузлів рішень, листових вузлів і кореневого вузла. Алгоритм дерева рішень ділить навчальний набір даних на гілки, які далі поділяються на інші гілки. Ця послідовність триває до тих пір, поки не буде досягнуто листовий вузол. Листковий вузол не може бути відокремлений далі [8].

Вузли в дереві рішень представляють атрибути, які використовуються для прогнозування результату. Вузли рішень забезпечують посилання на листи.

Одне дерево є слабким предиктором, проте його швидкість побудови компенсує цю обмеженість. Більше дерев додають надійності моделі і запобігають перенавчанню, але це також призводить до повільнішого процесу. Зменшення набору функцій може ефективно прискорити процес випадкового лісу.

Алгоритм випадкового лісу усуває обмеження використання одного дерева, це зменшує переобладнання наборів даних і також підвищує точність отриманих результатів. Алгоритм роботи представлений на рисунку 2.7:

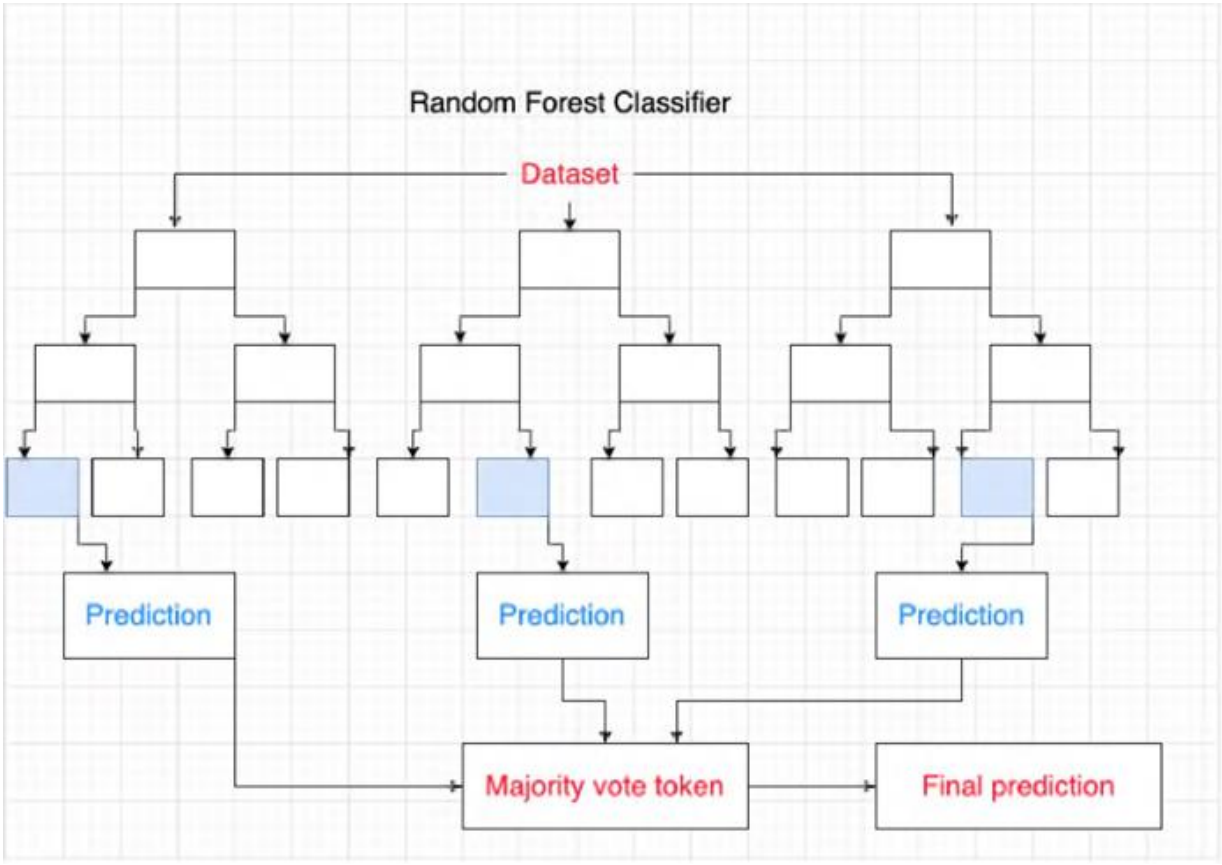


Рис. 2.7 Алгоритм роботи випадкового лісу

Особливості алгоритму Випадковий ліс:

- висока точність;
- забезпечує ефективний спосіб обробки відсутніх даних;
- він може створити обґрунтований прогноз без налаштування гіперпараметрів;
- він вирішує проблему переобладнання дерев рішень;

2.6 Градієнтний бустинг

Градієнтний бустинг представляє собою високоефективний метод машинного навчання, який оперує за принципом послідовного удосконалення прогнозів за допомогою ансамблю моделей. У цьому підході кожна нова модель в ансамблі спрямована на коригування помилок, що можуть бути допущені попередніми моделями, що робить його особливо потужним і гнучким.

В основі градієнтного бустінгу лежить ідея використання нової моделі для виправлення залишкових помилок, що залишились після прогнозування попередніми моделями. Це означає, що кожна нова модель не намагається просто відтворити цільовий показник напряму, але ставить перед собою завдання виправити чи зменшити помилки, що залишились від попередніх моделей.

Однією з ключових особливостей градієнтного бустінгу є використання градієнтного спуску для мінімізації функції втрат. Цей процес допомагає новій моделі адаптуватися до існуючих помилок, дозволяючи їй краще вирішувати завдання прогнозування з кожним новим внеском в ансамбль.

Градієнтний бустінг використовує градієнтний спуск для мінімізації функції втрат, що допомагає новій моделі адаптуватися до вже існуючих помилок. Це забезпечує покращення якості прогнозування з кожним новим додаванням моделі в ансамбль.

Навчальна вибірка являє собою сукупність пар $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ вектору вхідних змінних x та вихідної змінної y. Задача завдання заключається в тому, щоб оперуючи навчальною вибіркою знайти апроксимуючу функцію $\hat{F}(x)$ до функції $F(x)$, яка мінімізує очікуване значення певної заданої функції втрат за формулою:

$$\hat{F} = \arg\min \mathrm{E}_{x,y}\,[L(y, F(x))] \tag{2.6}$$

де $L(y, F(x))$ – функція втрат.

Метод градієнтного бустингу шукає апроксимуючу функцію $\hat{F}(x)$ у вигляді зваженої суми функцій $h_i(x)$ певного класу H, які рахуються слабкими моделями.

Тож, формула 2.7 відображає визначення функції $\hat{F}(x)$:

$$\hat{F}(x) = \sum_{i=1}^{M} \gamma_i\, h_i(x) + const \tag{2.7}$$

де M – кількість слабких моделей,

γ_i – ваговий коефіцієнт,

$h_i(x)$ – функція слабкої моделі.

Алгоритм знаходження $\hat{F}(x)$ зветься навчанням моделі, а сам процес визначення того, який вийде y з заданого входу даних x при допомозі побудованої моделі – передбаченням.

Отже, у випадку використання градієнтного бустингу з деревами рішень, в результаті навчання алгоритм генерує набір із M дерев рішень. Для проведення прогнозу, тобто визначення виходу y для нового об'єкта x, потрібно здійснити обчислення суми, яка визначається за допомогою наступної формули.

$$y = T_0(x) + \nu \cdot \sum_{m=1}^{M} T_m(x) \tag{2.8}$$

де T_0 – перше дерево рішень,

ν – коефіцієнт масштабування,

M – кількість побудованих дерев рішень,

T_m – m-е дерево рішень.

Фінальний класифікатор представлений у вигляді лінійної комбінації різних класифікаторів. Пошук оптимальних значень коефіцієнтів для цієї лінійної комбінації є досить витратним завданням, тому в градієнтному бустингу використовується жадібний алгоритм, який поетапно додає класифікатори.

У використанні градієнтного бустінгу ключово важливим є правильна параметризація моделі, вибір оптимальних гіперпараметрів та управління ризиком перенавчання для досягнення оптимального прогнозу. Головна перевага бустінгу полягає у його здатності до створення дуже точних моделей. Кожна нова модель, яка додається до ансамблю, фокусується на помилках попередніх, що дозволяє поступово уточнювати прогнози та покращувати загальну точність.

На рисунку 2.8 зображено схему роботи методу градієнтного бустингу

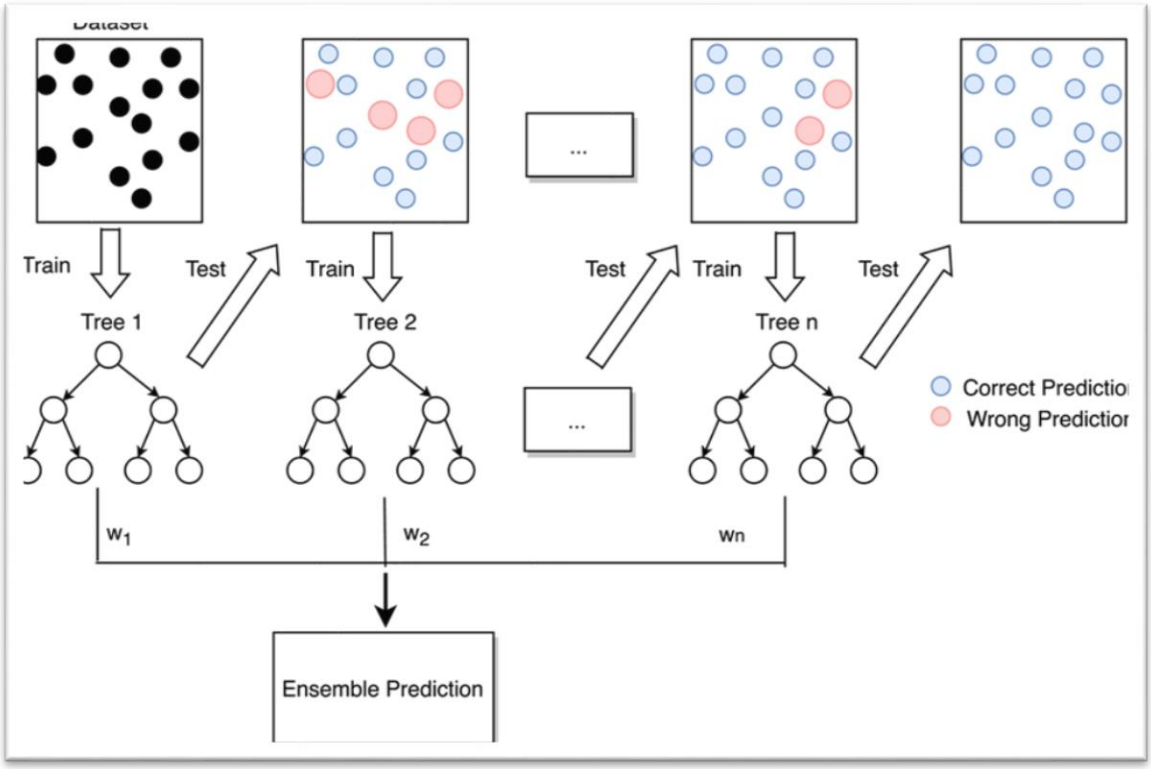


Рис. 2.8 Алгоритм Градієнтного бустингу

AdaBoost та градієнтний бустінг - це два з найпопулярніших методів бустінгу. AdaBoost (Adaptive Boosting) використовує важливість кожного прикладу у тренувальному наборі для покращення моделі. Градієнтний бустінг, у свою чергу, використовує градієнтний спуск для мінімізації функції втрат, покращуючи прогнози моделі через додавання нових елементів.

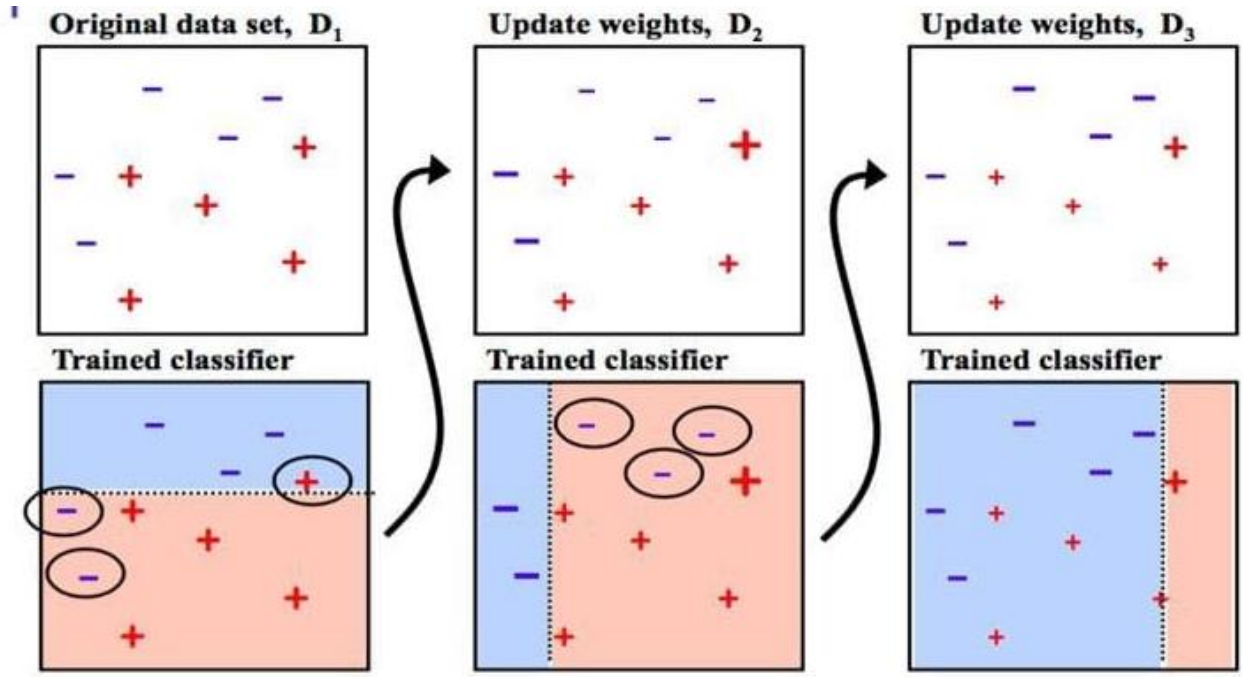


Рис. 2.9 Схема роботи AdaBoost

Градієнтний бустінг, завдяки своїй потужності та гнучкості, став одним із найефективніших методів в багатьох задачах машинного навчання. Основні можливості цього методу включають:

1. **Висока точність:** Градієнтний бустінг відомий своєю здатністю створювати надзвичайно точні моделі, які перевершують багато інших алгоритмів, особливо у завданнях класифікації та регресії.
2. **Робустність до шуму і викидів:** Цей метод зазвичай є менш чутливим до викидів у даних, оскільки кожна нова модель, яка додається, спрямована на коригування помилок попередніх моделей.
3. **Можливість використання з різними видами даних:** Градієнтний бустінг може працювати з різними типами даних, включаючи числові та категоріальні ознаки, без необхідності преобразування даних.
4. **Врахування взаємозв'язків між ознаками:** Метод здатний автоматично враховувати взаємодії між різними ознаками без необхідності вручну задавати взаємодію між ними.
5. **Масштабованість:** Градієнтний бустінг може ефективно працювати навіть з великими наборами даних, хоча і може потребувати більших обчислювальних

ресурсів у порівнянні з іншими методами.

6. **Робота з різними функціями втрат:** Метод може використовувати різні функції втрат в залежності від завдання, що робить його більш універсальним для різноманітних сценаріїв машинного навчання.

Градієнтний бустінг може забезпечити вражаючі результати у багатьох областях, а його гнучкість і здатність до адаптації роблять його важливим інструментом для різних завдань аналізу даних та прогнозування.

Градієнтний бустінг є потужним інструментом для аналізу даних і вирішення задач у багатьох галузях, особливо в області машинного навчання та прогнозування. Ця математична модель базується на ідеї поєднання багатьох слабких моделей для створення сильної та точної прогностичної моделі.

Важливість градієнтного бустінгу полягає в його здатності працювати з різноманітними типами даних і розв'язувати складні задачі прогнозування. Він дозволяє нам покращити точність прогнозів, поєднуючи слабкі моделі в ансамбль, кожна з яких виправляє недоліки попередньої.

Ця модель застосовується в багатьох сферах, від фінансів до медицини, де точні прогнози та аналіз даних грають важливу роль. В області реклами, наприклад, градієнтний бустінг дозволяє підбирати оптимальні стратегії рекламних кампаній, враховуючи різноманітні чинники та підходи до цільових аудиторій.

Завдяки своїй здатності працювати з великими обсягами даних і створювати складні моделі, градієнтний бустінг став важливим інструментом для вирішення задач, де потрібна висока точність прогнозування та аналізу даних.

Однією з ключових переваг градієнтного бустінгу є його здатність працювати з різними типами ознак (від категорійних до числових), а також автоматично враховувати важливість кожної ознаки у процесі навчання.

Така модель виявляється дуже потужною у багатьох сферах, де важливо точно передбачати результати або робити висновки з великих обсягів даних. Її використовують у фінансах для передбачення ризиків, у медицині для діагностики, у технологічних компаніях для аналізу користувацьких поведінок та багатьох інших сферах, де точність та вірогідність прогнозування є ключовими.

2.7 Критерії оцінки ефективності

В моделюванні для оцінки точності моделей використовуються різноманітні метрики, які дозволяють визначити, наскільки добре модель адаптується до реальних даних та як ефективно вона передбачає майбутні значення.

Важливість метрик для оцінки точності прогнозування моделей в моделюванні та інших галузях важко переоцінити. Ці метрики надають кілька ключових переваг у процесі розробки та вдосконалення прогнозуючих моделей.

По-перше, метрики створюють стандартизований інструмент для порівняння ефективності різних моделей. Результати, виражені в числовій формі, надають чітку основу для визначення того, яка модель є більш точною та ефективною в передбаченнях.

По-друге, метрики дозволяють визначити, як добре модель пристосовується до реальних даних та наскільки великою є її загальна помилка. Це допомагає дослідникам та практикам уточнити параметри моделі та виправити потенційні недоліки, забезпечуючи більш точні прогнози. Крім того, метрики стають ефективним інструментом для визначення адекватності моделі. Вони допомагають ідентифікувати можливі систематичні помилки та неправильності в залишках, що можуть виникнути через різноманітні причини, від невірного вибору функціональної форми до наявності автокореляції в залишках.

Серед ключових метрик варто виділити MAPE, RMSE, ex_var та durbin_watson:

1. **MAPE (Mean Absolute Percentage Error):** Вимірює середню абсолютну відсоткову похибку між прогнозом та фактичним значенням. Ця метрика часто використовується в економічних дослідженнях для оцінки точності прогнозів відносно величини фактичних даних. MAPE дозволяє зрозуміти, наскільки великою є середня відсоткова похибка прогнозів, допомагаючи оцінити ефективність моделі.
2. **RMSE (Root Mean Squared Error):** RMSE визначає середнє квадратичне відхилення між прогнозом та фактичним значенням. Ця метрика

використовується для вимірювання точності прогнозу в абсолютних величинах. RMSE надає інформацію про середню величину помилок прогнозу та служить для порівняння різних моделей.

3. **ex_var (Explained Variance):** Визначає частку дисперсії вихідного сигналу, яку можна пояснити моделлю. В інших словах, це відсоток дисперсії, який зумовлений впливом факторів моделі. Цей показник вказує на те, наскільки добре модель пояснює варіацію вихідних даних.
4. **Durbin-Watson Statistic:** durbin_watson є статистикою, яка вказує на наявність автокореляції в залишках моделі. Автокореляція може виникнути, коли залишки моделі залежать один від одного в часі. Цей показник допомагає виявляти проблеми в залишках моделі, що можуть впливати на її адекватність та точність. Значення близьке до 2 вказує на відсутність автокореляції.

Використання цих метрик сприяє об'єктивній оцінці результатів моделювання в економетриці та допомагає визначити, наскільки надійно та адекватно модель описує реальні дані.

3 ПРОГРАМНА РЕАЛІЗАЦІЯ ТА ЕКСПЕРИМЕНТАЛЬНІ РЕЗУЛЬТАТИ

3.1 Вибір інтрументів розробки

Мовою програмування для розробки даного проекту було обрано Python. Дана мова дійсно вражає своєю універсальністю та потужністю для аналізу даних і реалізації методів машинного навчання.

Використання мови програмування Python у сфері машинного навчання має численні переваги, які роблять її вельми популярною серед дослідників, розробників та спеціалістів у цій галузі. По-перше, Python має простий і зрозумілий синтаксис, що полегшує роботу з кодом і дозволяє швидше вивчення для початківців. Це зробило Python відмінним вибором для новачків у галузі машинного навчання.

По-друге, у Python існує велика кількість бібліотек та фреймворків, які сприяють впровадженню машинного навчання. Такі інструменти, як NumPy, pandas, scikit-learn, TensorFlow і PyTorch, надають потужний функціонал для обробки даних, створення моделей та їхнього навчання. Завдяки цим інструментам, розробники можуть швидко реалізувати складні алгоритми машинного навчання без значних труднощів.

Крім того, Python активно використовується у наукових галузях, і велика спільнота розробників активно долучається до розвитку нових інструментів та рішень у галузі машинного навчання. Це забезпечує актуальність та високий рівень підтримки мови, що робить Python переважною вибором для роботи в сфері машинного навчання.

Ось кілька ключових переваг Python у контексті розробки програмного продукту для прогнозування доходу від рекламних кампаній:

1. **Широкий функціонал:** Python має велику кількість бібліотек, таких як NumPy, Pandas, Matplotlib, SciPy, Scikit-learn тощо, які надають потужні

інструменти для аналізу даних, візуалізації та реалізації алгоритмів машинного навчання.

2. **Простота використання:** Чистий і зрозумілий синтаксис Python робить його більш доступним для новачків у світі програмування і сприяє швидкій розробці програмного забезпечення.
3. **Велике співтовариство:** Python має велике співтовариство користувачів і розробників. Це означає наявність безлічі ресурсів, форумів, туторіалів та відкритих бібліотек, що допомагає швидко вирішувати проблеми та знаходити підтримку.
4. **Швидкодія:** Незважаючи на те, що Python не є найшвидшою мовою програмування, багато бібліотек з високою продуктивністю реалізовані на C/C++, що дозволяє використовувати їх у Python, підвищуючи швидкість виконання деяких операцій.
5. **Машинне навчання:** Scikit-learn, TensorFlow, PyTorch і Keras - лише деякі з бібліотек, що надають широкий спектр алгоритмів машинного навчання та нейронних мереж, які можна легко використовувати та налаштовувати в Python.

Python славиться своєю високою кросплатформеністю, що робить його ідеальним вибором для розробки на різних операційних системах, таких як Windows, macOS та різні дистрибутиви Linux. Розробники можуть писати код на Python і впевнено використовувати його на будь-якій платформі без необхідності внесення значних змін чи адаптації. Це робить мову особливо зручною для команд розробників, що працюють у різних середовищах.

Щодо швидкості виконання, Python має ряд оптимізаційних інструментів та JIT-компіляторів, таких як PyPy, які дозволяють покращити продуктивність деяких застосунків. Однак у порівнянні з деякими іншими мовами, такими як C++ або Java, Python може виявлятися менш продуктивним на виконання обчислювально важких завдань. Незважаючи на це, для багатьох задач швидкість Python є прийнятною, а його зручний синтаксис і кросплатформенність

важливими перевагами для розробників у багатьох галузях.

Для написання проекту було вибрано середовище розробки Microsoft Visual Studio Code. Microsoft Visual Studio Code (VS Code) – це безкоштовний і відкритий редактор коду, розроблений компанією Microsoft для програмування та розробки. Він підтримує різноманітні мови програмування та технології.

Ось перелік основних перваг Microsoft Visual Studio Code:

1. **Безкоштовний та Відкритий:** VS Code можна завантажити та використовувати безкоштовно, і він має відкритий код. Це означає, що його можна використовувати громадськість, і розробники можуть вносити свої власні внески або розширювати його функціональність.
2. **Інтеграція з Git:** Вбудована підтримка системи контролю версій Git дозволяє легко відстежувати та керувати змінами у проектах.
3. **Розширене Відлагодження:** VS Code має вбудовану підтримку для відлагодження коду, що полегшує виявлення та виправлення помилок у програмному коді.
4. **Підтримка Мов Програмування:** Редактор підтримує велику кількість мов програмування, таких як JavaScript, Python, C++, Java, HTML, CSS, PHP і інші.
5. **Кросплатформений:** Підтримується на операційних системах Windows, macOS та Linux.
6. **Розширення та Маркетплейс:** Є можливість використовувати розширення, які доступні в маркетплейсі, для розширення функціональності редактора за потребою користувача.
7. **Спільнота та Підтримка:** VS Code має активну спільноту користувачів та регулярно отримує оновлення та вдосконалення.

Цей редактор широко використовується серед розробників завдяки своїй простоті використання, гнучкості та розширеним можливостям, що робить його популярним інструментом для роботи з програмним кодом.

Було використано кілька потужних бібліотек для обробки та аналізу даних, а

також для створення та налаштування моделей машинного навчання:

- Pandas: Це важливий інструмент для роботи з даними. Pandas дозволяє легко завантажувати, обробляти та аналізувати дані, використовуючи таблиці. Його функціонал дозволяє виконувати різноманітні операції, такі як фільтрація, групування та об'єднання даних.
- NumPy: Ця бібліотека стає невід'ємною частиною роботи з багатовимірними масивами та матрицями. NumPy включає велику кількість математичних функцій, що робить ефективну обробку та аналіз числових даних.
- Scikit-learn (sklearn): Scikit-learn - це золотий стандарт для роботи з алгоритмами машинного навчання. Вона надає широкий спектр інструментів для класифікації, регресії, кластеризації та іншого. Scikit-learn спрощує процес створення та оцінювання моделей.
- Matplotlib: Для візуалізації даних використовувався Matplotlib. Ця бібліотека дозволяє легко побудувати різноманітні графіки, що полегшує аналіз даних та представлення результатів.
- XGBoost: XGBoost - це бібліотека з градієнтним бустінгом, яка вирізняється швидкістю та ефективністю реалізації ансамблювання.
- Ensemble та Tree модулі Scikit-learn: Ці модулі включають ансамблі дерев рішень, які ефективно використовуються для класифікації та регресії.
- Linear Model модуль Scikit-learn: Модуль містить реалізації лінійних моделей, корисних для розв'язання завдань регресії та класифікації.
- Model Selection та Metrics модулі Scikit-learn: Model Selection надає інструменти для розділення даних на тренувальний та тестовий набори, а Metrics містить функції для оцінки продуктивності моделей.

Для того, щоб зробити продукт кращим для використання, було вирішено розробити сервер, який буде обробляти запити і віддавати відповіді.В запиті сервер буде отримувати параметри для моделі, а у відповідь відправлятися результат прогнозування моделі.

Стек технологій:

- Flask
- Docker
- Amazon AWS

Flask – фреймворк для створення веб-додатків на мові програмуванняPython, що використовує набір інструментів Werkzeug, а також шаблонізатор Jinja2. Відноситься до категорії так званих мікрофреймворків - мінімалістичних каркасів веб-додатків [12].

Docker – це операційна система для контейнерів. Подібно до того, яквіртуальна машина створює віртуальне уявлення апаратного забезпеченнясервера, контейнери створюють віртуальне уявлення серверної операційної системи. Після установки на кожен сервер Docker надає доступ до простих команд, необхідних для збірки, запуску або зупинки контейнерів [11]. Для створення докер-контейнера був написаний спеціальний файл “Dockerfile”, який зображений на рисинку 3.1.

```
1  FROM python:3.8
2  MAINTAINER Boiko Pavlo 'pboyko172839465@gmail.com'
3
4  WORKDIR /usr/src/app
5  COPY ./app/requirements.txt ./
6  RUN pip3 install --no-cache-dir -r requirements.txt
7  COPY ./app .
8
9  ENTRYPOINT  [ "python","-m", "flask_app"]
10
```

Рис. 3.1 Вміст Dockerfile

Amazon AWS – це найпоширеніша в світі хмарна платформа з широкими можливостями, що надає більше 200 повнофункціональних сервісів для центрів обробки даних по всій планеті [13]. За допомогою AWS було створеноповноцінний віртуальний комп’ютер, який завжди доступний через Інтернет. На AWS було запущено Docker контейнер, в якому функціонуватиме сервер для взаємодії з моделлю.

3.2 Опис вхідних даних для тренування моделі

Повнота та широкий набір параметрів вхідних даних є важливими аспектами для тренування моделей прогнозування та їхньої успішної роботи. Ці параметри дозволяють моделі адекватно взаємодіяти з реальними умовами та робити точні, надійні прогнози.

Повні та репрезентативні дані дозволяють моделі краще усвідомлювати особливості та закономірності в досліджуваному явищі чи домені. Наприклад, у сфері метеорології для прогнозування погоди важливо мати повні та актуальні дані про температуру, вологість, вітряні умови тощо. Без повної інформації модель може упустити ключові фактори, що призведе до неточних прогнозів.

Широкий набір параметрів вхідних даних дозволяє враховувати різноманітні аспекти та взаємозв'язки в системі. Він розширює можливості моделі в аналізі та прогнозуванні, забезпечуючи комплексний підхід до розв'язання завдань. Наприклад, у фінансовому прогнозуванні важливо враховувати різні фактори, такі як економічні показники, політичні події та інші, для точного прогнозу ринкових тенденцій.

В роботі будуть використовуватися реальні дані, отримані від компанії “AddMetrics”. Дані завантажені з Facebook кабінетів, де ведеться статистика реклами за 2023 рік.

Щоденна статистика для тренування моделі наведена на рисунку 3.2.

project_id	acc_id	country_code	date	spend	impression	reach	click	install	purchase	conversion_values_usd	ROAS
1	act_884357531647058	MX	2021-04-25	2126.49	216159	123090	123090	352	34	779.1	0.36
2	act_884357531647058	MX	2021-04-25	529.134	3021	100	100	21	7	100	0.18
3	act_884357531647058	MX	2021-04-25	12	190	21	21	10	1	12.9931	1

Рис. 3.2 Приклад щоденних даних

Отже, вхідні дані включатимуть в себе такі змінні, як:

- project_id – ідентифікатор проекта для якого працює реклама;
- acc_id – ідентифікатор Facebook аккаунта, де працює реклама.;
- country_code – країна, для якої зібрана інформація;
- date – дата, за яку зібрана інформація;
- spend – витрати;
- impression – покази, скільки реклама була показана;
- reach – охоплення, число людей, які хоча б раз побачили рекламу.Відмінність охоплення від показів полягає в тому, що останні можуть включати багаторазові перегляди реклами однією людиною;
- click – число переходів по посиланнях в рекламних оголошеннях, в результаті яких було здійснено перехід на цільові сторінки;
- install – кількість завантажень;
- purchase – кількість покупок;
- conversion_values_usd - цінність конверсії покупок. Загальна вартість, повернута після покупки в мобільному додатку;
- ROAS – показник рентабельності рекламних витрат.

Вхідні дані використовуються під час тренування моделей прогнозування для створення внутрішнього представлення про взаємозв'язки та закономірності в досліджуваному явищі. Ці дані надають моделі інформацію про різноманітні параметри та їхні взаємодії, створюючи основу для подальшого аналізу. Чим більшим та різноманітнішим є спектр вхідних даних, тим більше можливостей для моделі усвідомити різні аспекти досліджуваного явища.

Враховуючи вхідні дані, модель виробляє абстракції та робить припущення щодо залежностей між різними параметрами. Тренувальний процес полягає в тому, щоб модель навчилася адаптуватися до різноманітних умов та динаміки даних, що покращує її прогностичні можливості. Змістовне та повне представлення вхідних даних є важливим, оскільки воно визначає ефективність та точність прогнозів

моделі.

Крім того, вхідні дані визначають параметри та структуру моделі, впливаючи на її здатність до виявлення складних взаємозв'язків та роботи з новими даними. Усі ці аспекти взаємодії з вхідними даними формують основу для ефективного та адаптивного прогнозування.

На рисунку 3.3 показаний приклад щогодинних даних.

project_id	campaign_name	acc_id	date	spend	conversion_values_usd	ROAS	hour
1	WW_APP_iOS14_TOP_M_BC/16	act_884357531647058	2021-04-25	2126.49	779.1	0.36	11
2	WW_APP_iOS14_TOP_M_BC/16	act_884357531647058	2021-04-25	529.134	100	0.18	01
3	WW_APP_iOS14_TOP_M_BC/16	act_884357531647058	2021-04-25	12	12.9931	1	21

Рис. 3.3 Приклад щогодинних даних

Опис параметрів вхідних даних:

- project_id – ідентифікатор проекту, для якого працює реклама;
- acc_id – ідентифікатор Facebook аккаунта, де працює реклама;
- campaign_name – назва рекламної кампанії;
- date – дата, за яку зібрана інформація;
- spend – витрати;
- hour – година, за яку зібрана інформація;
- conversion_values_usd – цінність конверсії покупок в мобільному додатку;
- ROAS – показник рентабельності рекламних витрат.

3.3 Обробка вхідних даних для прогнозування

Підготовка вхідних даних для тренування моделі - це ключовий етап у створенні ефективних та точних прогностичних моделей в області машинного навчання. Основною метою цього процесу є забезпечення якості та коректності

вхідних даних для того, щоб модель могла належним чином вивчати та узагальнювати патерни.

В рамках цього завдання видалення непотрібних колонок з таблиці вхідних даних визначається як важливий крок. Це включає в себе врахування різних аспектів, таких як значущість та релевантність інформації, яку надає кожна колонка. Вибір та збереження лише ключових параметрів сприяє підвищенню ефективності моделі, спрощенню її структури та скороченню часу навчання. Важливо ретельно вибирати колонки, щоб забезпечити те, що вони дійсно впливають на результативність моделі та приносять інформаційну цінність.

По-перше, необхідно видалити дублікати. Дублікати можуть виникнути через невірне збирання даних або інші причини. Їхня присутність може призвести до перекривання даних та спричинити неправильні результати під час тренування моделі. Таким чином, перед початком тренування важливо визначити та видалити дублікати, щоб забезпечити чистоту даних.

По-друге, слід видалити колонки, які не несуть інформації, корисної в процесі прогнозування. Деякі колонки можуть містити інформацію, яка не є необхідною для тренування моделі. Це можуть бути, наприклад, ідентифікатори чи статичні дані, які не змінюються і не впливають на результати. Вилучення таких колонок допомагає спростити дані та зменшити їхню складність, що, в свою чергу, поліпшує ефективність тренування моделі.

Отже, видалення непотрібних колонок є важливим етапом в підготовці даних для тренування моделі, сприяючи підвищенню точності та робастності отриманих результатів.

Конвеєр обробки даних – це важливий інструмент перед підготовкою та використанням даних у моделях машинного навчання. Його етапи включають:

Нормалізація та масштабування даних, що забезпечує однаковий масштаб для всіх ознак та поліпшує ефективність моделі. Видалення відсутніх значень – обробка або вилучення відсутніх даних для покращення якості моделі. Кодування категоріальних даних, тобто перетворення категоріальних ознак у числові,

забезпечуючи коректну обробку моделлю. Особливості інженерії – створення нових ознак для підвищення інформативності моделі. Видалення зайвих ознак, які можуть нести мало корисної інформації для моделі. Перехресна перевірка та валідація моделей, включаючи розбиття даних та перехресну перевірку для ефективності моделі. Балансування класів (за потреби) для уникнення впливу нерівномірного розподілу класів. Видалення викидів, що можуть вплинути на точність моделі.

Перед тим як дані будуть готові до роботи, в машинному навчанні необхідним є процес очистки даних. Очистка даних включає в себе наступні пункти:

- Виключення знаків пунктуації: Це означає усунення всіх знаків, що не є буквами або цифрами, таких як коми, крапки, лапки і так далі. Підходить для поліпшення сприйняття тексту моделлю, зменшення шуму та скорочення розміру словника.

- Вилучення чисел: При аналізі тексту числа можуть бути неінформативними, тому вилучення їх сприяє кращому розумінню текстового контексту.

- Вилучення "стоп-слів": Це ті слова, які зазвичай не мають значущого семантичного змісту (наприклад, "та", "і", "або" українською мовою). Вони можуть бути вилучені для скорочення обсягу тексту та покращення ефективності алгоритмів.

- Перетворення на нижній регістр: Приведення всього тексту до нижнього регістру спрощує уніфікацію слів та зменшує варіативність форм слів, що полегшує розпізнавання та аналіз тексту моделлю.

- Кількість слів: Включення цього параметра може бути корисним для відстеження кількості слів після обробки тексту та використання цієї інформації у моделі.

Після проведення всіх етапів підготовки та аналізу даних колонки "original_spend" виявлено, що значенню "0" відповідає значна кількість даних - аж 2 495 325 записів. Однак у контексті аналізу витрат, ситуація, коли значення "original_spend" менше 0, є неприродною. Такі дані можуть бути результатом помилки вводу або інших аномалій. Логічним вирішенням цього аспекту є видалення усіх записів, де значення "original_spend" менше 0. Після проведення цієї операції кількість записів зменшилась до 934 425, що є значущим кроком для подальшого аналізу та моделювання. Такий підхід дозволяє уникнути впливу аномальних чи невірних даних на вірогідність та точність моделі.

Отже, видаливши всі непотрібні колонки, залишилось 13 колонок, які будуть використовуватись для тренування моделі:

- campaign_id;
- project_id;
- country_code;
- date;
- original_spend;
- impression;
- reach;
- click;
- conversion_values_usd;
- fb_id;
- campaign_name;
- start_time;
- currency_code.

Залишені дані після вилучення зайвих колонок стають основою для тренування прогностичної моделі, оскільки вони містять значущу та корисну інформацію. Цей етап відіграє ключову роль у підготовці даних, оскільки відбирає лише ті параметри, які є суттєвими для навчання моделі та утворення патернів.

Важливо, щоб обрані дані відображали реальні зв'язки та властивості досліджуваного явища, щоб модель могла належним чином усвідомити та прогнозувати майбутні вхідні дані. Такий підхід сприяє підвищенню точності та ефективності моделі під час її тренування.

id	campaign_id	project_id_1	country_code	date	original_spend	impression	reach	click	conversion_values_usd	project_id	fb_
45023369	23844215753110197	18	AU	2021-02-08	0.02	1	0	0	0.00000	18	act_90961744605915
44849476	23844215753110197	18	BA	2021-02-05	0.00	2	0	0	0.00000	18	act_90961744605915
44640328	23844215753110197	18	BG	2021-02-01	0.00	1	0	0	0.00000	18	act_90961744605915
44583938	23844215753110197	18	BM	2021-01-31	0.00	1	0	0	0.00000	18	act_90961744605915
44746848	23844215753110197	18	BQ	2021-02-03	0.00	1	0	0	0.00000	18	act_90961744605915
...	...	...	...	...	...	...	...	...	...	...	
48492329	23848110230990530	14	US	2021-04-25	101.01	10568	234	234	44.39230	14	act_37796674291288
48441708	23848110453850530	102	US	2021-04-24	191.68	8267	108	108	0.00000	102	act_37796674291288
48492331	23848110453850530	102	US	2021-04-25	0.38	23	0	0	13.11700	102	act_37796674291288
48441709	23848110462570530	102	US	2021-04-24	574.44	21107	583	583	17.19000	102	act_37796674291288
48492332	23848110462570530	102	US	2021-04-25	0.55	40	0	0	8.57138	102	act_37796674291288

Рис. 3.4 DataFrame після обробки даних

У сучасному програмуванні на мові Python використовуються різноманітні типи даних, що дозволяє ефективно взаємодіяти з різними аспектами інформації.

Один із ключових типів даних - це цілі числа (int). Вони використовуються для представлення дискретних, цілих значень, таких як кількість об'єктів, порядкові номери чи будь-які інші цілочисельні параметри.

Другий важливий тип - числа з плаваючою комою (float), які використовуються для представлення десяткових числових значень. Вони є необхідними для точного відображення параметрів, які можуть мати десяткову частину, таких як координати, вага чи вартість товарів.

Третім ключовим типом є об'єкт (object), який може представляти різноманітні структури даних або навіть власні класи. Об'єктний тип дозволяє створювати складні дані та об'єкти, що розширює можливості моделювання для більш різноманітних задач.

Ця різноманітність типів даних у Python надає програмістам гнучкість та потужність при роботі з різноманітними видами інформації та завдань.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5562905 entries, 0 to 5562904
Data columns (total 14 columns):
 #   Column                 Dtype
---  ------                 -----
 0   id                     int64
 1   campaign_id            int64
 2   country_code           object
 3   date                   object
 4   original_spend         float64
 5   impression             int64
 6   reach                  int64
 7   click                  int64
 8   conversion_values_usd  float64
 9   project_id             int64
 10  fb_id                  object
 11  campaign_name          object
 12  start_time             object
 13  currency_code          object
dtypes: float64(2), int64(6), object(6)
memory usage: 594.2+ MB
```

Рис. 3.5 Типи даних в колонках

3.4 Векторизація тексту

TfidfVectorizer представляє собою один із методів векторизації тексту, спрямований на перетворення текстових даних у форму, зрозумілу для моделей машинного навчання. Цей метод ґрунтується на концепції TF-IDF, що скорочено від "term frequency-inverse document frequency" — частота терміну-інверсійна частота документа.

Основна мета застосування TfidfVectorizer полягає в конвертації тексту у вектори числових значень, які можуть бути зрозумілі та використані моделями машинного навчання. Цей процес включає кілька ключових кроків, спрямованих на створення числового подання текстової інформації для подальшого використання в алгоритмах машинного навчання.

Однією з перших етапів є обчислення TF-IDF для кожного терміну у кожному документі. Це визначає, наскільки важливим є певне слово у конкретному документі у порівнянні з іншими документами. Далі, отримані TF-IDF значення об'єднуються для створення векторів, що представляють кожен документ у просторі числових ознак.

Таким чином, TfidfVectorizer стає ефективним інструментом у попередній обробці текстових даних, забезпечуючи їхнє числове представлення, яке можна успішно використовувати в подальших завданнях машинного навчання.

Ось основні кроки алгоритму:

- **TF - Term Frequency (Частота терміну):** Це частота, з якою терміни з'являються в документі. Чим частіше термін зустрічається у документі, тим вище його значення. TF обчислюється як кількість входжень терміну у документ поділена на загальну кількість слів у документі.
- **IDF - Inverse Document Frequency (Інверсна частота документа):** Це міра того, наскільки унікальний термін є у колекції документів. Чим менше часто термін зустрічається у всіх документах, тим більше його значення. IDF обчислюється як логарифм відношення загальної кількості документів до кількості документів, у яких зустрічається термін.
- **TF-IDF:** Це добуток TF та IDF. Це значення оцінює важливість терміну для конкретного документа у контексті всього корпусу тексту. Він показує, наскільки термін є важливим для певного документа у порівнянні з іншими документами у колекції.

TfidfVectorizer – це інструмент, який відіграє ключову роль у процесі числового представлення текстових даних для подальшого використання в моделях машинного навчання. Основна мета цього алгоритму полягає в обчисленні значень TF-IDF (term frequency-inverse document frequency) для всіх слів у тексті, створюючи вектор для кожного документа, де кожен компонент вектору відображає TF-IDF для відповідного слова.

На початковому етапі використання TfidfVectorizer включає обчислення

частоти та ваги кожного слова у кожному документі. Цей процес дозволяє визначити, наскільки значущим є кожен термін у контексті кожного документа порівняно з іншими документами в наборі даних.

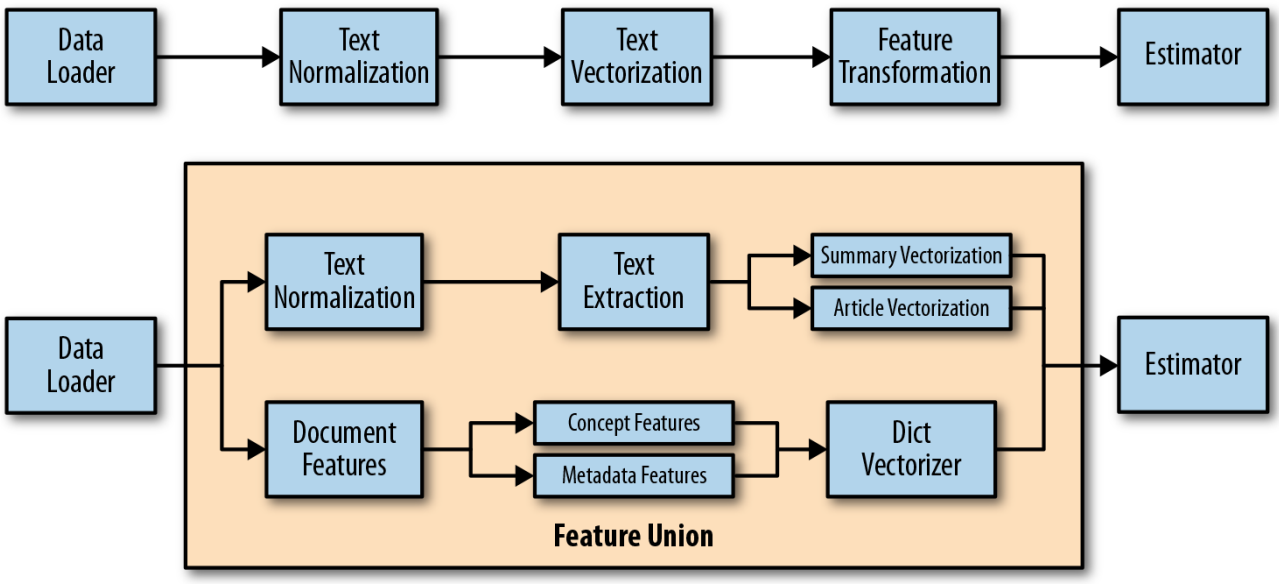


Рис.3.6 Алгоритм веторизації тексту

Ключовим результатом є створення векторів для кожного документа, де кожен вектор містить компоненти, представлені значеннями TF-IDF для відповідних слів. Такий числовий формат дозволяє моделям машинного навчання ефективно працювати з текстовою інформацією, забезпечуючи можливість прогнозування або класифікації на основі введених даних.

Важливо відзначити, що TfidfVectorizer є популярним алгоритмом для отримання значущого числового представлення тексту, що використовується в різних завданнях машинного навчання. [10].

В сучасному аналізі даних та машинному навчанні великою проблемою є наявність пропущених значень в даних. SimpleImputer стає ключовим інструментом для попередньої обробки даних, особливо коли важливо ефективно вирішити це питання. Цей інструмент пропонує різні стратегії заповнення пропусків, враховуючи при цьому різноманітність даних та особливості завдань.

Основні стратегії заповнення пропусків:

1. **Найчастіше значення:** Ця стратегія використовується для категоріальних або текстових даних. Заміна пропущених значень найпоширенішим значенням в стовпці забезпечує збереження характеру найчастіше зустрічаючогося елемента в кожній категорії.
2. **Середнє і медіана:** Застосовується для числових даних, особливо коли пропуски мають визначений паттерн або є характерними для числових ознак. Використання середнього або медіанного значення допомагає ефективно заповнити пропуски в числових даних, які не можуть бути видалені.
3. **Константа**: Цей метод використовується, коли важливо зберегти інформацію про відсутність даних. Заміна пропущених значень сталим значенням, таким як "невідомо" чи "відсутньо", корисна у випадках, коли пропуски не випадкові і їх наявність має вагому вплив на модель.

Обрана стратегія заповнення пропусків повинна бути узгоджена із суттєвістю даних та враховувати можливий вплив на якість моделі. У даному випадку обрана стратегія найчастіше повторюваних значень логічна, оскільки враховувала як текстові, так і числові дані, зберігаючи важливий контекст категоріальних ознак

3.5 Результат ефективності прогнозування

Після завершення процесу підготовки даних, наступним кроком було запущено алгоритми для тренування моделей машинного навчання. Цей етап включав в себе використання різноманітних алгоритмів, таких як класифікації, регресії або кластеризації, залежно від конкретної задачі.

Спочатку проводилося налаштування параметрів моделей для забезпечення їхньої оптимальності, щоб досягти найкращих результатів на навчальних даних. Після цього розпочиналася фаза тренування, під час якої моделі вивчали закономірності та структуру даних для подальшого використання в передбаченні або категоризації нових вхідних даних.

Процес тренування включав ітеративні цикли, в ході яких здійснювалося впровадження та вдосконалення моделей. За результатами тренування проводилася оцінка продуктивності за допомогою тестового набору даних для перевірки загальної ефективності та уникнення перенавчання моделей. Такий підхід дозволяв забезпечити високу точність та надійність побудованих моделей машинного навчання.

Для тренування моделей використовувались наступні алгоритми:

- GradientBoostingRegressor
- RandomForestRegressor
- XGBRFRegressor
- XGBRegressor

Обрані алгоритми для тренування моделей, такі як GradientBoostingRegressor, RandomForestRegressor, XGBRFRegressor та XGBRegressor, демонструють високу ефективність у розв'язанні конкретних завдань машинного навчання. Вони виявилися досить гнучкими та здатними адаптуватися до різноманітних видів даних, надаючи високу точність передбачень та добре впоравшись з різноманітними видами модельних завдань.

GradientBoostingRegressor та RandomForestRegressor відзначилися високою точністю у вирішенні завдань регресії, враховуючи взаємодію між різними ознаками даних. У свою чергу, XGBRFRegressor та XGBRegressor, як бібліотеки з градієнтним бустінгом, забезпечили швидку та ефективну реалізацію методу ансамблювання, що сприяло високій якості передбачень.

Загалом, обрані алгоритми відзначаються оптимальним балансом між складністю моделі та її здатністю адаптуватися до різноманітних умов, роблячи їх ефективними засобами для завдань машинного навчання.

Числові показники результатів тренування моделі наведені на рисунку 3.7

```
GradientBoostingRegressor: -0.4138003166592 57
RandomForestRegressor: -0.20495520668602 77
XGBRegressor: -0.29263438534999 18
XGBRFRegressor: -0.92553600178286 51
CPU times: user 10.8 s, sys: 25.3 s, total: 36.1 s
Wall time: 8h 31min 43s
```

Рис. 3.7 Числові показники тренування моделі

В рамках вивчення та прогнозування доходності реклами було використано кілька передових алгоритмів машинного навчання. Серед обраного набору алгоритмів виділяються GradientBoostingRegressor, RandomForestRegressor, XGBRFRegressor та XGBRegressor, які визначаються своєю високою ефективністю у вирішенні завдань регресії та адаптивністю до різноманітних видів даних.

Під час тренування моделі виявилося, що алгоритми GradientBoostingRegressor та RandomForestRegressor продемонстрували високу точність та добре справилися з різними характеристиками даних, забезпечуючи оптимальні результати у завданнях регресії. Однак при аналізі візуального відображення результатів виявилось, що XGBRegressor виявився найефективнішим у вирішенні конкретного завдання прогнозування доходності реклами.

Застосування алгоритму XGBRegressor призвело до вражаючих результатів, де висока точність та швидкодія дозволили отримати передбачення з великою достовірністю. Його ефективність може бути пояснена використанням градієнтного бустінгу та оптимізацією параметрів, що дозволило досягти високого рівня адаптації до особливостей рекламних даних. Такий висновок надає підтримку вибору XGBRegressor для досягнення найкращих результатів у прогнозуванні доходності реклами в даному дослідженні.

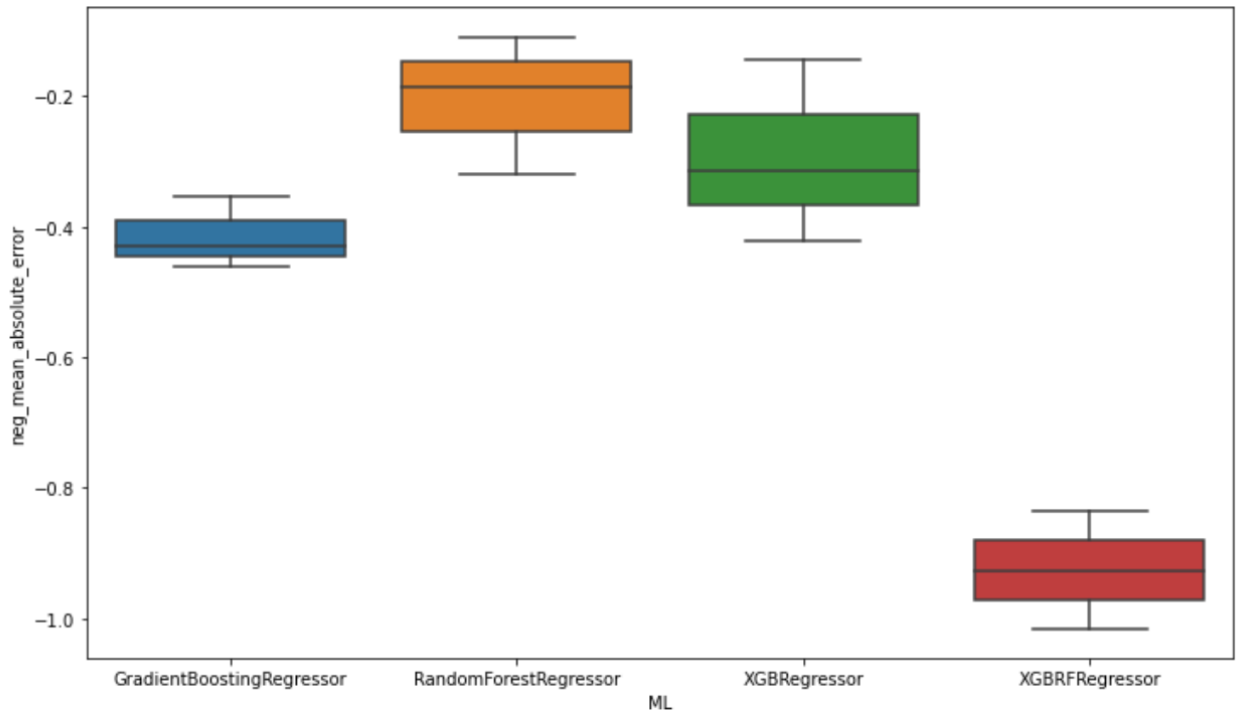


Рис.3.8 Візуальне відображення результатів в графічному інтерфейсі

У результаті порівняльного аналізу різних методів прогнозування в машинному навчанні та оцінки їхньої ефективності за допомогою відповідних метрик, виявлено, що XGBRegressor видається найбільш оптимальним в контексті даного дослідження. Результати отримані з використанням цього методу перевершують конкурентів, визначаючи його як найефективніший для прогнозування ефективності реклами та підвищення доходності рекламної компанії.

XGBRegressor продемонстрував найкращі показники точності та надійності прогнозування, що підкреслює його високий потенціал в обробці конкретної вибірки даних. Враховуючи ці результати, доцільність використання XGBRegressor у сфері прогнозування ефективності реклами стає очевидною. Цей метод може стати ключовим інструментом для підвищення точності та успішності рекламних кампаній, сприяючи позитивним змінам у доходності та стратегіях маркетингу компанії.

В ході підбору параметрів для тренування моделі, було виявлено, що найкраща модель виходить з використанням алгоритму “XGBRegressor” з

параметрами, які показані на рисунку 3.9.

```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=1, eval_metric='mae', gamma=0,
             gpu_id=-1, importance_type='gain', interaction_constraints='',
             learning_rate=0.300000012, max_delta_step=0, max_depth=3,
             min_child_weight=1, missing=nan, monotone_constraints='()',
             n_estimators=100, n_jobs=0, num_parallel_tree=1, random_state=0,
             reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
             tree_method='exact', validate_parameters=1, verbosity=None)
```

Рисунок 3.9 Параметри XGBRegressor

Показники метрик, що використовуються для оцінки ефективності моделі прогнозування наведено нижче. Виявилось, що відповідно до нормативних значень, алгоритм XGBRegressor показав себе чудово з наступними результатами:

	name	r2	MAPE	RMSE	ex_var	durbin_watson
0	XGBRegressor_train	0.999931	7.204061	2.382237	0.999931	2.006304
0	XGBRegressor_test	0.990706	7.272750	20.588067	0.990706	2.001499

Рисунок 3.10 Показники метрик в результаті тренування

ВИСНОВКИ

В результаті виконання магістерської роботи було проведено тренування моделей прогнозування на основі історичних даних з використанням методів машинного навчання.

У роботі було проаналізовано та описано типи та методи сучасного прогнозування ефективності реклами. Проведено аналіз різних методів машинного навчання для прогнозування ефективності реклами. Виявлено, що для того, щоб передбачати прибутковість реклами, оптимальною є модель, побудована на основі алгоритму "XGBRegressor". Оцінюючи цю модель за декількома метриками, отрмиано наступні результати: 0.99 за r2, 0.07 за MAPE, 20.58 за RMSE та 2.001499 за durbin_watson.

Ці результати підтвердили достатню ефективність моделі у прогнозуванні ефективності рекламних кампаній. Було побудовано веб-сервер для обробки даних, використовуючи технології Docker та AWS, щоб забезпечити зручний доступ до функціоналу моделі.

Застосування цього методу для прогнозування доходності реклами дає змогу краще управляти рекламним бюджетом та зрозуміти ефективність своїх рекламних кампаній, сприяючи зростанню бізнесу та ефективності маркетингової стратегії.

Розроблена модель прогнозування ефективності реклами має потенціал зберегти значну частину рекламного бюджету та допомогти бізнесу ефективно розширюватися.

Застосування методів машинного навчання для цієї задачі дозволило побудувати модель, яка ефективно прогнозує дохід від рекламних кампаній. Результати цієї моделі були інтегровані в програмне забезпечення із використанням серверів.

ПЕРЕЛІК ПОСИЛАНЬ

1. Оновлення iOS 14.5: додана можливість розблокувати iPhone задопомогою Apple Watch, а також нові голоси Siri і засоби управління конфіденційністю URL: https://www.apple.com/ru/newsroom/2021/04/ios-14-5- offers-unlock-iphone-with-apple-watch-diverse-siri-voices-and-more/ (датазвернення: 2023.09.05)
2. Estelle Laziuk Daily iOS 14.5 Opt-in Rate URL: https://www.flurry.com/blog/ios-14-5-opt-in-rate-att-restricted-app-tracking-transparency-worldwide-us-daily-latest-update/apple-watch-diverse-siri-voices- and-more/ (дата звернення: 2023.09.10)
3. Engineering to Improve Marketing Effectiveness (Part 1) URL: https://netflixtechblog.com/engineering-to-improve-marketing-effectiveness-part- 1-a6dd5d02bab7 (дата звернення: 2023.09.01)
4. Engineering to Improve Marketing Effectiveness (Part 2) — Scaling Ad Creation and Management URL: https://netflixtechblog.com/https-medium-com-netflixtechblog-engineering-to-improve-marketing-effectiveness-part-2- 7dd933974f5e. (дата звернення: 2023.09.1)
5. Facebook Business Manager URL: https://ru-ru.facebook.com/business/tools/business-manager.
6. Aurelien Geron , Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems. Sebastopol; O'Reilly Media, 2019. 25-290 p.
7. Christopher M. Bishop, Pattern Recognition and Machine Learning.Cambridge; Springer.138-172 p.
8. What Is a Data Pipeline? URL: https://hazelcast.com/glossary/data-pipeline/ (дата звернення: 2023.10.13)

9. Sklearn.impute.SimpleImputer URL: https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html (дата звернення: 2023.10.15)

10. TF-IDF Vectorizer scikit-learn URL: https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a (дата звернення: 2023.10.26)

11. Що таке Docker? URL: https://aws.amazon.com/ru/docker/ (дата звернення: 2023.10.15)

12. Flask’s documentation URL: https://flask.palletsprojects.com/en/2.0.x/ (дата звернення: 2023.10.27)

13. Хмарні обчислення за допомогою AWS URL: https://aws.amazon.com/ru/what-is-aws/ (дата звернення: 2023.11.17)

14. 18 метрик і KPI інтернет-маркетингу, які ви повинні знати URL: https://www.owox.ru/blog/articles/digital-marketing-metrics-and- kpis/#h92e30e3bd (дата звернення: 2023.11.21)

15. Що таке ROAS: формула розрахунку і приклади URL: https://www.owox.ru/blog/articles/roas/ (дата звернення: 2023.11.21)

16. Відкритий курс машинного навчання. Тема 10. Метод найшвидшого бустінг URL: https://habr.com/ru/company/ods/blog/327250/ (дата звернення: 2023.12.20)

17. Estelle Laziuk Daily ML Rate URL: https://www.flurry.com/blog/ios-14-5-opt-in-rate-att-restricted-app-trackingtransparency-worldwide-us-daily-latest-update/apple-watch-diverse-siri-voicesand-more/ (дата звернення: 2023.11.11)

18. Engineering to Improve Marketing Effectiveness (Part 1) URL: https://netflixtechblog.com/engineering-to-improve-marketing-effectiveness-part1-a6dd5d02bab7 (дата звернення: 2023.12.01)

19. Engineering to Improve Marketing Effectiveness (Part 2) — Scaling Ad Creation and Management URL: https://netflixtechblog.com/https-medium-comnetflixtechblog-engineering-to-improve-marketing-effectiveness-part-27dd933974f5e. (дата звернення: 2023.11.05)

20. Melnick K. Learn How To Build A Gaming PC In VR. URL: https://vrscout.com/news/learn-how-to-build-a-gaming-pc-in-vr/.
21. Прокопенко С. Історія віртуальної реальності. Як наш світ захоплює віртуальна реальність. gwara media. URL: https://gwaramedia.com/istoriia-virtualnoi-realnosti/ (дата звернення: 01.03.2023).
22. Facebook Business Manager URL: https://ruru.facebook.com/business/tools/business-manager.
23. Aurelien Geron , Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems. Sebastopol; O'Reilly Media, 2019. 25-290 p.
24. Christopher M. Bishop, Pattern Recognition and Machine Learning. Cambridge; Springer.138-172 p.
25. What Is a Data Pipeline? URL: https://hazelcast.com/glossary/data-pipeline/ (дата звернення: 2023.11.15)
26. Sklearn.impute.SimpleImputer URL: https://scikitlearn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html (дата звернення: 2023.11.23)
27. Петров, В. Ф., et al. "Виникнення технологій віртуальної реальності і їх введення в медицину." *Україна. Здоров'я нації* 4 (2022): 134-138.
28. Вернигора, А. В. "Віртуальна реальність в реальному будівництві." 226-227.
29. TF-IDF Vectorizer scikit-learn URL: https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a (дата звернення: 2023.11.12)
30. Що таке Docker? URL: https://aws.amazon.com/ru/docker/ (дата звернення: 2023.11.15)
31. Flask's documentation URL: https://flask.palletsprojects.com/en/2.0.x/ (дата звернення: 2023.11.25)
32. Хмарні обчислення за допомогою AWS URL: https://aws.amazon.com/ru/what-is-aws/ (дата звернення: 2023.11.25)
33. 18 метрик і KPI інтернет-маркетингу, які ви повинні знати URL: https://www.owox.ru/blog/articles/digital-marketing-metrics-andkpis/#h92e30e3bd

(дата звернення: 2023.11.21)

34. Що таке ROAS: формула розрахунку і приклади URL: https://www.owox.ru/blog/articles/roas/ (дата звернення: 2023.12.21)
35. Відкритий курс машинного навчання. Тема 10. Метод найшвидшого бустінг URL: https://habr.com/ru/company/ods/blog/327250/ (дата звернення: 2023.12.21)

ПРЕЗЕНТАЦІЙНІ МАТЕРІАЛИ

(Презентація)

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-ТЕЛЕКОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

НАВЧАЛЬНО–НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра інженерії програмного забезпечення

МАГІСТЕРСЬКА РОБОТА

«ПІДВИЩЕННЯ РЕЗУЛЬТАТИВНОСТІ РЕКЛАМНОЇ КОМПАНІЇ НА ОСНОВІ МЕТОДІВ МАШИННОГО НАВЧАННЯ»

Виконав: студент групи ПДМ-64, Кухаренко Юрій Дмитрович

Керівник: к. ф.-м. н., доцент кафедри ІПЗ Садовенко Володимир Сергійович

Київ - 2023

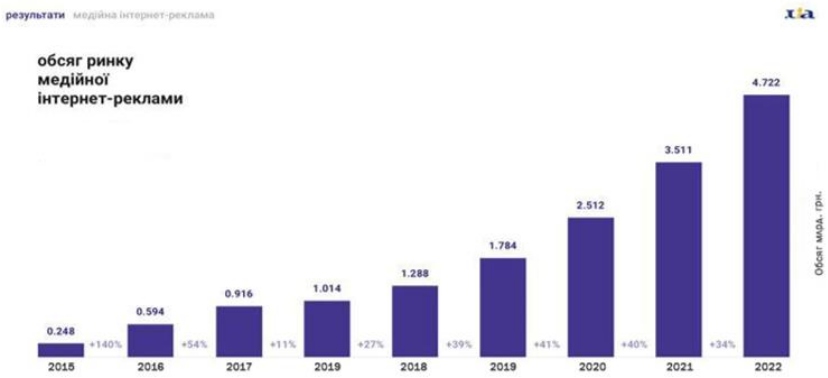
МЕТА, ОБ’ЄКТ, ПРЕДМЕТ ДОСЛІДЖЕННЯ

Мета роботи: збільшення доходності рекламної компанії на основі методів машинного навчання.

Об’єкт дослідження: процес прогнозування ефективності реклами.

Предмет дослідження: методи прогнозування машинного навчання для прогнозування ефективності реклами.

АКТУАЛЬНІСТЬ ТЕМИ



Графік росту обсягу сфери інтернет-реклами в Україні.

3

МЕТОДИ МОДЕЛЮВАННЯ

Метод	Переваги	Недоліки
Градієнтний бустинг	Висока точність, ефективно моделює складні залежності в даних, наявність вбудованої регуляризації	Схильність до перенавчання, довгий час навчання
Випадковий ліс	Уникає перенавчання, ефективний у роботі з великим обсягом даних	Невисока точність, нездатність моделювати складні залежності, вразливість до винятків
XGBRegressor (XGBoost Regressor)	Висока швидкість та точність, стійкість до перенавчання, посилена регуляризація	Чутливість до гіперпараметрів, ресурсоємність
XGBRFRegressor (XGBoost Random Forest Regressor)	Висока швидкість навчання, стійкість до перенавчання, ефективність на великих наборах даних	Менша точність, обмеженість в моделюванні, чутливість до гіперпараметрів

4

МАТЕМАТИЧНІ РОЗРАХУНКИ ТА СХЕМА АЛГОРИТМУ ГРАДІЄНТНИЙ БУСТИНГ

- Функція втрат:

$$L(y_i, h(x_i)) = (y_i - h(x_i))^2$$

L – функція втрат,
y – рішення з навчальної вибірки,
$h(x)$ – результат і-го дерева рішень.

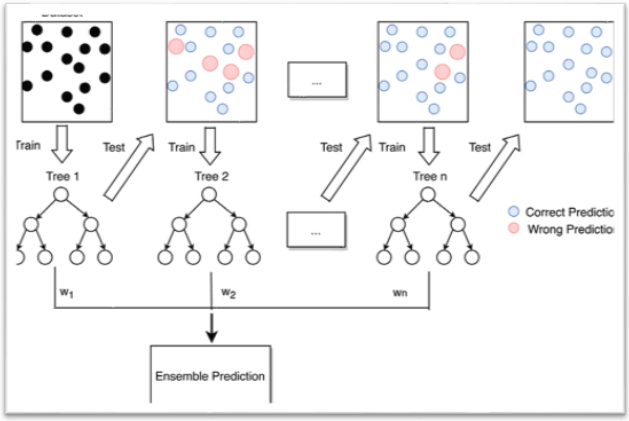


Схема роботи алгоритму Градієнтний бустинг

- Загальна помилка:

$$Q = \sum_{i=1}^{N} L(y_i, h(x_i)) = \sum_{i=1}^{N} (y_i - h(x_i))^2$$

5

АЛГОРИТМ РОБОТИ МЕТОДУ

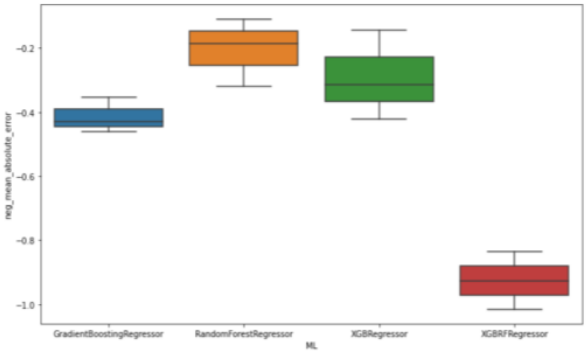
6

ВХІДНІ ДАНІ

Назва параметру	Значення
Project_id	Ідентифікатор проєкта для якого працює реклама
Acc_id	Ідентифікатор акаунта, де працює реклама
Country_code	Країна, для якої зібрана інформація
Date	Дата, за яку зібрана інформація
Spend	Витрати
Impression	Покази
Reach	Охоплення
Click	Число переходів по посиланнях
Install	Кількість завантажень
Purchase	Кількість покупок
Conversion_values_usd	Цінність конверсії покупок
ROAS	Показник рентабельності рекламних витрат

ПРАКТИЧНИЙ РЕЗУЛЬТАТ

- Візуальне відображення



- Числове відображення

```
GradientBoostingRegressor: -0.413800316659257
RandomForestRegressor: -0.2049552066860277
XGBRegressor: -0.2926343853499918
XGBRFRegressor: -0.9255360017828651
CPU times: user 10.8 s, sys: 25.3 s, total: 36.1 s
Wall time: 8h 31min 43s
```

РЕЗУЛЬТАТИ ТРЕНУВАННЯ МОДЕЛІ ЗА ДОПОМОГОЮ РІЗНИХ МЕТОДІВ

Метод	MAPE	RMSE	Explained variance	Durbin-Watson
Gradient Boosting Regressor	**11.97**	**4.87**	**0.67**	**1.70**
Random Forest Regressor	**9.12**	**3.66**	**0.86**	**1.89**
XGBRF Regressor	**8.77**	**3.26**	**0.86**	**2.26**
XGB Regressor	**7.20**	**2.38**	**0.99**	**2.006**

- **MAPE** – похибка прогнозу у відсотках від фактичного значення.
- **RMSE** – квадратичне відхилення(чим менше тим метод ефективніший).
- **Explained variance** – відсоток пояснених моделлю варіацій(0 -1).
- **Durbin-Watson** – виявлення автокореляції:
  ≈ 2 – відсутність автокореляції;
  < 2 – позитивна автокореляція;
  > 2 – від'ємна автокореляція.

9

ВИСНОВКИ

1. Проаналізовано існуючі методи машинного навчання для тренування моделей прогнозування.
2. Було натреновано декілька моделей, щоб виявити який з алгоритмів буде найефективнішим для використання саме у сфері реклами.
3. Створено програмну реалізацію з візуальним відображенням ефективності кожного окремого алгоритму.
4. Найкращу модель було обрана за алгоритмом "XGBRegressor". Вона оцінювалася за декількома метриками, та показала наступні результатів: 7.20 за MAPE, 2.38 за RMSE та 2.05 за durbin-watson. Ці результати є хорошим результатом за критеріями оцінки алгоритмів машинного навчання, щоб використовувати натренеровану модель в прогнозуванні ефективності рекламних кампаній.

10

АПРОБАЦІЯ

- **Тези**

1. Кухаренко Ю.Д. Інтеграція методів машинного навчання в сфері реклами. // VI Міжнародної наукової конференції «Науковий простір: актуальні питання, досягнення та інновації». - м.Київ, 2023.

2. Кухаренко Ю.Д., Підвищення доходності реклами за рахунок методів прогнозування // VI Міжнародної наукової конференції «Науковий простір: актуальні питання, досягнення та інновації». - м.Київ, 2023.

11

ДЯКУЮ ЗА УВАГУ!

12