Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)
Кафедра Інформатики
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

ДОСЛІДЖЕННЯ МЕТРИК ТА ІНСТРУМЕНТУ GOOGLE ANALYTICS ДЛЯ ПІДВИЩЕННЯ ВІДВІДУВАННЯ САЙТІВ
(тема)

Виконав:
студент 2 курсу, групи ІНФм-20-1
Колосок Е.В.
(прізвище, ініціали)

Спеціальності 122 Комп'ютерні науки
(код і повна назва спеціальності)

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва освітньої програми)

Керівник доц. Руденко Д. О.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ Кобилін О.А.
(підпис) (прізвище, ініціали)

2021 р.

Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)

Кафедра Інформатики
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва освітньої програми)

ЗАТВЕРДЖУЮ:
Зав. кафедри
_____
(підпис)
«____»_____ 2021 р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Колоску Едуарду Володимировичу
(прізвище, ім'я, по батькові)

1.Тема роботи: Дослідження метрик і інструменту Google Analytics для підвищення відвідуваності сайтів

затверджена наказом по університету від «22» жовтня 2021 року № 1574Ст.

2. Термін подання студентом роботи до екзаменаційної комісії 30 листопада 2021 р.

3. Вихідні дані до роботи Дослідження метрик та інструменту Google Analytics для підвищення відвідування сайтів

4. Перелік питань, що потрібно опрацювати в роботі

1. Сучасний стан питання успіху веб-сайту. 2. Робота з даними Google Analytics для отримання та опрацювання даних. Метрики Google Analytics. 3. Опрацювання даних методами машинного навчання 4. Програмна реалізація та аналіз результатів прогнозування Відвідуванн

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) Теплові карти. Аудиторії. Демографія. Поведінка. Мета. Аналіз даних за допомогою звіту «Карта відвідувань»

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Консультант з дотримання діючих стандартів та норм	Доцент Бєлова Н.В.		

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Терміни виконання етапів роботи	Приміт ка
1	Отримання завдання на кваліфікаційну роботу	02.09.2021	
2	Аналіз завдання, підбір літератури	10.10.21-14.10.21	
3	Аналіз літератури з досліджуваної проблеми	15.10.21-18.10.21	
4	Програмна реалізація	01.11.21-09.11.21	
5	Оформлення пояснювальної записки	10.11.21-25.11.21	
6	Перевірка на плагіат	25.11.2021	
7	Рецензування	29.11.2021	
8	Підготовка презентації та доповіді	30.11.2021	
9	Занесення роботи в електронний архів	30.11.2021	
10	Попередній захист кваліфікаційної роботи	30.11.2021	

Дата видачі завдання 22 жовтня 2021 р.

Студент _____
(підпис)

Керівник роботи _____ _____доц. Руденко Д.О._____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ/ABSTRAC

Пояснювальна записка до кваліфікаційної роботи 69 с., 19 рисунків, 2 додатки, 21 джерел.

СЕССІЇ, СЕРЕДНІЙ ПОКАЗНИК ТРИВАЛОСТІ СЕССІЇ, ПОКАЗАТЕЛЬ ВІДМОВ, ДОХІД, ТРАНЗАКЦІЯ, КЛАСТЕРИЗАЦІЯ, GOOGLE ANALYTICS

Об’єктом дослідження є веб-система інтернет-магазину ювелірних виробів ручної роботи.

Метою дослідження є прогнозування збільшення доходу від користувачів на основі їх поведінки на сайті. Дані такого характеру дають змогу оптимізувати веб-систему під ту групу користувачів, які найвірогідніше куплять товар на веб-сайті.

Використано способи сбору даних про поведінку користувачів; робота с даними Google Analytics; розробка алгоритму опрацювання даних; реалізація алгоритму t-SNE; реалізація алгоритму k-means; вирішення задачі прогнозування конверсії на основі поведінки користувачів.

У результаті роботи була спроектована та реалізована модель прогнозування відвідуваннь на основі поведінкових метрик користувачів веб-системи.

SESSIONS, AVG. SESSION DURATION, BOUNCE RATE, REVENUE, TRANSACTION, CLUSTERIZATION, GOOGLE ANALYTICS

The object of the study is the web system of the online store of handmade jewelry.

The purpose of the study is to predict an increase in revenue from users based on their behavior on the site. This type of data allows you to optimize the web system for the group of users who are most likely to buy the product on the website.

Methods of collecting data on user behavior are used; work with Google Analytics data; development of data processing algorithm; implementation of the t-SNE algorithm; implementation of the k-means algorithm; solving the problem of conversion prediction based on user behavior.

As a result, a model for forecasting visits based on the behavioral metrics of web system users was designed and implemented.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

SEO – search engine optimization

CPC – ціна за клік – це сума, фактично сплачена за клік

RPC – це середній доход, який отримується за кожен клік по пошуковій рекламі

ROAS – це маркетингова метрика, яка вимірює ефективність рекламної кампанії

CTR – коефіцієнт, що показує, як часто люди, які бачать оголошення, клікають на нього

ВСТУП

Google Analytics - метричний сервіс, що дозволяє отримувати дані про кількість користувачів, аналізувати їх поведінку та дані про джерела.

Можливості Google Analytics:

– дозволяє відстежувати тренди;

– дає інформацію про джерело відвідувачів (з яких сторінок переходять) і яка їхня поведінка на сторінці;

– дає розуміння того, як конвертувати відвідувачів в покупців і клієнтів;

– відстежує, де і з якої причини користувачі залишають сторінку;

– показує, по яких запитах відвідувачі знаходять сайт;

– показує, який канал трафіку приносить більше доходу;

– дає список найпопулярніших сторінок сайту;

– дозволяє контролювати ефективність рекламних кампаній;

– дає розуміння того, який контент стимулює користувачів до дії або привертає увагу і багато іншого.

Google Analytics дозволяє розвивати ресурс, грунтуючись на отриманих даних від сервісу, а також дозволяє відстежувати ефективність проведеної рекламної кампанії.

Відвідуваність сайту – це число відвідувачів сайту за певний період часу, зазвичай за добу. Це важливий показник на який звертають увагу пошукові системи, тому його постійно відстежують системи статистики сайту. При визначенні рівня відвідуваності сайту в розрахунок беруться як унікальні відвідувачі, так і загальна кількість переглянутих ними сторінок сайту.

Підвищення відвідуваннь веб-системи прямо пропорційно до росту прибутку бізнесу. Саме тому впровадження раціонального підходу до оптимізації відвідуваннь веб-системи є критичним з точки зору розвитку бізнесу.

1 СУЧАСНИЙ СТАН ПИТАННЯ УСПІХУ ВЕБ-САЙТУ

1.1 Задача пошуку цільової аудиторії

Успіх сайту залежить від розміру потоку клієнтів, який дають пошукові системи, тому першочергово потрібно визначити портрет цільової аудиторії, яка буде відвідувати сайт. Маючи чітке розуміння того, хто ваші потенційні клієнти, скільки їм років, чим вони цікавляться, де проводять час, які у них страхи та радості, дозволяють бізнесу створити онлайн ресурс, який буде цікавий та необхідний майбутнім покупцям.

Існує безліч підходів до визначення цільової аудиторії, але один з них найефективніший та деталізований – карта емпатії. Карта емпатії – інструмент візуалізації ідей, що дозволяє поставити себе на місце користувача, поглянути на проблему, яку вирішує продукт, його очима. Карта емпатії — це схема, яка проілюстрована на рисунку 1.1, в центрі якої розміщується представник певного сегменту користувачів, по різні боки від нього – 4 блоки («думаю та відчуваю», «говорю та роблю», «бачу», «чую»). Висновки наводяться в двох додаткових блоках: «проблеми та больові точки» і «цінності та досягнення». Інформація розподіляється по блокам наступним чином:

– думаю та відчуваю: що турбує користувача? Якими словами він думає про проблему? Щодо чого сумнівається? Цю інформацію краще шукати там, де користувачі скаржаться: наприклад, на форумах. Якщо у продукту є служба підтримки, варто поспілкуватися з її представниками / почитати, що туди пишуть користувачі;

– кажу та роблю: як користувач поводиться публічно? Що кажуть? Якими способами вирішують проблему? Цю інформацію варто шукати в соціальних мережах. Особливо в цьому плані корисний Linkedin: там можна відносно легко знайти представників найрізноманітніших

професій і вивчити, які події/конференції/навчальні заходи вони відвідують, в яких спільнотах складаються, які питання в них піднімають, якими навичками володіють;

– бачу: на що схоже середовище, в якому знаходиться користувач? З якими пропозиціями і альтернативами вашого продукту він стикається? Джерело інформації для цього блоку – сама система та системи-конкуренти, онлайн і оффлайн реклама, статті у фахових співтовариствах, офіційні огляди;

– чую: як середовище, в якому знаходиться користувач, впливає на нього? Що кажуть йому колеги, знайомі, авторитетні для нього джерела? Які медіаканали мають вплив на користувача? На відміну від блоку «бачу», інформація тут не обов'язково відповідає дійсності. Але користувач їй довіряє. Де шукати інформацію для цього блоку: чутки та думки на форумах (а власне з соціальних мереж можна дізнатися, що це за форуми), історії успіху, стереотипи та навіть міські легенди;

– проблеми та больові точки: що турбує користувача? Чого він побоюється? Що може стати причиною того, що він відмовиться від вашого продукту? Часто джерелом інформації для цього стає блок «Думаю та відчуваю». Всі ці страхи та сумніви потрібно буде розвіяти, і зробити це можна різними способами: від «правильного» тексту в інтерфейсі до індивідуальних консультацій;

– цінності та досягнення: що допоможе користувачеві позбутися від проблем і сумнівів? За які можливості продукту він готовий платити? Які цінності ми повинні транслювати? Висновки з цього блоку впливають на продукт з самих різних сторін: вони можуть спричинити як дрібні зміни в інтерфейсі або в тексті, так і додавання / виключення певних функціональних особливостей, і іноді – навіть зміна в позиціонуванні продукту. (Рисунок 1.1)[2]

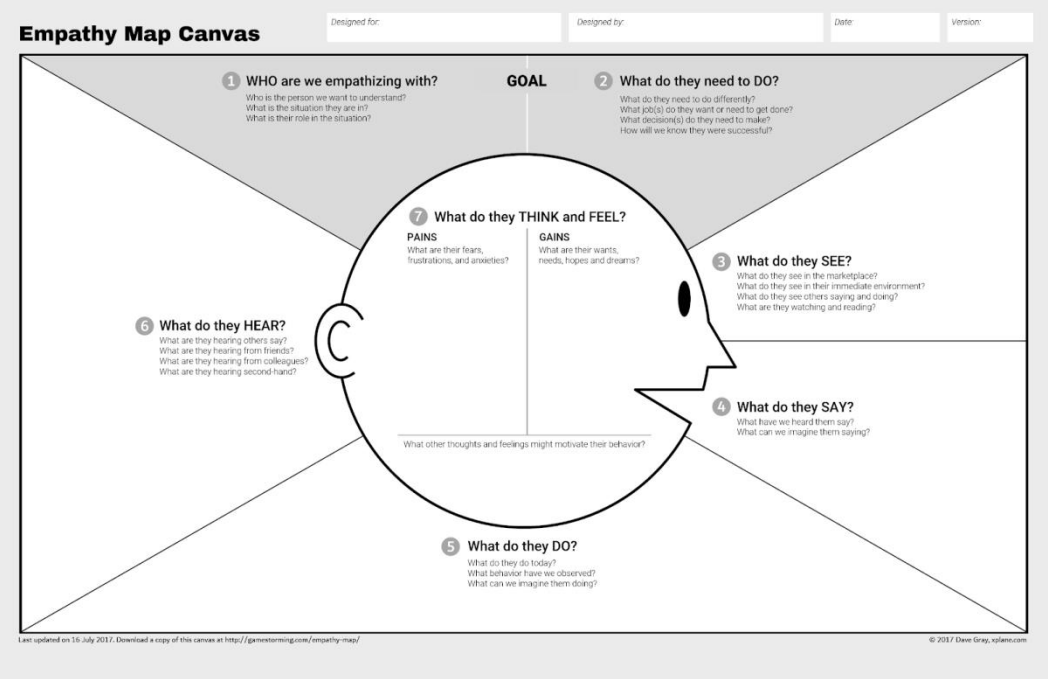


Рисунок 1.1 – Приклад карти емпатіїї (Empathy Map Canvas)

На основі отриманих даних про основні цільову аудиторію починається побудова архітектури та дизайну сайту. Схильності користувачів, їх вибір і переваги все більше стають важливими для систем пошуку. Пошукові системи враховують кількість сторінок, переглянутих користувачем, кількість часу, проведеного на сайті, а також коло інтересів і активність участі конкретного користувача в мережі.

Успіх бізнесу онлайн залежить від безлічі факторів. В першу чергу це конкурентоспроможність. Бізнесу у непопулярній ніші набагато простіше вийти в “топ” онлайн, чим зайняти перші позиції по давно відомим областям, таким як продаж електронної техніки або доставка їжі. Тому аналіз конкурентних сайтів часто займає одну з головних ролей в хорошому просуванні свого бізнесу онлайн. Переглядаючи їх структуру, різні рішення та поведінку, можна вигідно застосувати отриману інформацію.

Також на успіх в діджиталізації бізнесу впливає якість виконаної веб-системи. В першу чергу це швидкість завантаження сторінки, дружній і зрозумілий user interface і адаптивність під мобільні пристрої. Падіння

швидкості безпосередньо впливає на кількість відмов – в сучасному світі люди відмовляються заходити на повільні сайти. Люди не знають точного часу завантаження, але можуть оцінити швидкість завантаження сайту за своїми відчуттями. В середньому близько 83% відвідувачів залишають веб-сайт, якщо він завантажується більше 3 секунд. З липня 2018 року всі сторінки, які завантажуються повільніше, ніж за 3 секунди, були штучно опущені в результатах пошуку. В нормах швидкості для завантаження сайту є великі діапазони, які пов'язані з різницею баз даних і інших внутрішніх особливостей. Зазвичай для десктопних версій це менше 1 секунди, а для мобільних пристроїв менше 1,6 секунди.

Працювати для підтримки рейтингу серед пошукових систем потрібно завжди, додаючи матеріали, створюючи нові сектори сайту, не забуваючи оновлювати старі. Посилання на сайт також мають велику важливість, пошукові системи їх високо оцінюють.

1.2 Трекінгові системи

Неможливо оцінити важливість даних при прийнятті бізнес-рішень і проведенні маркетингових кампаній. Тому існує десятки онлайн сервісів і додатків для збору та відстеження даних по веб-сайтам.

Як відстежувати продажі з сайту? З якого каналу приходять покупці? Який контент варто розвивати? На ці та інші питання про взаємодію користувачів з сайтом дасть відповідь Google Analytics.

1.2.1 Google Analytics

Google Analytics — метричний сервіс, що дозволяє отримувати дані про кількість користувачів, аналізувати їх поведінку та дані про джерело.[3]

Можливості Google Analytics:

– дозволяє відстежувати тренди;

– дає інформацію про джерело відвідувачів (з яких сторінок переходять) і яка їхня поведінка на сторінці;

– дає розуміння того, як конвертувати відвідувачів в покупців і клієнтів;

– відстежує, де і з якої причини користувачі залишають сторінку;

– показує, по яких запитах відвідувачі знаходять сайт;

– показує, який канал трафіку приносить більше доходу;

– дає список найпопулярніших сторінок сайту;

– дозволяє контролювати ефективність рекламних кампаній;

– дає розуміння того, який контент стимулює користувачів до дії або привертає увагу і багато іншого.

Google Analytics дозволяє розвивати ресурс, грунтуючись на отриманих даних від сервісу, а також дозволяє відстежувати ефективність проведеної рекламної кампанії.

1.2.2 Теплові карти

Коли виникає питання підвищення відвідуваннь сайту, на допомогу приходять теплові карти. Теплова карта сайту – це інструмент, який використовує колірну палітру для візуалізації даних на графіку. Наприклад, якщо ви проглядуєте веб-сторінку та хочете знати, які елементи привертають найбільше уваги, теплова карта покаже цю

інформацію на підставі призначених для користувача даних відвідувачів цієї сторінки. Теплова карта використовує колірний спектр від теплих до холодних тонів для демонстрації ділянок сторінки сайту, які привертають увагу користувачів (Риунок 1.2)[3].

Рисунок 1.2 – Приклад роботи теплових карт

Існує кілька різновидів теплових карт, в залежності від функціоналу сервісу, який надає дані.

Класична теплова карта. За допомогою теплового градієнта показує ділянки сторінки, які залучають найбільше уваги.

Скролл-карта. Яка довжина контенту? Чи читають користувачі все, що написано? Де найкраще розмістити CTA-кнопку? Чи варто використовувати лонгріди? Чи використовувати нескінченну прокрутку

або додати пагінацію? На всі ці питання можна отримати відповідь за допомогою скролл-карти.

Конфетті. Досить специфічна версія класичної теплової карти. На карті відображаються ділянки з кліками користувачів, які варіюються в залежності від кольору, типу дії.

До популярних сервісів теплових карт відносять Hotjar та Feng GUI.

Hotjar – це онлайн-сервіс для аналітики сайту та накопичення зворотного зв'язку. Він створений допомогти зрозуміти потреби відвідувачів і збільшити відвідування. Система підійде для UX і UI дизайнерів, розробників, менеджерів, аналітиків, маркетологів і ін.

Можливості Hotjar:

– карти кліків;
– відстеження поведінки відвідувачів;
– воронки конверсій;
– аналітика форм;
– зворотній зв'язок;
– опитування;
– проактивний чат;
– порівняння кліків;
– створення FAQ на основі відгуків;
– робота над формами для генерації лідів;
– перегляд реакцій відвідувачів на довжину контенту;
– прямий чат з відвідувачами;
– A / B-тести;
– визначення необхідності видалення, переміщення або зміни контенту;
– аналіз джерел трафіку;
– поради щодо поліпшення UX, в тому числі на різних дозволах екрану.

За допомогою Feng GUI карти візуального сприйняття, система показує елементи зображення, які будуть помічені людиною в першу чергу. Дана технологія, звичайно, не є останньою інстанцією – є ще безліч інших методів аналізу і просто хороший смак дизайнера. Однак, Feng-GUI може бути корисна веб-дизайнерам для тестування прототипів сайтів, а також фотографам, відео-операторам і навіть архітекторам, які з її допомогою можуть визначати фокальні точки в просторі. Технологія, використовувана Feng-GUI, заснована на визначенні особливостей (saliency detection), яка застосовується в нейронному аналізі (neuroscience). Відповідно до цієї теорії, живі організми звертають увагу в першу чергу на ті групи об'єктів, які виділяються із загальної кількості.

1.3 Механізми просування сайтів

Кожному сайту необхідно та важливо перебувати в «топі» пошукових систем, щоб якомога більша частина аудиторії змогла відвідати його. Для цього найчастіше використовують SEO-просування, але це займає досить багато часу. Зазвичай, щоб отримати результати від SEO оптимізації сайту, потрібно більше 3-х місяців. Більш швидкий спосіб отримання релевантного трафіку на сайт – реклама в Google і соціальних мережах, так званий Pay per Click медіа-канал. Social Media Marketing дозволяє розширити можливості фірми виходом інформації в соціальні мережі та спільноти, а за допомогою Email маркетингу можна відправляти потенційним клієнтам актуальну інформацію про продукт або сервіси. Нижче розглянемо кожен з каналів просування сайтів більш докладно.

1.3.1 Search Engine Optimization

Пошукова оптимізація (SEO) – це процес роботи над сайтом, його внутрішніми факторами, що впливають на ранжування в пошукових системах – структурою, контентом, кодом HTML, його зовнішніми факторами ранжування – посиланнями на сайт з метою збільшення релевантності ресурсу визначеними, заздалегідь відомим ключовими словами, збільшення популярності сайту для пошукових систем і, відповідно, збільшення позицій в пошукових результатах для залучення більшої кількості відвідувачів на сайт.

За часів 1990-х пошукові системи надавали великого значення тексту на сторінці, ключовими словами в мета-тегах і іншим внутрішнім чинникам, якими власники сайтів могли легко маніпулювати. Це призвело до того, що у видачі багатьох пошукових систем перші кілька сторінок зайняли сайти, які були повністю присвячені рекламі, що різко знизило якість роботи пошукових систем і привело багатьох з них до занепаду. З появою технології PageRank більше ваги стало додаватися зовнішнім факторам, що допомогло Google вийти в лідери пошуку у світовому масштабі, ускладнивши оптимізацію за допомогою одного лише тексту на сайті.

Одним з важливих факторів при розробці методики просування сайту є вивчення алгоритмів роботи пошукових систем, так як на їх основі складаються алгоритми SEO-оптимізації та просування сайтів.

Алгоритм ранжирування Google є одним з найбільш інтелектуальних і складних. Просування сайтів в Google, особливо на початковому етапі, набагато складніше, ніж в інших. Розкрутка сайту в Google процес не із легких, так як на нові web-ресурси накладається фільтр (так звана «пісочниця»). Google при індексуванні та ранжируванні сайту використовує більше двохсот факторів, на жаль оптимізатор може вплинути далеко не на всі. Головна перевага пошукової системи Google —

це її стабільність щодо своїх конкурентів в плані зміни алгоритму та апдейтів. Інформація, тільки що розміщена на сайті, може протягом декількох хвилин потрапити в основну видачу. Пошукові роботи Google в три рази швидші, ніж роботи інших пошукових систем. Фільтри майже не змінюються з моменту початку їх впровадження.

Головними факторами, на які в силах впливати оптимізатор, є посилальна маса і унікальне наповнення сайту. Релевантність сайту щодо пошукового запиту підвищується у випадках: використання ключових слів у заголовках. У тезі [title] необхідно прописувати ключову фразу, що відповідає даній сторінці з використанням додаткових слів.

Облік зовнішніх посилань пошукова система Google враховує в наступному порядку: кількість, авторитетність сайту-донора (тобто наскільки пошукова система довіряє сайту), тематика сайту. Найвагомішими посиланнями для пошукової системи Google є так звані наскрізні посилання. Це означає, що посилання розміщені абсолютно на всіх сторінках сайта. Виходячи з цього випливає, що одне наскрізне посилання, в залежності від кількості сторінок на сайті, має вагу більшу, ніж десять або навіть сто посилань з інших ресурсів.

Поняття Google PageRank є одним з ключових моментів в роботі пошукової машини Google. Поряд з безліччю інших параметрів, що впливають на видачу сайтів в результатах пошуку, знання моделі розрахунку PageRank необхідно як для розуміння процесу пошуку, так і для використання оптимізаторами при просуванні своїх сайтів в пошуковій системі.

PageRank — це числова величина-міра «важливості» сторінки в пошуковій системі Google. Дана величина цілком і повністю залежить тільки від кількості і від якості вхідних посилань. Більше значення для пошукової системи Google є саме якість. PageRank — це алгоритм розрахунку авторитетності кожної, окремо взятої сторінки,

використовуваний пошуковою системою Google. PageRank на даний момент не є головним фактором в ранжування сайту, але є дуже важливим.

Алгоритми пошукової системи Google враховують не всі посилання, а тільки з якісних ресурсів. Так само існує ймовірність, що посилання не тільки не буде врахованим, а може навіть зробити негативний вплив на ранжування сайту (пошукова песимізація). Основна формула для розрахунку PageRank:

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)}\right), \quad (1.1)$$

де $PR(T_i)$ – значення PageRank для сторінки;

d – демпфуючий коефіцієнт. Показує, яку частку ваги, передає сайт-донор на сайт-акцентор. Прийнято приймати його рівним 0.85, це означає, що сторінка може передати максимум 85% ваги. У деяких джерелах d показує ймовірність, з якою користувач здійснить перехід на один із сайтів-акценторов, а не закриє браузер, що, фактично, те ж саме. Точне значення цього коефіцієнта відоме тільки в Google, експерименти показали, що він дорівнює 0,85;

n – кількість сторінок, які не знаходяться під фільтром, що посилаються на сайт-акцентор;

T_i – сторінки, що посилаються;

$C(T_i)$ – кількість зовнішніх посилань на сторінці T_i.

Через те, що число посилань на сайті може бути велике, а кількість сторінок у пошуку Google, досягає величезних значень (близько десяти трильйонів штук), і з кожною годиною кількість проіндексованих сторінок збільшується, то уявлення *PR* в абсолютних значеннях було б грубою помилкою. Для правильного розрахунку була введена формула TLPR – ToolBar PageRank – це значення може варіюватися в інтервалі від 0 до 10.

Для того, щоб укласти все значення ваги сторінок в проміжку від нуля до десяти використовується логарифмічна шкала.

Визначається ToolBar PageRank за формулою:

$$TLPR = \log_{base}\left(PR\right)*a\ , \qquad (1.2)$$

де *base* – основа логарифма, значення залежить від кількості сторінок в пошуковій видачі сайту;

a – коефіцієнт приведення, що задовольняє нерівності $0 < a \leq 1$.

З вищесказаного помилково робити висновки, що при $TLPR = 0$ значення PageRank теж дорівнює 0. За формулою PR можна зробити висновок, що при $n = 0$ мінімальне значення PRmin = $(1–d) = 0{,}15$. Дане значення еквівалентно $TLPR = -1$. При значеннях $TLPR < 0$ приймається значення $PR = N/A$ (значення не визначене), навіть при такому значенні виявляється вплив на поширення ваги посилань на сайті-акценторе. Тулбарне значення використовується тільки для відображення веб-майстром в Google Toolbar і вплив на ранжирування не надає.

Вплив на позиції сайту надає тільки реальний PR сторінки. Використовуючи дану інформацію легко порахувати, скільки потрібно посилань для досягнення сайта того чи іншого PR. Так само є можливість прогнозувати PR. В цілому, існує безліч варіантів збільшення ваги своїх сторінок, але головна – це якісні посилання з інших сайтів. Для цього враховуються посилання з каталогів, сайтів, форумів, блогів. Однак не слід розміщувати посилання на всіх ресурсах, так як крім PR враховуються ще безліч факторів, що впливають на сторінки в результатах пошуку, наприклад Trust Rank.

Trust Rank вказує рейтинг довіри пошукової системи до сайту. Даний рейтинг впливає на ранжування сайту в пошукових системах. Вихідні

посилання з нашого сайту повинні або бути закриті або відсутні повністю, так як з цих посилань перетікає вага на інші сайти, чого нам, в принципі не потрібно.

Положення сайту в пошуковій системі залежить від зовнішніх і внутрішніх факторів. До внутрішніх факторів відноситься робота, яка перебуває під контролем власника сайту. Для поліпшення роботи необхідно приводити в порядок тексти та розмітки сторінок, поліпшення якості та кількості тексту на сайті, перевіряти унікальність тексту, його стилістичне оформлення та інше.

До зовнішніх факторів належать релевантність сайту, яка визначається кількістю цитування його веб-ресурсами.

1.3.2 Pay per Click

Контекстна реклама (англ. «PPC marketing» від «pay-per-click») – вид інтернет-реклами, в якому рекламодавці несуть витрати, коли користувачі переходять на сайт через рекламні оголошення. Рекламодавці задають ставки на ключові слова та типи аудиторій, які пошукова система зіставляє з пошуковими запитами користувачів і попередньо визначеними списками користувачів, а потім відбувається показ реклами.

На ціну кліка може впливати безліч факторів. По-перше, це саме оголошення, а саме його якість, показник ефективності (Quality Score). Необхідно, щоб оголошення відповідало запиту, який користувач вводить в рядку пошукової системи. Важливо, щоб запитам і самому оголошенню відповідала початкова сторінка, на яку потрапляє користувач після того, як клікає по оголошенню.

Також необхідно враховувати конкурентну активність за ключовими словами, регіони, в яких ви хочете показувати ваше оголошення і час доби, в яке хочете рекламуватися.

Контекстна реклама найчастіше використовується для:

– підвищення кількості продажів;
– залучення потенційний кліентів / заявок / дзвінків;
– просування бренду.

Ключова особливість контекстної реклами – релевантність. Користувачі постійно шукають певні товари, послуги та інформацію. У рекламодавців є можливість показувати рекламні оголошення точно в той момент, коли відбувається пошукова сесія. Наприклад, якщо користувач шукає «юридичну фірму в Харкові», то рекламодавець може у відповідь на цей запит показати рекламне оголошення, яке розповідає про юридичні фірми в Харкові.

До основних рекламних платформ відносяться Google Ads, Яндекс Діректа та Bing Ads. Google Ads реклама показується в пошуку Google, на пошукових партнерах Google і сайтах у Медійній мережі (КМС, англ. «Display Network»). Google Ads займає перше місце в Україні (частка 62%). Google Ads запущений в 2001 році. Яндекс Директ реклама показується в пошуку Яндекса, на пошукових партнерах Яндекса (в тому числі в пошуку @ mail.ru) і сайтах в рекламній мережі Яндекса (РМЯ). Яндекс Директ посідає друге місце в Україні (частка 29%). Яндекс Директ був запущений в 2001 році (контекстна реклама в пошуку Яндекса стала показуватися в 1998 році). Рекламу в Яндекс Директі використовують як невеликі рекламодавці зі сфери малого бізнесу, так і представники великого бізнесу. Bing Ads — це рекламна платформа Microsoft з оплатою за клік, що забезпечує показ пошукової реклами в системах Bing, Yahoo і MSN.

1.3.3 Social Media Marketing

SMM, або Social Media Marketing, – це сукупність дій, спрямованих на просування певної торгової марки, компанії, організації чи ідеї на ринку

за допомогою активної роботи в соціальних мережах різних типів та спрямувань.[3] Фактично під SMM мається на увазі регулярне, емоційно та інформативно насичене спілкування з конкретною цільовою аудиторією в соціальних мережах. Для цього варто використовувати саме ті соціальні мережі, які здатні найширше розкрити основну ідею бізнесу та максимально повно відповідати потребам та інтересам кола потенційних клієнтів та прихильників.

Для того, щоб SMM було вдалим, по-перше потрібне ретельне вивчення аудиторії соціальної мережі. Наприклад, якщо оптимізатор розкручує сайт пейнтбольного клубу, то на форумі вишивання хрестиком успіх оптимізації буде вельми сумнівний. Втім, все залежить від таланту оптимізатора. По-друге, у великих соціальних мережах, подібних Facebook і Twitter, потрібно вибирати свою цільову аудиторію. На це потрібно чимало часу, але його витрати, як правило, повертаються дієвою увагою не всіх користувачів підряд, а тільки зацікавлених осіб. По-третє, за свідченням фахівців, найбільшого успіху досягають оптимізатори, які витрачають зусилля не тільки на оптимізацію свого сайту, але і на розвиток того ресурсу, в якому здійснюють просування SMM.

1.3.4 Email Marketing

Email маркетинг – важливий інструмент роботи з цільовою аудиторією та просування в інтернеті, що сприяє прямому спілкуванню між бізнесом і покупцями. Метою таких зусиль є зміцнення лояльності та зростання кількості продажів.[2]

Незважаючи на тривалість існування та появу багатьох різних нових маркетингових практик, поштовий маркетинг залишається важливим інструментом в арсеналі інтернет-маркетолога.

В основі цієї практики лежать різні типи електронних повідомлень, основними з яких є наступні:

– welcome email. Вітальним є електронне повідомлення, яке адресат отримує після підтвердження підписки. Таке є гарантом ефективності електронної розсилки, від правильності виконання якого залежить зацікавленість одержувачів в подальшій комунікації та загальне враження про бізнес. Не варто недооцінювати важливість welcome emails – вони відкриваються в 4 рази частіше звичайних, а цільові дії в них виконуються в 5 разів частіше;

– інформаційне послання (Informational Letter). Один з найбільш частих форматів електронних повідомлень. Може зміцнити лояльність за рахунок якісного контенту. Поширювати варто корисні матеріали, розширення та багато іншого;

– дайджест (Digest). Може здатися схожим на інформаційні emails, проте має дещо інший формат – є коротким оглядом нової інформації (наприклад, email зі списком найбільш популярних товарів за тиждень);

– commercial Letter. Як правило, призначене для прямих продажів. Контент такого може бути пропозицією певного оффера та його опис, рекомендаціями товарів, заснованих на перевагах / попередніх покупках користувача, або інформацією про промо-акції. Маркетолог тут повинен бути максимально обережним, тому що занадто агресивна стратегія може викликати в одержувачів невдоволення, внаслідок чого emails будуть частіше відзначатися як спам. В результаті це не тільки знизить ефективність кампанії, а й зменшить показник доставлених листів;

– розсилка. Серія листів, які поступово збільшують зацікавленість одержувача до офферу, тим самим підвищуючи готовність до угоди. Згідно зі статистикою, серійні email-кампанії мають коефіцієнт КЕП в 2-3 рази вище, ніж стандартні комерційні послання. Варто зауважити, що серійна розсилка повинна бути заснована на ретельно та

правильно складеній стратегії – це єдина умова, при якому кампанія може принести позитивний результат. Суть полягає в тому, щоб створити бесіду, кожне з повідомлень якої буде корисним, цікавим і викликати інтерес до наступного повідомлення, в той же час збільшуючи привабливість оффера;

– тригери (Triggers). Emails, що відправляються за певних умов або виконанні користувачем певної дії: наприклад, якщо користувач SaaS-рішення для менеджменту фінансів вперше генерує місячний звіт бюджету, йому відправляється міні-керівництво з рекомендаціями щодо правильного аналізу та порадами щодо економії грошей;

– транзакційні (Transactional). Листи, що відправляються при виконанні користувачем певної транзакції;

– пряма комунікація. У даному випадку представник компанії відправляє користувачеві особисте послання. Email є персоналізований, його контент написаний в манері простого спілкування. В ньому може бути запропонована допомога, висловлена подяка за лояльність, представлені рекомендації.

1.4 Постановка задачі

Дане дослідження присвячене прогнозу росту відвідваннь веб-системи у зв’язку з кількісними та поведінковими метриками. Датасет був отриманий за допомогою багатофункціонального сервісу для аналізу веб-систем та додатків – Google Analytics. Прогнозування відвідуваннь веб-сайту буде вирахуване на основі кількісних та поведінкових метрик. До кількісних метрик можна віднести кількість користувачів веб-системи, до поведінкових метрик – такі параметри, як середня тривалість сесії, показник відмов, кількість проглянутих сторінок за одну сесію. Дискретизаційний період – 3 місяці.

Кожного користувача можно описати такими атрибутами як Sessions, Avg. Session Duration, Bounce Rate, Revenue, Transaction. Так як в п’ятимірному вимірі описати користувача не вийде, тому було реалізовано алгоритм t-SNE, який п’ятимірний вимір переводить в двовимірний. Таке перетворення дає можливість візуально відобразити датасет та зрозуміти, чи є якісь групи серед користувачів.

Наступним шагом є кластеризація. Основна мета – розділити користувачів на групи, близькі за поведінкою.

Останнім кроком є побудова прогнозуючої модель: залежніть доходу сайту від поведінкових факторів користувачів.

Реалізація підходу буде здійснена за допомогою інтерпретованої об'єктно-орієнтованої мови програмування високого рівня зі строгою динамічною типізацією – Python. Структури даних високого рівня разом із динамічною семантикою та динамічним зв'язуванням роблять її привабливою для швидкої розробки програм, а також як засіб поєднування наявних компонентів. Python підтримує модулі та пакети модулів, що сприяє модульності та повторному використанню коду.

2 РОБОТА З ДАНИМИ GOOGLE ANALYTICS: ЯК ОТРИМАТИ ТА ЯК ОПРАЦЮВАТИ ДАНІ

2.1 Метрики Google Analytics

Підвищення відвідуваннь веб-системи прямо пропорційно до росту прибутку бізнесу. Саме тому впровадження раціонального підходу до оптимізації відвідуваннь веб-системи є критичним з точки зору розвитку бізнесу.

Для отримання даних по веб-сайту був використаний багатофункціональний сервіс для аналізу веб-систем та додатків – Google Analytics. Трекінговий javascript код був встановлений на веб-сайт. Кожного разу, коли користувач завантажує сторінку сайту, в його браузері виконується код відстеження. Під час першого візиту він записує в браузер відвідувача cookie-файл, який містить унікальний ідентифікатор користувача – Client ID. Завдяки cookie-файлам всі наступні заходи з того ж браузера будуть зараховані системою Google Analytics як повторні відвідування. [3]

Датасет містить дані про сесії кожного користувача та включає в себе кількісні та поведінкові метрики.

До кількісних метрик можна віднести загальну кількість користувачів та кількість активних користувачів веб-системи, до поведінкових метрик – такі параметри, як середня тривалість сесії, показник відмов, кількість переглянутих сторінок за одну сесію.

Щоб система Google Analytics могла співвідносити користувачів із трафіком, з кожним зверненням до системи надсилається унікальний ідентифікатор, пов’язаний із користувачем. Роль ідентифікатора може виконувати одиничний основний файл cookie з назвою _ga, що зберігає ідентифікатор клієнта Google Analytics. Разом з ідентифікатором клієнта

можна також використовувати функцію User ID, щоб точніше визначати користувачів на всіх пристроях, де вони переглядають сайт або використовують додаток.

Середня тривалість сеансу – це загальна тривалість усіх сеансів (у секундах), поділена на кількість сеансів. [2]

Тривалість окремого сеансу обчислюється по-різному залежно від наявності звернень до взаємодій на останній сторінці сеансу. [3]

Щоб обчислити середню тривалість сеансу, система Analytics підсумовує тривалість кожного сеансу протягом вказаного діапазону дат і ділить цю суму на загальну кількість сеансів.

Відмова – це сеанс із переглядом однієї сторінки сайту. Система Analytics реєструє як відмову сеанс, під час якого ініціюється лише один запит до сервера Analytics (наприклад, коли користувач відкриває одну сторінку сайту та покидає його, не ініціювавши за цей сеанс жодного іншого запиту).

Показник відмов – це відсоток сеансів, під час яких користувачі переглянули тільки одну сторінку сайту, не ініціюючи додаткових запитів до сервера Analytics. Цей показник обчислюється діленням кількості сеансів із переглядом тільки однієї сторінки на загальну кількість сеансів на сайті.

Для сеансів із переглядом однієї сторінки реєструється тривалість 0 секунд: оскільки звернення тільки одне, система Analytics не може зафіксувати іншу тривалість сеансу.

Якщо для успішної взаємодії з сайтом користувачі мають переглядати більше однієї сторінки, то високий показник відмов свідчить про необхідність покращення. Наприклад, якщо з домашньої сторінки користувач може перейти до інших розділів сайту (наприклад, статей новин, сторінок продуктів, сторінок оформлення покупки), але високий відсоток користувачів переглядає тільки її, високий показник відмов небажаний.

З іншого боку, якщо сайт складається з однієї сторінки (на зразок блогу) або надає доступ до інших типів вмісту на тій самій сторінці, високий показник відмов не свідчить про будь-які проблеми.

Якщо загальний показник відмов високий, можна провести докладний аналіз, щоб визначити, чи він однаково високий на всіх рівнях, чи тільки для одного-двох каналів, окремих пар "джерело/засіб" або кількох сторінок.

Наприклад, якщо проблема виникла лише з кількома сторінками, потрібно перевірити, наскільки їх вміст відповідає маркетинговим заходам, за допомогою яких залучаються відвідувачі. Також потрібно перевірити, чи зручно користувачам здійснювати на цих сторінках цінні для бізнесу дії.

Якщо високий показник відмов спостерігається для певного каналу, треба оцінити відповідність своїх маркетингових заходів. Наприклад, якщо користувачі, перейшовши на сайт із Медійної мережі, одразу покидають його, потрібно переконатися, що оголошення відповідають вмісту сайту.

Якщо масштаб проблеми ширший, треба перевірити застосований код відстеження, щоб переконатися, що всі сторінки правильно позначаються тегами. Також можна заново обміркувати дизайн сайту, перевірити відповідність текстів, графіки, кольорів, закликів до дії та видимість важливих елементів сторінок.

Сеанс – це час, який користувач приділив сайту чи додатку. Якщо користувач не виконує жодних дій протягом 30 хвилин, усі подальші дії за умовчанням реєструються для нового сеансу. Якщо ж користувач залишає сайт і повертається протягом 30 хвилин, це вважається продовженням початкового сеансу.[3]

2.2 Архітектура Google Analytics

Архітектура Google Analytics є ієрархічною та представлена на рисунку 2.1. Акаунт – це точка доступу до GA, верхній рівень ієрархії. Один акаунт може містити один або кілька ресурсів. (Рисунок 2.1)[9]

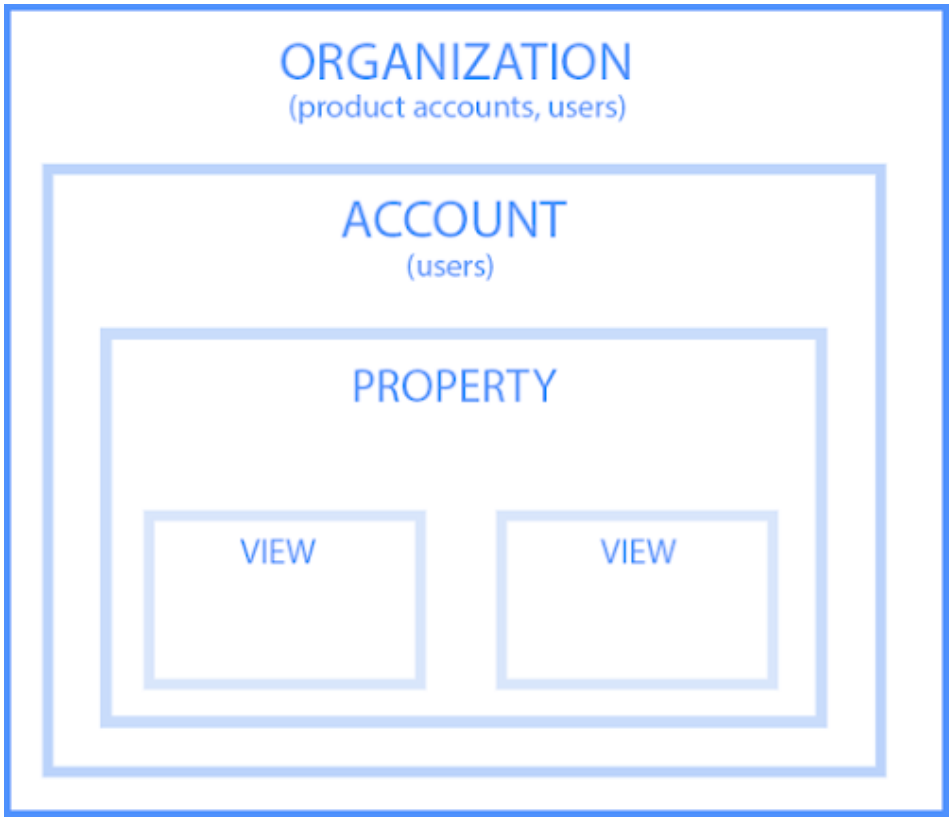


Рисунок 2.1 – Архітектура Google Analytics

Ресурсом може бути веб-сайт, мобільний додаток або пристрій (наприклад, кіоск або касовий термінал). Коли потрібно додати ресурс до облікового запису, GA створює код відстеження, необхідний для збору даних по ресурсу.

Представлення – це набір даних про ресурс. У ресурса може бути кілька представлень, наприклад:

– представлення з усіма даними по сайт;

– представлення, в яке потрапляють тільки сесії с джерелом — Google Ads;

– представлення з даними тільки по трафіку для субдомена.

Налаштування облікового запису дозволяє налаштувати доступ до даних для всіх ресурсів і представлень, що знаходяться в обліковому записі.

Керування користувачами дає право на перегляд, редагування, запрошення нових користувачів для обраних нових користувачів. Варто мати на увазі, що надаючи права на рівні облікового запису, ви відкриваєте доступ до всіх ресурсів і уявленнь, що знаходяться в ньому.

Всі фільтри. Створення глобальних фільтрів, які можна застосувати до всіх ресурсів і уявленням, що знаходяться в цьому акаунті.

Історія змін. Показує ким, коли і які зміни були зроблені. Подібний лог відмінно рятує в спірних ситуаціях, коли хтось ще має доступ до акаунта і може вносити свої правки.

Кошик. Містяться всі віддалені об'єкти протягом 35 днів. Після чого вони остаточно видаляються.

Налаштування ресурсу. Тут можна знайти ідентифікатор, присвоєний розглянутого ресурсу, змінити його назву, відкоригувати URL-адресу, вибрати, яке з представлень використовувати за замовчуванням і змінити галузь.

Розширені налаштування дозволяють заборонити автоматичну позначку GCLID (Google Click Identifier – ідентифікатор кліка Google) посилань для Google Ads та DoubleClick (платформа для програми покупки медійної реклами).

Функція поліпшеної атрибуції посилань дозволяє розрізняти прямі посилання на одну сторінку. Використання цього налаштування дозволяє одержувати більш точну картину поведінки користувачів, використовуючі посторінкову аналітику.

Вибір режиму для звіту за статистикою сторінки дозволяє аналізувати візуальну карту кліків різних елементів сторінок сайту.

Важлива частина налаштування відображення: зв'язок ресурсу з Search Console. Для того, щоб отримувати дані з Search Console в Google Analytics, необхідно додай сайт в панель вебмайстрів, після чого підтвердити їх зв'язок.

Керування користувачами дозволяє давати права на перегляд і редагування іншим користувачам на рівні ресурсу. Запрошені люди матимуть доступ до конкретного ресурсу та всім уявленням, пов'язаним з цим ресурсом.

Код відстеження – у цих налаштуваннях знаходиться сам скрипт Google Analytics, який необхідно інтегрувати з сайтом або додатком, що відслідковується.

Збір даних – можна заборонити або дозволити збір додаткових даних про трафік.

Зберігання даних дозволяє налаштовувати термін зберігання призначених для користувача даних: відомості з файлів cookie, ідентифікатори користувачів і рекламодавців. За замовчуванням термін зберігання дорівнює 26 місяцям, але можна задати власний термін зберігання, що дорівнює 14, 26, 38, 50 місяців або поставити без терміну дії.

User ID дозволяє зв'язувати дані про дії користувачів з різних пристроїв і відомості про повторні відвідування. Дозволяє більш точно визначати кількість користувачів в звітах.

Налаштування сеансу дозволяє вручну задати тривалість, тобто після якого часу буде автоматично закритий сеанс. Подібні налаштування будуть корисні власникам онлайн-кінотеатрів, де час очікування повинен бути більше ніж 30 хвилин, заданих за замовчуванням.

Джерела звичайних результатів пошуку дозволяє додати більше пошукових систем в список джерел результатів пошуку за замовчуванням.

Таким чином, користувачі, які переходять на сайт з пошукової системи, внесеної до списку пошукових систем за замовчуванням, відображаються в звітах як безкоштовний трафік.

Список виключених джерел переходу – після додавання обраних доменів, трафік з них буде виключений зі звітів про трафік переходів.

Список виключених пошукових запитів – після додавання обрані пошукові запити будуть виключені як джерело звичайного трафіку.

Зв'язок з Google Ads – налаштування зв'язку для отримання даних в звіти по Google Ads.

Зв'язок з AdSense – після зв'язування акаунтів AdSense і Google Analytics їх дані будуть доступні користувачам обох акаунтів.

Встановлення зв'язку з Ad Exchange – після зв'язування акаунтів дані будуть доступні користувачам обох акаунтів.

Налаштування аудиторії дозволяє об'єднувати групи користувачів за певними параметрами, за якими можна проводити більш глибокий аналіз відвідувачів сайту. На даний момент створення аудиторії без зв'язку з Google Ads не є можливим.

Спеціальні параметри. За допомогою спеціальних параметрів і показників ця функція збирає дані, які не відслідковуються в Google Analytics автоматично. Наприклад, обсяг двигуна.

Імпорт даних дозволяє завантажити в систему аналітики даних з зовнішніх джерел для складання більш повної картини поведінки відвідувачів.

Налаштування доступу. Аналогічно з попередніми налаштуваннями доступу, за винятком того, що тепер можна налаштувати доступ виключно до конкретного представлення.

Налаштування цілей дозволяє відстежувати дії користувачів, які максимально важливі для сайта. Наприклад, оформлення замовлення, заповнення форми, перехід на сторінку. Без цих відомостей практично

неможливо оцінити ефективність онлайн-бізнесу та проведених маркетингових кампаній.

Групи контенту дозволяє об'єднати однотипні сторінки в групу для аналізу кожної з них.

Здійснюємо угрупування:

– за допомогою спеціального коду відстеження, який необхідно розмістити на сторінках сайту;

– за допомогою угруповання за правилами відповідності. Діє набір логічних операцій І / АБО, що дозволяють додавати необхідні умови;

– вилучення параметрів з заголовків і URl-адрес сторінок. Сторінки витягуються по URL, заголовку або назві екрану.

Налаштування фільтрів дозволяє налаштувати параметри фільтрації даних, що потрапляють в звіти. Ця функція може бути корисною для відсіювання офісного трафіку, щоб не засмічувати дані при тестуванні налаштування цілей.

Налаштування групи каналів дозволяє створити власні групи каналів трафіку для зручності відстеження. Корисно, якщо йде реклама в декількох джерелах, а необхідно проаналізувати в звіті по асоціативним переходам який саме канал успішно відпрацював.

Налаштування електронної торгівлі дозволяє отримувати в звіти детальну інформацію про транзакції, вартості замовлень і багато іншого. Для збору даних з електронної торгівлі, крім активації його в Google Analytics, необхідно впровадження спеціального JavaScript-коду для збору інформації по проведених транзакціях.

Обчислювані показники дозволяють виводити в звіти додаткові дані, засновані на спеціальних формулах, в яких проводиться обчислення зі стандартних значень Google Analytics. Наприклад за допомогою спеціальної формули можна отримувати ROI відразу в звітах, але при цьому повинна бути налаштована електронна торгівля.

Сегменти – додаткова настройка сегментації дозволяє відокремити частину даних по заданих параметрах. Її використовують для аналізу тенденцій ринку, а також ефективному розподілу рекламного бюджету.

Нотатки дозволяють зробити позначку, яка буде видна на загальній діаграмі трафіку. Це може бути зручним для позначки важливих подій, які можуть позначитися на зміні трафіку.

Моделі атрибуції озволяють налаштувати розподіл цінності серед точок взаємодії в шляху відвідуваннь.

Налаштування сповіщень дозволяє налаштувати тригери, при спрацьовуванні яких буде відправлено повідомлення на пошту.

Представлені найбільш важливі тригери:

– відсутність трафіку на веб-сайт. Створюється за умовою: якщо значення трафіку з усіх джерел менше 1 – спрацьовує оповіщення, дозволяє отримати оповіщення про непрацюючий сайт;

– кількість конверсій зменшилася на 20% в порівнянні з попереднім тижнем;

– показник відмов менше 5% – дозволяє виявити дублікат Analytics ID;

– не було транзакцій (не працює в режимі онлайн) за тиждень – за умови, що електронна комерція налаштована, дозволяє вчасно відреагувати на проблему з кодом електронної торгівлі.

2.3 Інтеграція Google Analytics з веб-сайтом

Для того, щоб інтегрувати Google Analytics на сайт, потрібно першочергово перейти на сайт Google Analytics і зареєструватися. Далі створити акаунт:

– вибрати, що відстежувати: веб-сайт або мобільний додаток;

– придумати назву облікового запису, що відображає його зміст;

– ввести назву сайту;

– вказати URL сайту;

– вибрати відповідну галузь;

– вказати країну та часовий пояс (в якому велика частина аудиторії сайту) – вони будуть відображатися у звітах.

Після успішно створеного акаунта буде згенерований код відстеження Google Analytics (Рисунок 2.2)[4], який потрібно встановити на кожній сторінці веб-сайту безпосередньо після тега <head>.

```
<!-- Global Site Tag (gtag.js) - Google Analytics -->
<script async src="https://www.googletagmanager.com/gtag/js?id=GA_TRACKING_ID"></script>
<script>
  window.dataLayer = window.dataLayer || [];
  function gtag(){dataLayer.push(arguments);}
  gtag('js', new Date());

  gtag('config', 'GA_TRACKING_ID');
</script>
```

Рисунок 2.2 – Трекіновий javascript код Google Analytics

Для установки Google Analytics на сайт можна використовувати кілька простих методів:

– використовувати безпосередньо скрипт Global Site Tag;

– скористатися диспетчером тегів;

– використовувати різні плагіни.

Наступним кроком буде настройка першого представлення. В першу чергу необхідно провести такі налаштування:

– задати валюту подання, яка буде відображатися в звітах;

– увімкнути фільтрацію ботів, щоб їх візити на сайт не спотворювали статистику по трафіку;

– включити відстеження пошукових запитів на сайті, якщо потрібно знати, що шукають відвідувачі.

2.4 Налаштування відстежень конверсій

Для оцінки ефективності роботи веб-сайту і маркетинговий кампанії потрібно налаштувати відстеження конверсій в Google Analytics.

Правильна настройка відстеження конверсій в Google Analytics дозволить дізнатися точний відсоток аудиторії, яка виконує певну цільову дію. Даний показник також називається «рівень конверсії», і є найважливішим в оцінці ефективності сайту або джерел трафіку.

У Google Analytics існує пять основних видів цілей:

– цільова сторінка;
– тривалість візиту;
– кількість сторінок або екранів за сеанс;
– подія;
– розумні цілі.

«Цільова сторінка» дозволяє визначити точну кількість відвідувачів певної сторінки або послідовності сторінок / документів (активується після відвідин відвідувачем посадкової сторінки).

«Тривалість візиту» дозволяє визначити кількість часу, що проведений користувачами на сайті. В основному ця ціль використовується на сайтах «landing» і подібних односторінкових джерелах, де важливо дізнатися кількість відмов.

«Кількість сторінок або екранів за сеанс». Як умову, тут можна поставити певну кількість відкритих відвідувачем сторінок на сайті. За допомогою цієї цілі можна визначити, чи цікавий веб-сайт для користувачів.

«Подія» – одна з найцікавіших і важливих цілей, яка дозволяє відслідковувати велику кількість дій відвідувачі або певний трафік. Налаштування подій дозволить проконтролювати:

– переходи по посиланнях;
– заповнення форм;

– підписки на соцмережі;
– кліки на банери;
– перегляди відео.

2.5 Використані дані Google Analytics для аналізу

Датасет, отриманий за допомогою Google Analytics, містить такі атрибути як Client ID, Sessions, Bounce Rate, Avg. Session Duration, Revenue, Transactions, Goal Converion Rate. Кількіть сесій, дохід, транзакції та коефіцієнт конверсій відносять до кількістних метрик, а середню тривалість сесії та показник відмов до якісних:

– client ID – це унікальний ідентифікатор, який відправляється в Google Analytics разом з кожним зверненням з сайту та дозволяє зв'язати ці звернення з одним користувачем. Коли користувач заходить на сайт, бібліотека analytics.js, яка підключається при установці лічильника, відправляє в Google Analytics запит, який містить різну інформацію про факт відвідування: url відвідується сторінки, її назва, referrer, який призвів користувача, розмір вікна, мова, кодування, ідентифікатор користувача, номер лічильника, куди це все треба скласти;

– sessions – це група взаємодій користувачів із сайтом протягом певного періоду часу. Наприклад, окремий сеанс може включати кілька переглядів сторінок, подій, соціальних взаємодій та транзакцій електронної комерції. Один відвідувач може відкрити кілька сеансів. Такі сеанси можуть відбутися протягом одного або кількох днів, тижнів чи місяців. Після завершення одного сеансу можна почати новий. Якщо користувач не виконує жодних дій протягом 30 хвилин, усі подальші дії за умовчанням реєструються для нового сеансу. Якщо ж користувач залишає сайт і повертається протягом 30 хвилин, це вважається продовженням початкового сеансу;

– середня тривалість сеансу (avg. session duration) – це загальна тривалість усіх сеансів (у секундах), поділена на кількість сеансів. Щоб обчислити середню тривалість сеансу, система Analytics підсумовує тривалість кожного сеансу протягом указаного діапазону дат і ділить цю суму на загальну кількість сеансів;

– bounce rate – це відсоток сеансів, під час яких користувачі переглянули тільки одну сторінку сайту, не ініціюючи додаткових запитів до сервера Analytics. Цей показник обчислюється діленням кількості сеансів із переглядом тільки однієї сторінки на загальну кількість сеансів на сайті;

– revenue – загальна сума доходу або обороту від транзакцій;

– transactions – загальна кількість транзакцій, що проведена одним користувачем;

– goal conversion rate – це відношення числа відвідувачів, які здійснили запропоновану на цільовій сторінці дію (заповнили лід-форму, зробили покупку, підписалися на розсилку і т. п.) до загальної кількості відвідувачів, виражене у відсотках.

3 ОПРАЦЮВАННЯ ДАНИХ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

3.1 Кластерний аналіз

Кластеризація (або кластерний аналіз) – це задача розбиття множини об'єктів на групи, які називаються кластерами. Усередині кожної групи повинні виявитися «схожі» об'єкти, а об'єкти різних групи мають бути якомога більш відмінні. Головна відмінність кластеризації від класифікації полягає в тому, що перелік груп чітко не заданий та визначається в процесі роботи алгоритму.[4]

Кластеризація в Data Mining набуває цінність тоді, коли вона виступає одним з етапів аналізу даних, побудови закінченого аналітичного рішення. Аналітику часто легше виділити групи схожих об'єктів, вивчити їх особливості та побудувати для кожної групи окрему модель, ніж створювати одну загальну модель на всіх даних. Таким прийомом постійно користуються в маркетингу, виділяючи групи клієнтів, покупців, товарів і розробляючи для кожної з них окрему стратегію.

Більшість алгоритмів кластеризації припускають порівняння об'єктів між собою на основі певної міри близькості (подібності). Мірою близькості називається величина, що має межу і зростає зі збільшенням близькості об'єктів. Заходи подібності знаходяться за спеціальними правилами, а вибір конкретних заходів залежить від завдання, а також від шкали вимірювань.

Для того, щоб визначити схожість або подібність об'єктів, потрібно скласти вектор характеристик для кожного об'єкта – як правило, це набір числових значень, наприклад, ріст або вага людини. Однак існують також алгоритми, що працюють з якісними (категорійними) характеристиками.

Після того, як визначили вектор характеристик, можна провести нормалізацію, щоб всі компоненти давали однаковий внесок при розрахунку «відстані». У процесі нормалізації всі значення приводяться до деякого діапазону, наприклад, [-1, -1] або [0, 1].

Для кожної пари об'єктів вимірюється відстань між ними – ступінь схожості. Існує безліч метрик, ось лише основні з них:

Евклідова відстань – найбільш поширена функція відстані, являє собою геометричну відстань в багатовимірному просторі:

$$p(x, x') = \sqrt{\sum_i^n \left(x_i - x_i^{'}\right)^2}. \quad (3.1)$$

Квадрат евклидової відстані застосовується для додання більшої ваги більш віддаленим один від одного об'єктів. Ця відстань обчислюється таким чином:

$$p(x, x') = \sum_i^n \left(x_i - x_i^{'}\right)^2 \quad (3.2)$$

Відстань міських кварталів (Манхеттенський відстань) – це відстань є середнім різниць по координатах. У більшості випадків ця міра відстані приводить до таких же результатів, як і для звичайної відстані Евкліда. Однак при використанні цієї відстані вплив окремих великих різниць (викидів) зменшується, тому що вони не зводяться в квадрат. Формула для розрахунку манхеттенського відстані:

$$p(x, x') = \sum_i^n x_i - x_i^{'} \vee . \quad (3.3)$$

Відстань Чебишева – може виявитися корисною, коли потрібно визначити два об'єкти як «різні», якщо вони розрізняються за якоюсь однією координатою. Відстань Чебишева обчислюється за формулою:

$$p(x, x') = \max\left(\left|x_i - x_i'\right|\right). \quad (3.4)$$

Ступінна відстань застосовується, коли необхідно збільшити або зменшити вагу, що відноситься до розмірності, для якої відповідні об'єкти сильно відрізняються. Статична відстань обчислюється за такою формулою:

$$p(x, x') = \sqrt[r]{\sum_i^n \left(x_i - x_i'\right)^p}\ . \quad (3.5)$$

де r і p – параметри, що визначаються користувачем. Параметр p відповідальний за поступове зважування різниць за окремими координатами, параметр r відповідальний за прогресивне зважування великих відстаней між об'єктами. Якщо обидва параметри – r і p – дорівнюють двом, то це відстань збігається з відстанню Евкліда.

Вибір метрики повністю лежить на дослідникові, оскільки результати кластеризації можуть істотно відрізнятися при використанні різних методів.

Як об'єднувати між собою кластери? Як обчислювати «відстані» між ними?

Існує кілька метрик:

– одиночний зв'язок (відстані найближчого сусіда). У цьому методі відстань між двома кластерами визначається відстанню між двома найбільш близькими об'єктами (найближчими сусідами) в різних кластерах. Результуючі кластери мають тенденцію об'єднуватися в ланцюжки;

– повний зв'язок (відстань найбільш віддалених сусідів). У цьому методі відстані між кластерами визначаються найбільшою відстанню між будь-якими двома об'єктами в різних кластерах (тобто найбільш віддаленими сусідами). Цей метод зазвичай працює дуже добре, коли об'єкти походять з окремих груп. Якщо ж кластери мають подовжену форму або їх природний тип є «ланцюговий», то цей метод непридатний;

– незважене попарне середнє. У цьому методі відстань між двома різними кластерами обчислюється як середня відстань між усіма парами об'єктів в них. Метод ефективний, коли об'єкти формують різні групи, проте він працює однаково добре і в випадках протяжних («ланцюгового» типу) кластерів;

– виважене попарне середнє. Метод, ідентичний методу невиваженого попарного середнього, за винятком того, що при обчисленнях розмір відповідних кластерів (тобто число об'єктів, що містяться в них) використовується в якості вагового коефіцієнта. Тому даний метод повинен бути використаний, коли передбачаються нерівні розміри кластерів;

– незважений центроїдний метод.У цьому методі відстань між двома кластерами визначається як відстань між їх центрами тяжкості;

– зважений центроїдний метод (медіана). Цей метод ідентичний попередньому, за винятком того, що при обчисленнях використовуються ваги для обліку різниці між розмірами кластерів. Тому, якщо є або підозрюються значні відмінності в розмірах кластерів, цей метод виявляється більш використовуваним за попередній.

Застосування кластерного аналізу в загальному вигляді зводиться до наступних етапів:

– відбір вибірки об'єктів для кластеризації;

– визначення множини змінних, за якими будуть оцінюватися об'єкти у вибірці, при необхідності – нормалізація значень змінних;

– обчислення значень міри схожості між об'єктами;

– застосування методу кластерного аналізу для створення груп схожих об'єктів (кластерів);

– представлення результатів аналізу.

Після отримання та аналізу результатів можливе коригування обраної метрики та методу кластеризації до отримання оптимального результату.

3.2 Алгоритм k-means

Алгоритм k-means – найбільш простий, але в той же час досить неточний метод кластеризації в класичній реалізації. Він розбиває множину елементів векторного простору на заздалегідь відоме число кластерів k. Дія алгоритму така, що він прагне мінімізувати середньоквадратичне відхилення на точках кожного кластера. Основна ідея полягає в тому, що на кожній ітерації переобчислюється центр мас для кожного кластера, отриманого на попередньому кроці, потім вектори розбиваються на кластери знову відповідно до того, який з нових центрів виявився ближчим за обраною метрикою. Алгоритм завершується, коли на якийсь ітерації не відбувається зміни кластерів.[5]

В якості міри близькості використовується Евклідова відстань:

$$\mathrm{p}(\mathrm{x},\mathrm{y}) = \|\mathrm{x}-\mathrm{y}\| = \sqrt{\sum_{\mathrm{p}=1}^{\mathrm{n}} \left(\mathrm{x}_{\mathrm{i}} - \mathrm{x}_{\mathrm{i}}\right)^2}, \qquad (3.6)$$

де $\mathrm{x},\mathrm{y} \in \mathrm{R}^{\mathrm{n}}$.

Отже, розглянемо ряд спостережень, $(x^1, x^2, \dots, x^m), j \in R^n$. Метод k-середніх розділяє m спостережень на k груп (або кластерів) $(k \leq m)$ $S = \{S_1, S_2, \dots, S_k\}$, щоб мінімізувати сумарне квадратичне відхилення точок кластерів від центроїдів цих кластерів:

$$\min\left[\sum_{i=1}^{k}\sum_{x^{(s)}\in S_i} \vee x^{(j)} - \mu_i \vee^2\right], \qquad (3.7)$$

де $x^{(j)} \in R^n, \mu_i \in R^n$, μ_i – центроїд для кластера S_i.

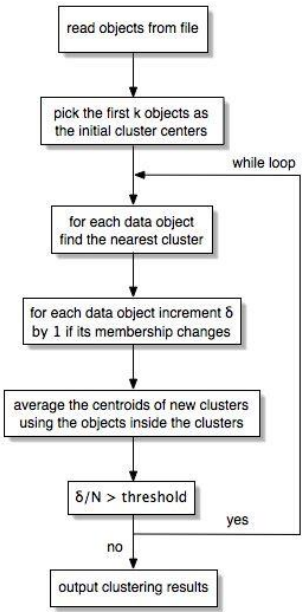
Розглянемо початковий набір k середніх (центроїд) $\mu_1, \ldots, \mu_k$ в кластерах $S_1, S_2, \ldots, S_k$.

На першому етапі центроїди кластерів вибираються випадково або за певним правилом (наприклад, вибрати центроїди для максимізації початкових відстанів між кластерами). Відносимо спостереження до тих кластерів, чиє середнє (центр ваги) до них найближче. Кожне спостереження належить тільки до одного кластеру, навіть якщо його можна віднести до двох і більше кластерів. Потім центр ваги кожного і-го кластера переобчислюють за таким правилом:

$$\mu_j = \frac{1}{s_j}\sum_{x^{(j)}\in S_i} x^{(j)}. \qquad (3.8)$$

Таким чином, алгоритм *k*-середніх полягає в перерозрахунку на кожному кроці центроїда для кожного кластера, отриманого на попередньому кроці. (Рисунок 3.1)

Алгоритм зупиняється, коли значення μ_j не змінюються $\mu_i^{\text{крок}t} = \mu_i^{\text{крок}t+1}$.




```
N:  number of data objects
K:  number of clusters

objects[N]:  array of data objects
clusters[K]:  array of cluster centers
membership[N]:  array of object memberships


kmeans_clustering( )
1   while  δ/N > threshold
2      δ ← 0
3      for i ← 0 to N-1
4         for j ← 0 to K-1
5            distance ← | objects[i] - clusters[j] |
6            if distance < dmin
7               dmin ← distance
8               n ← j
9         if membership[i] ≠ n
10           δ ← δ + 1
11           membership[i] ← n
12        new_clusters[n] ← new_clusters[n] + objects[i]
13        new_cluster_size[n] ← new_cluster_size[n] + 1
14     for j ← 0 to K-1
15        clusters[j][*] ← new_clusters[j][*] / new_cluster_size[j]
16        new_clusters[j][*] ← 0
17        new_cluster_size[j] ← 0
```

Рисунок 3.1 – Реалізація алгоритму k-means

Перевагою алгоритму є швидкість і простота реалізації.

Проблеми алгоритму *k*-means:

– необхідно заздалегідь знати кількість кластерів. Можна запропонувати метод визначення кількості кластерів, який грунтувався на знаходженні кластерів, розподілених по якомусь закону (в даному випадку все зводилося до нормального закону). Після цього виконувався класичний алгоритм k-means, який давав більш точні результати;

– алгоритм дуже чутливий до вибору початкових центрів кластерів. Класичний варіант працює за принципом випадкового вибору кластерів, що дуже часто є джерелом похибки. Як варіант вирішення необхідно проводити дослідження об'єкта для більш точного визначення центрів початкових кластерів;

– алгоритм не справляється із завданням, коли об'єкт належить до різних кластерів в рівній мірі або не належить жодному.[5]

3.3 Метод найближчих сусідів (*kNN-k* Nearest Neighbours)

Метод найближчих сусідів (*kNN-k* Nearest Neighbours) – метод вирішення завдань класифікації та завдань регресії, заснований на пошуку найближчих об'єктів з відомими значення цільової змінної.

Метод заснований на припущенні про те, що близьким об'єктам в просторі ознак відповідають схожі мітки.

Для нового об'єкта x метод передбачає знайти найближчі до нього об'єкти $x_1, x_2, \ldots, x_k$ і побудувати прогноз по їх міткам.

Для класифікації кожного з об'єктів тестової вибірки необхідно послідовно виконати наступні операції:

– обчислити відстань до кожного з об'єктів навчальної вибірки;

– відібрати k об'єктів навчальної вибірки, відстань до яких мінімальна;

– клас, що класифікує об'єкт – це клас, який найчастіше визначається серед *k* найближчих сусідів.

Складність навчання – O (1). Технічно правильною відповіддю також є O (n), так як потрібно запам'ятовувати навчальну вибірку.

Складність прогнозування – O(n) для кожного об'єкта. Якщо по фіксованій навчальній вибірці потрібно незалежно зробити прогноз для k об'єктів, складність O (kn).

Перевагою алгоритму є швидкість і простота реалізації і не потрібно переднавчання. До недоліків можна віднести високу складність одного прогнозу.

3.4 Регресійний аналіз

Якщо є кореляційна залежність *F (y) = F (x, y)* між змінними *y* та *x*, виникає необхідність визначити функціональний зв'язок між двома величинами. Залежність середнього значення $\mu_1 y = f(x)$ називається регресією *y* по *x*.

Основу регресійного аналізу становить метод найменших квадратів (МНК), відповідно до якого в якості рівняння регресії береться функція *y=f(x)* така, що сума квадратів різниць є мінімальною (3.9):

$$s = \sum_{i=1}^{n} \left[y_1 - f\left(x\right)_i \right]^2 \qquad (3.9)$$

Даний статистичний метод дослідження широко використовується для прогнозування, де його використання має істотну перевагую. Але іноді це може призводити до ілюзії чи хибним висновкам, тому рекомендується акуратно його використовувати, оскільки, наприклад, кореляція це не є причинно-наслідковий зв'язок. Відома велика кількість методів для проведення регресійного аналізу, такі як лінійна та звичайна регресії за методом найменших квадратів, які є параметричними. Їх суть в тому, що функція регресії визначається в термінах кінцевого числа невідомих параметрів, які оцінюються з даних. Непараметрична регресія дозволяє її функції лежати в певному наборі функцій, які можуть бути нескінченновимірними. Як статистичний метод дослідження, регресійний аналіз на практиці залежить від форми процесу генерації даних і від того, як він ставиться до регрессионному підходу. Так як справжня форма процесу даних, що генерують, як правило, невідоме число, регресійний аналіз даних часто залежить в деякій мірі від припущень про цей процес. Ці припущення іноді перевіряються, якщо є достатня кількість доступних

даних. Регресивні моделі часто бувають корисні навіть тоді, коли припущення помірно порушені, хоча вони не можуть працювати з максимальною ефективністю.

3.5 Лінійний регресійний аналіз

У лінійної регресії особливістю є те, що залежна змінна, якою є Y_i, являє собою лінійну комбінацію параметрів. Наприклад, в простій лінійній регресії для моделювання n точок використовується одна незалежна змінна, x_1, і два параметра, β_0 i β_1.

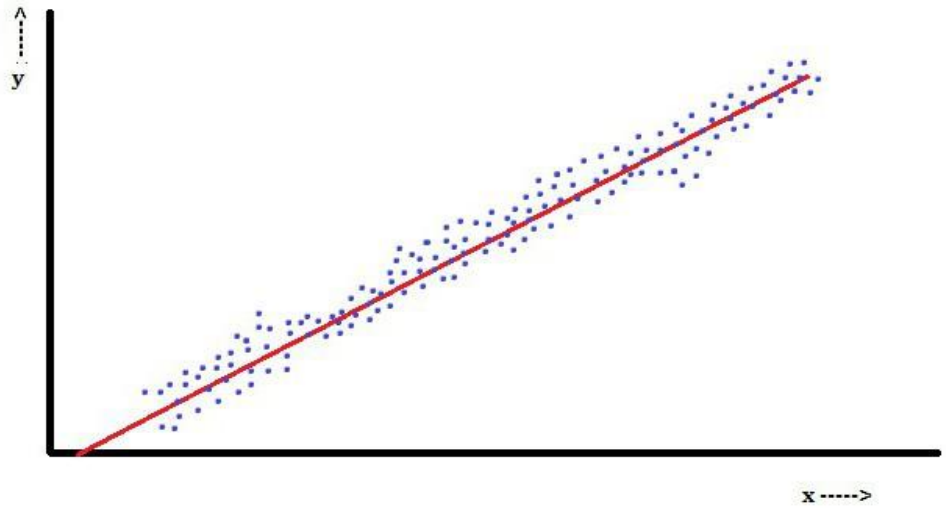


Рисунок 3.2 – Графік лінійної регресії

При множиній лінійній регресії існує кілька незалежних змінних або їх функцій. При випадковій вибірці з популяції її параметри дозволяють отримати зразок моделі лінійної регресії (Рисунок 3.2).В даному аспекті найпопулярнішим є метод найменших квадратів. За допомогою нього отримують оцінки параметрів, які мінімізують суму квадратів залишків. Такого роду мінімізація (що характерно саме лінійної регресії) цієї функції призводить до набору нормальних рівнянь і набору лінійних рівнянь з параметрами, які вирішуються з отриманням оцінок параметрів. При

подальшому припущенні, що помилка популяції зазвичай поширюється, дослідник може використовувати ці оцінки стандартних помилок для створення довірчих інтервалів і проведення перевірки гіпотез про її параметри.

Для того, щоб використовувати модель лінійної регресії, необхідні деякі припущення щодо розподілу і властивостей змінних:

– лінійність. Збільшення, або зменшення вектора незалежних змінних в *k* разів, призводить до зміни залежної змінної також в *k* разів;

– матриця коефіцієнтів володіє повним рангом, тобто вектори незалежних змінних лінійно незалежні;

– екзогенність незалежних змінних:

$$E\left[\epsilon_i \left| x_{j1},\ x_{j2},\ldots x_{jk} \right|\right] = 0. \qquad (3.10)$$

Ця вимога означає, що математичне очікування похибки жодним чином не можна пояснити за допомогою незалежних змінних;

– однорідність дисперсії та відсутність автокореляції. Кожна *εi* має однакову і кінцевої дисперсії *σ2* і не корелює з іншою *εi*. Це відчутно обмежує застосування моделі лінійної регресії, необхідно упевнитися в тому, що умови дотримані, інакше виявлений взаємозв'язок змінних буде невірно інтерпретований.

3.6 Алгоритм t-SNE

Алгоритм t-SNE, який також відносять до методів множинного навчання ознак, був опублікований в 2008 році голландським дослідником Лоуренсом ван дер Маатеном і спеціалістом в галузі нейронних мереж Джеффрі Хінтоном. Класичний SNE був запропонований Хінтоном і Ровейсом в 2002 році. У статті 2008 року описується кілька механізмів, які

дозволили спростити процес пошуку глобальних мінімумів і підвищити якість візуалізації. Одним з них стала заміна нормального розподілу на розподіл Стьюдента для даних низької розмірності. Крім того, була зроблена вдала реалізація алгоритму, яка потім впроваджувалась в інші популярні середовища.

Почнемо з «класичного» SNE і сформулюємо завдання. У нас є набір даних з точками, описаними багатовимірними змінними з розмірністю простору істотно більше трьох. Необхідно отримати нову змінну, яка існує в двовимірному або тривимірному просторі, яка б в максимальному ступені зберігала структуру та закономірності в вихідних даних. SNE починається з перетворення багатовимірної евклідової дистанції між точками в умовні ймовірності, що відображають схожість точок. Математично це виглядає наступним чином:

$$p_{j\vee i} = \frac{\exp\left(-\vee x_i - x_j \vee^2 / 2\sigma_i^2\right)}{\sum_{k\neq}\exp\left(-\vee x_i - x_k \vee^2 / 2\sigma_i^2\right)}. \qquad (3.11)$$

Ця формула показує, наскільки точка *Xj* близька до точки *Xi* при гауссовому розподілі навколо *Xi* з заданим відхиленням σ. Сігма буде різною для кожної точки. Вона вибирається так, щоб точки в областях з більшою щільністю мали меншу дисперсію. Для цього використовується оцінка перплексії:

$$Perp\left(P_i\right) = 2^{H(P_i)}, \qquad (3.12)$$

де $H(P_i)$) – ентропія Шеннона в бітах.

$$H\left(P_i\right) = -\sum_j p_{j\vee i}\log_2 p_{j\vee i}. \qquad (3.13)$$

В даному випадку перплексія може бути інтерпретована як згладжена оцінка ефективної кількості «сусідів» для точки *Xi*. Вона задається як параметр методу. Рекомендовано використовувати значення в інтервалі від 5 до 50. Сігма визначається для кожної пари *Xi* і *Xj* за допомогою алгоритму бінарного пошуку.

Для двовимірних або тривимірних випадків пари *Xi* і *Xj*, назвемо їх для ясності *Yi* і *Yj*, не становить труднощів оцінити умовну ймовірність. Стандартне відхилення пропонується встановити в *1 / √2*:

$$q_{j||i} = \frac{exp(-\left\| y_i - y_j \right\|^2}{\sum_{k\neq} exp\left(-\left\| y_i - y_k \right\|^2\right)}. \tag{3.14}$$

Якщо точки відображення Y_i і Y_j коректно моделюють схожість між вихідними точками високої розмірності X_i і X_j, то відповідні умовні ймовірності $p_{j\vee i}$ і $q_{j\vee i}$ будуть еквівалентні. Як очевидною оцінки якості, з яким $q_{j\vee i}$ відображає $p_{j\vee i}$, використовується дивергенція або відстань Кульбака-Лейблера. SNE мінімізує суму таких відстаней для всіх точок відображення за допомогою градієнтного спуску. Функція втрат для даного методу буде визначатися формулою:

$$\text{Cost} = \sum_i \text{KL}(\text{P}_i \,||\text{Qi}) = \sum_i \sum_j \text{p}_{j|i} \log \frac{\text{p}_{j|i}}{\text{q}_{j|i}} \tag{3.15}$$

При цьому градієнт виглядає:

$$\frac{\partial \text{Cost}}{\partial \text{y}_i} = 2\sum_j \left(\text{p}_{j\vee i} - \text{q}_{j\vee i} + \text{p}_{i\vee j} - \text{q}_{i\vee j}\right)\left(\text{y}_i - \text{y}_j\right). \tag{3.16}$$

Всі точки відображення з'єднані пружинами. Жорсткість пружини, що з'єднує точки *i* і *j*, залежить від різниці між схожістю двох точок в

багатовимірному просторі і двох точок у просторі відображення. У цій аналогії, градієнт – це результуюча сила, що діє на точку в просторі відображення. Якщо систему «відпустити», через якийсь час вона прийде в рівновагу, це і буде шуканий розподіл. Алгоритмічно, пошук рівноваги пропонується робити з урахуванням моментів:

$$Y^{t} = Y^{(t-1)} + n\frac{\partial Cost}{\partial Y} + \alpha(t)\left(Y^{(t-1)} - Y^{(t-2)}\right). \quad (3.17)$$

де η – параметр, що визначає швидкість навчання (довжину кроку);

α – коефіцієнт інерції.

Використання класичного SNE дозволяє отримати хороші результати, але можуть бути труднощі в оптимізації функції втрат і проблемою скупченості.

t-SNE якщо і не вирішує ці проблеми зовсім, то значно полегшує. Функція втрат t-SNE має два принципові різниці. По-перше, у t-SNE симетрична форма подібності в багатовимірному просторі і простіший варіант градієнта. По-друге, замість гауссовського розподілення для точок з простору відображення використовується t-розподіл (Стьюдента), що полегшує оптимізацію і вирішує проблему скупченості.

В якості альтернативи мінімізації суми дивергенції Кульбака-Лейблера між умовними ймовірностями $p_{i\vee j}$ і $q_{i\vee j}$ пропонується мінімізувати одиночну дивергенцію між спільною ймовірністю P в багатовимірному просторі і спільною ймовірністю Q в просторі відображення:

$$Cost = KL(P||Q) = \sum_{i}\sum_{j} p_{ij} \log\frac{p_{ij}}{q_{ij}}, \quad (3.18)$$

де p_{ij}, $q_{ij} = 0$, $p_{ij} = p_{ji}$, $q_{ij} = q_{ji}$ для будь-яких i та j, а p_{ij} визначається за формулою:

$$p_{ij} = \frac{p_{j \vee i} + p_{i \vee j}}{2n}, \tag{3.19}$$

де n – кількість точок в наборі даних.

Градієнт для симетричного SNE виходить значно простіше, ніж для класичного:

$$\frac{\partial \text{Cost}}{\partial y_i} = 4 \sum_j \left(p_{ij} - q_{ij}\right)\left(y_i - y_j\right) \tag{3.20}$$

Проблема скупченості полягає в тому, що відстань між двома точками в просторі відображення, відповідними двом середньовіддаленим точкам в багатовимірному просторі, має бути істотно більше, ніж відстань, яке дозволяє отримати гауссовий розподіл. Проблему вирішують хвости Стьюдента. В t-SNE використовується t-розподіл з одним ступенем свободи. Спільна ймовірність для простору відображення в цьому випадку буде визначатися формулою:

$$q_{ij} = \frac{\left(1 + \vee y_i - y_j \vee^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \vee y_k - y_t \vee^2\right)^{-1}}. \tag{3.21}$$

А відповідний градієнт:

$$\frac{\partial \text{Cost}}{\partial y_i} = 4 \sum_j \left(p_{ij} - q_{ij}\right)\left(y_i - y_j\right)\left(1 + \vee y_i - y_j \vee^2\right)^{-1}. \tag{3.22}$$

Повертаючись до фізичної аналогії, результуюча сила буде істотно стягувати точки простору відображення для довколишніх точок багатовимірного простору, і відштовхувати – для віддалених.

Алгоритм t-SNE (Рисунок 3.2) в спрощеному вигляді можна представити таким псевдокодом.

```
Data: набір даних x = {x1, x2, …, xn},
параметр функції втрат: перплексія Perp,
Параметри оптимізації: кількість ітерацій T, швидкість навчання n, момент α(t).
Result: представлення даних Y(T) = {y1, y2, …, yn} (в 2D або 3D).
begin
    порахувати попарну схожість pj | i с перплексією Perp
    встановити pij = (pj|i + pi|j)/2n
    ініціалізувати Y(0) = {y1, y2, …, yn} точками нормального розподілення (mean=0,
    for t = 1 to T do
        порахувати схожість точок в просторі відображення qij
        порахувати градієнт δCost/δy
        встановити Y(t) = Y(t-1) + nδCost/δy + α(t)(Y(t-1) - Y(t-2))
    end
end
```

Рисунок 3.2 – Реалізація алгоритму t-SNE на псевдокоді

Щоб поліпшити результат, пропонується використовувати два механізми. Перший називається «ранньої компресією». Його завдання – змусити точки в просторі відображення на початку оптимізації бути якомога ближче один до одного. Коли дистанція між точками відображення невелика, переміщати один кластер через інший істотно легше. Так набагато простіше досліджувати простір оптимізації та «націлитися» на глобальні мінімуми. Рання компресія створюється за рахунок додаткового L2-штрафу в функції втрат, який пропорційний сумі квадратів дистанцій точок відображення від початку координат.

Другий механізм менш очевидний – «раннє гіперусилення». Полягає він в множенні на початку оптимізації всіх pij на деяке ціле число, наприклад на 4. Сенс в тому, щоб для великих p_{ij} отримати більші q_{ij}. Це

дозволить для кластерів у вихідних даних отримати щільні та широко рознесені кластери в просторі відображення.

4 ПРОГРАМНА РЕАЛІЗАЦІЯ ТА АНАЛІЗ РЕЗУЛЬТАТІВ ПРОГНОЗУВАННЯ ВІДВІДУВАННЬ

4.1 Опис досліджувальної системи

Для збору даних про користувачів була використана веб-система інтернет-магазину ювелірних виробів ручної роботи, яка представлена на рисунках 4.1 та 4.2. Бізнес базується в Лос-Анджелесі, штат Каліфорнія, Сполучені Штати Америки. Система побудована на платформі електронною комерції Shopify і працює за принципом B2C (Бізнес-споживач). Товар надходить від підприємства до фізичної особи шляхом роздрібної торгівлі. Shopify – це провідна платформа в електронній комерції, що дозволяє підприємцям створювати власні інтернет-магазини. Веб-система включає в себе 1, 700 посадочних сторінок та 1048 товарів.

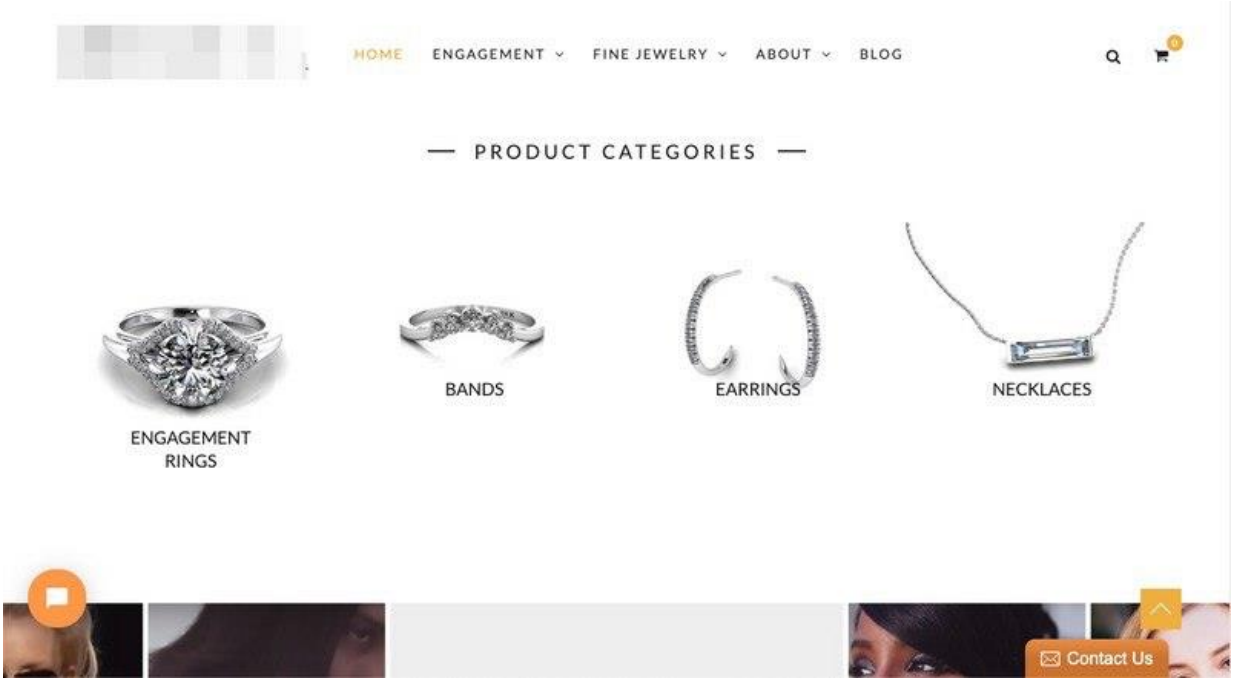


Рисунок 4.1 – Домашня сторінка досліджуваної системи (десктоп версія)

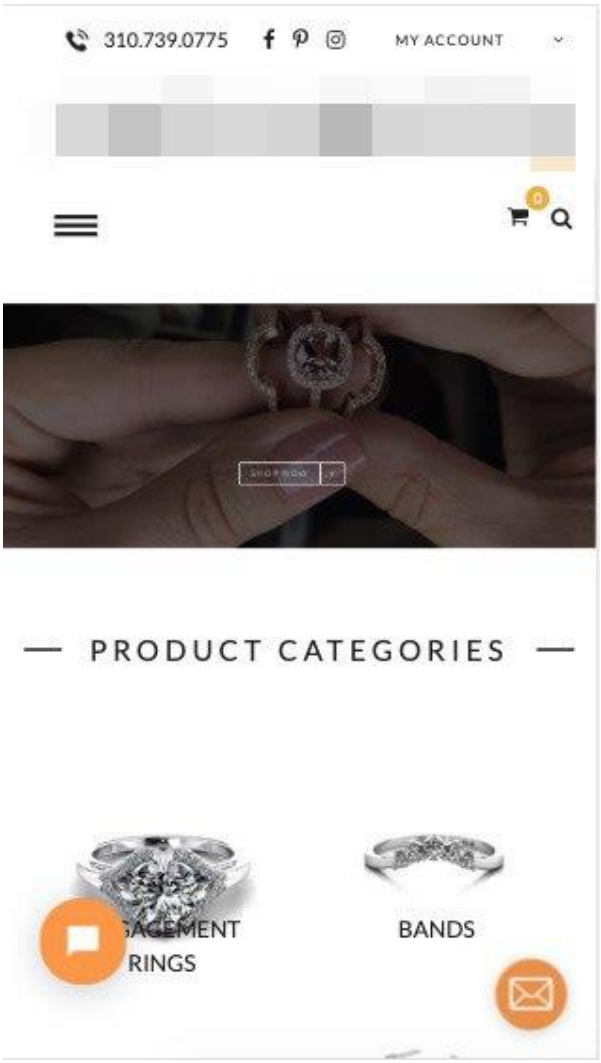


Рисунок 4.2 – Домашня сторінка досліджуваної системи (мобільна версія)

4.1.1 Технічні характеристики веб-системи

Швидкість завантаження сторінок. Люди не знають точного часу завантаження, але можуть оцінити швидкість завантаження сайту за своїми відчуттями. В середньому близько 83% відвідувачів залишають веб-сайт, якщо він завантажується більше 3 секунд. Було порівняно результати Google Analytics (рисунок 4.3) та незалежного онлайн-додатку Pingdom (рисунок 4.4).

Рисунок 4.3 – Швидкість завантаження сторінок (Google Analytics)

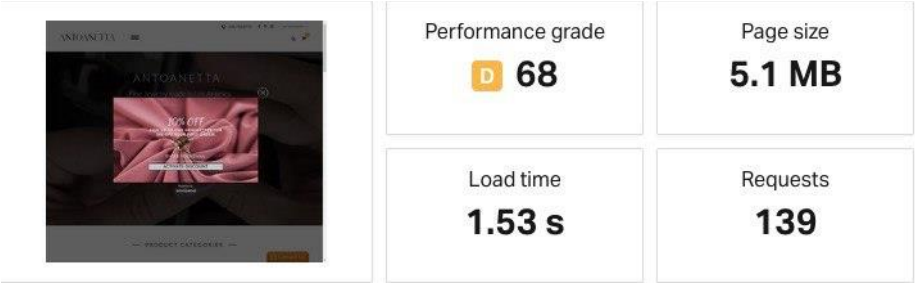


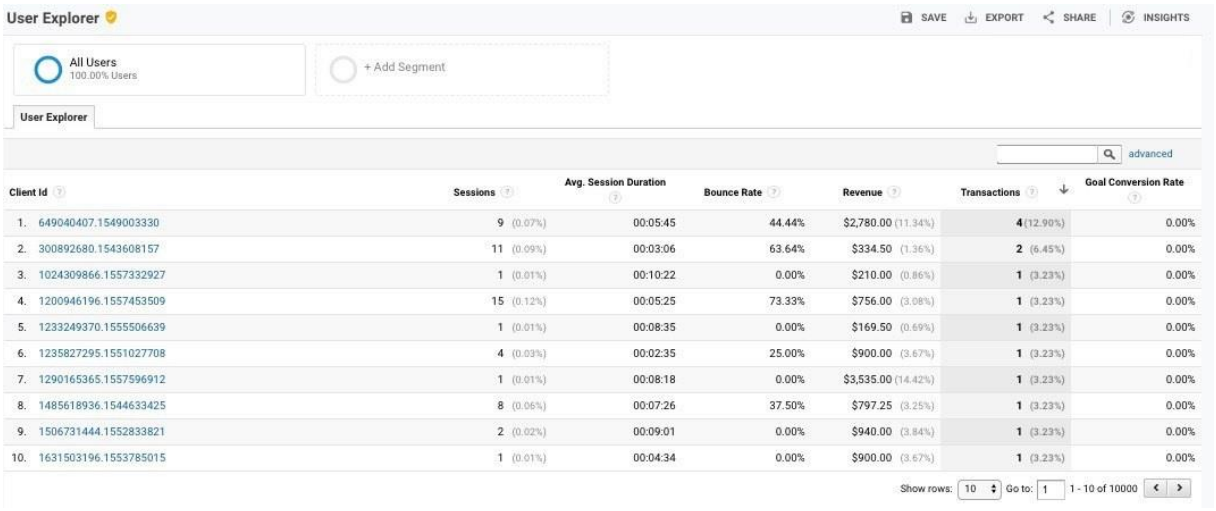
Рисунок 4.4 – Швидкість завантаження сторінок (Pingdom)

Як проілюстровано на рисунку 4.5[4], середній трафік за місяць становить 1, 164 користувачів. Середній час сесії – 00:01:42, а показник відмов приблизно 68%. Середнєстатистичний користучав відвідує 4 сторінки за одну сесію та середнє число проглянутих сесій, завершених одним користувачем, становить 1.27.

Рисунок 4.5 –Характеристики сайту (Google Analytics)

4.2 Збір даних

На веб-сайті заздалегідь було встановлено код відстеження Google Analytics і налаштована вся конфігурація відповідно до розділу 2.2. Датасета був отриманий за допомогою спеціального звіту User Exploler (рисунок 4.6)[21]. Дані, які описують поведінку користувачів, були зібрані за останні пів року.



Client Id	Sessions	Avg. Session Duration	Bounce Rate	Revenue	Transactions	Goal Conversion Rate
1. 649040407.1549003330	9 (0.07%)	00:05:45	44.44%	$2,780.00 (11.34%)	4 (12.90%)	0.00%
2. 300892680.1543608157	11 (0.09%)	00:03:06	63.64%	$334.50 (1.36%)	2 (6.45%)	0.00%
3. 1024309866.1557332927	1 (0.01%)	00:10:22	0.00%	$210.00 (0.86%)	1 (3.23%)	0.00%
4. 1200946196.1557453509	15 (0.12%)	00:05:25	73.33%	$756.00 (3.08%)	1 (3.23%)	0.00%
5. 1233249370.1555506639	1 (0.01%)	00:08:35	0.00%	$169.50 (0.69%)	1 (3.23%)	0.00%
6. 1235827295.1551027708	4 (0.03%)	00:02:35	25.00%	$900.00 (3.67%)	1 (3.23%)	0.00%
7. 1290165365.1557596912	1 (0.01%)	00:08:18	0.00%	$3,535.00 (14.42%)	1 (3.23%)	0.00%
8. 1485618936.1544633425	8 (0.06%)	00:07:26	37.50%	$797.25 (3.25%)	1 (3.23%)	0.00%
9. 1506731444.1552833821	2 (0.02%)	00:09:01	0.00%	$940.00 (3.84%)	1 (3.23%)	0.00%
10. 1631503196.1553785015	1 (0.01%)	00:04:34	0.00%	$900.00 (3.67%)	1 (3.23%)	0.00%

Show rows: 10 Go to: 1 1 - 10 of 10000

Рисунок 4.6 – Приклад звіту User Exploler

Відстеження конверсій було налаштоване за допомогою Enhanced Ecommerce, яке ілюструється на рисунку 4.7. [21] Enhanced Ecommerce – розширений модуль електронної комерції, який дозволяє отримати безліч

додаткової інформації про дії користувача, взаємодії з товарами, його шляху до завершення процесу покупки.

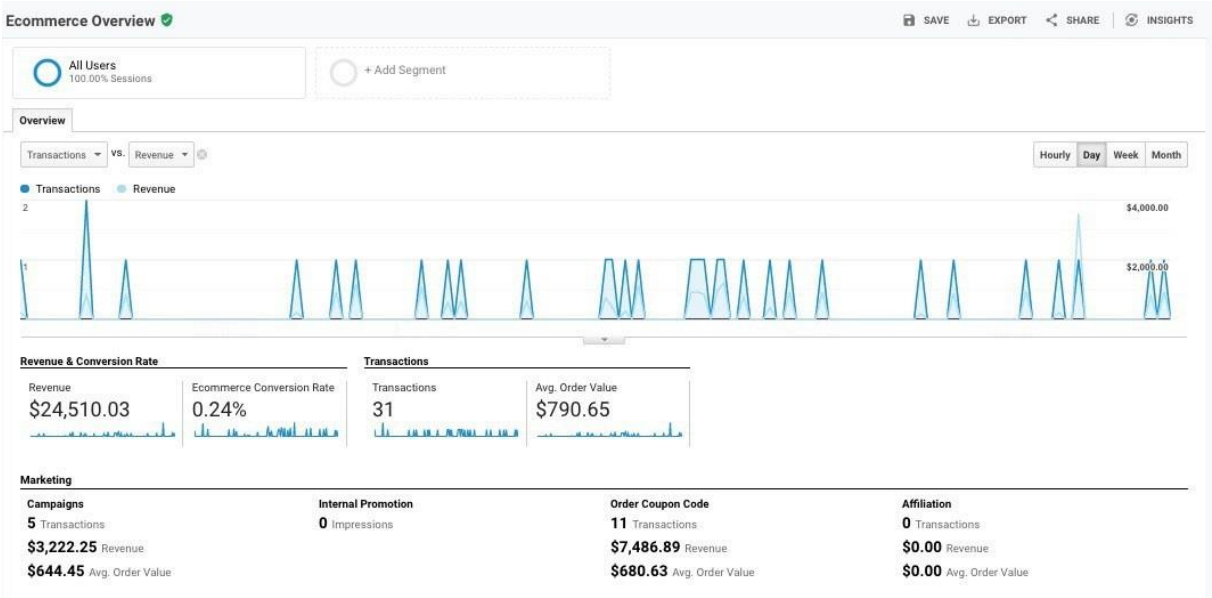


Рисунок 4.7 – Приклад звіту Enhanced Ecommerce

4.3 Аналіз результатів

Отриманий датасет був кластеризований по поведінці користувачів для визначення груп. Після того було проаналізовано дані за допомогою алгоритму t-SNE для проєктування багатовимірних даних на двовимірні для аналізу структури даних, а останнім кроком було побудування регресійної моделі, що дозволяє прогнозувати рівень доходу від користувача на основі його поведінки на сайті.

4.1.2 Результат роботи алгоритму t-SNE

Як результат відпрацьованого алгоритму t-SNE (рисунок 4.8), було спроєктовано пятивимірний датасет у двовимірний.

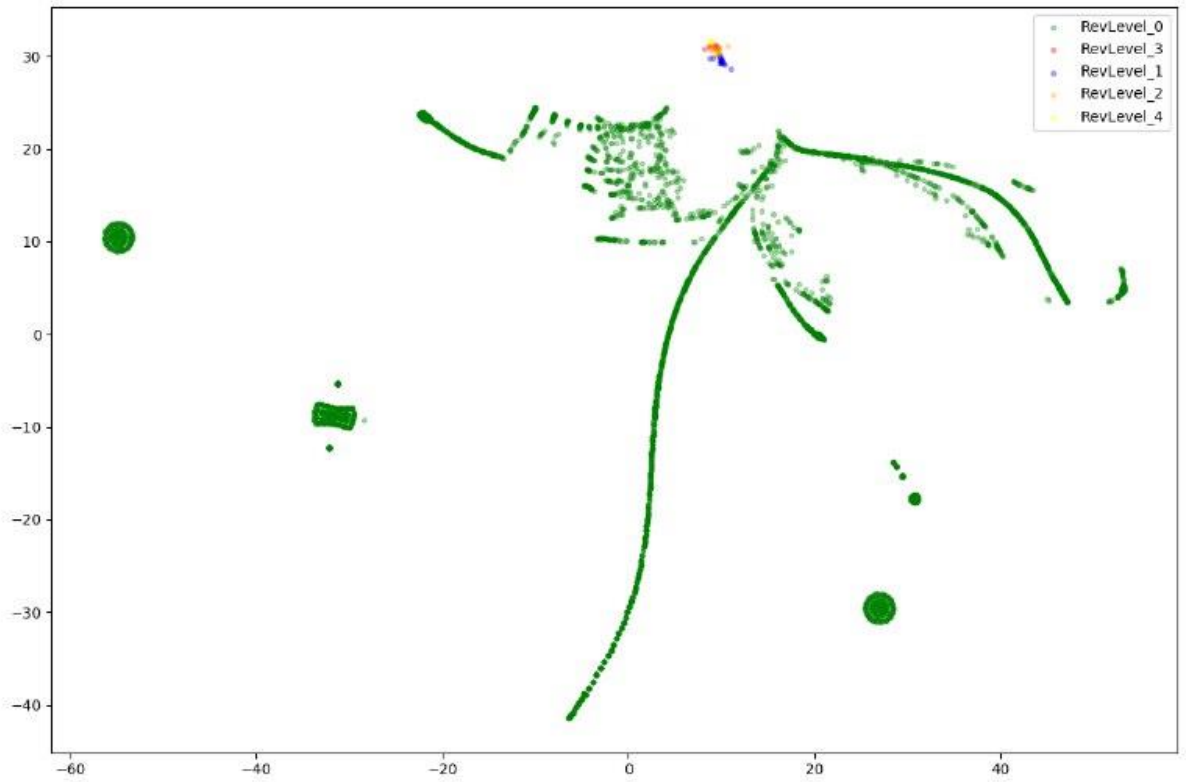


Рисунок 4.8 – Результат алгоритму t-SNE

Основна мета – зрозуміти, чи є якісь групи серед користувачів чи ні. В результаті було розподілено користувачів за витраченними коштами (revenue) на веб-сайті:

– RevLevel 0 --- Revenue = $0;
– RevLevel 1 --- Revenue < $500;
– RevLevel 2 --- Revenue > $500 та < $1000;
– RevLevel 3 --- Revenue > $1000 та < $1500;
– RevLevel 4 --- Revenue > $1500.

4.1.3 Результат роботи алгоритму кластеризації

Після того, як було спроєктовано датасет на двовимірний простір, стояла задача розбиття користувачів сайту на групи людей, схожих за своєю поведінкою. Поведінку користувача можна описати поведінковими

метриками відповідно до розділу 2.5. Для кластеризації було вибрано метод k-means, який мінімізує середньоквадратичне відхилення на точках кожного кластера. Основна ідея полягає в тому, що на кожній ітерації переобчислюють центр мас для кожного кластера, отриманого на попередньому кроці, потім вектори розбиваються на кластери знову відповідно до того, який з нових центрів виявився ближчим за обраною метрикою. Алгоритм завершується, коли на якийсь ітерації не відбувається зміни кластерів. Основною перевагою алгоритму є простота реалізації та швидкість роботи.

Як результат правильно відпрацьованого алгоритму було отримано 8 кластерів користувачів веб-сайту (рисунок 4.9).

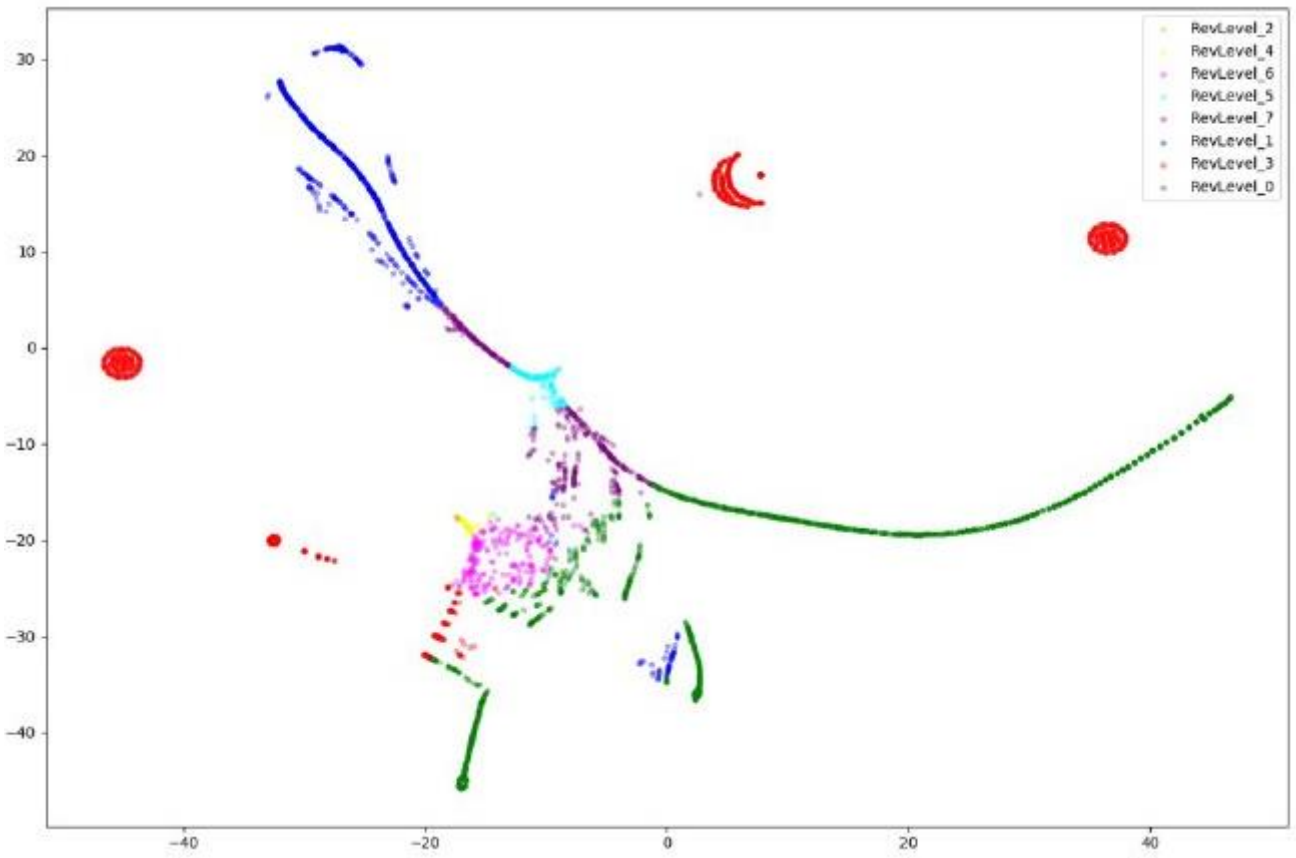


Рисунок 4.9 – Результат алгоритму k-means

4.1.4 Прогнозуюча модель

Після того, як було розбито користувачів на групи, була побудована прогнозуюча модель на основі лінійної регресії. (Рисунок 4.10)

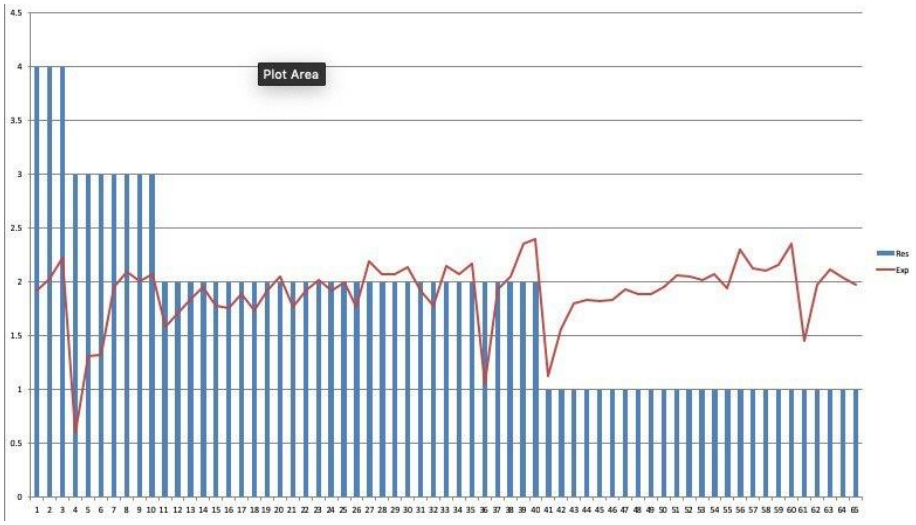


Рисунок 4.10 – Результат алгоритму множинної регресії

В результаті була отримана модель, яка проілюстрована на рисунку 4.11:

```
RevLev = 2.04 - 0.11 * [Session] - 0.11 * [Avg.Session Duration] - 0.181 * [Bounce Rate]
```

Рисунок 4.11 – Прогнозуюча модель

Завдяки одержаній моделі можна спрогнозувати, який дохід отримає веб-система на основі поведінкових метрик користувача.

ВИСНОВКИ

У рамках каліфікаційної роботи була спроєктована та реалізована модель прогнозування відвідуваннь на основі поведінкових метрик користувачів веб-системи. В ході дослідження проблематики був розглянутий такий список питань:

– способи збору даних про поведінку користувачів;
– робота с даними Google Analytics;
– розробка алгоритму опрацювання даних;
– реалізація алгоритму t-SNE;
– реалізація алгоритму *k*-means;
– вирішення задачі прогнозування конверсії на основі поведінки користувачів.

Основна задача полягає в тому, щоб спрогнозувати дохід від користувачів на основі їх поведінки на сайті. Дані такого характеру дають змогу оптимізувати веб-систему під ту групу користувачів, які найвірогідніше куплять товар на веб-сайті.

Після даних досліджень можна сказати, що розроблений алгоритм роботи з даними є доцільний в питанні опрацювання даних Google Analytics. Перевагами методу є швидкість та простота реалізації; на основі отриманної моделі досить легко оцінити наскільки якісний, з точки зору конверсіонності, є користувач. Недолік методу – треба працювати з датасетом, який містить багато даних про історії відвідуваннь.

Також була проведена порівняльна характеристика реалізованого методу кластеризації. Виявлено, що чим більша кількість кластерів, тим більш точніший результат можна отримати при роботі з моделлю.

Для реалізації програми на мові Python була використана відкрита програмна бібліотека для обробки та аналізу даних Pandas та відкрита бібліотека для машинного навчання Scikit-learn. Бібліотеки можуть зчитувати дані у форматах CSV, Excel, HDF, SQL, JSON, HTML і форматі

Stata, що дозволяє зручно працювати з різними типами даних. Бібліотеки розповсюджується за ліцензією BSD, це означає, що їх можна вільно та безкоштовно використовувати як у відкритих проєктах з відкритим кодом, так і в закритих, комерційних проєктах.

У якості подальшого напрямку розширення та дослідження є можливість вивчення інших методів кластеризації та прогнозування для того, щоб розробити більш універсальний алгоритм, який буде відпрацьовувати для будь-якого датасету. А також розробити інтерфейс застосунку для більш зручної роботи.

ПЕРЕЛІК ПОСИЛАНЬ

1. Руденко Д. О., Колосок Е. В. «ОГЛЯД МОЖЛИВОСТЕЙ GOOGLE ANALYTICS ДЛЯ РОБОТИ З ДАНИМИ». — 2021

2. Cameron Davidson-Pilon. Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference [Text] // Addison-Wesley Data – 2015.

3. Dr. Anasse Bari, Mohamed Chaouchi, Tommy Jung. Predictive Analytics For Dummies [Text] // For Dummies. – 2016. – Т. 40. – №. 2. – С. 5.

4. Google Analytics. Materials and tutorials [Електронний ресурс]. – Режим доступу: www/ URL: https://support.google.com/analytics/. — 2021 (Дата звернення: 8.10.2021)

5. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning, 2nd edition [Text]. — Springer, 2009.

6. John K. Kruschke. Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan [Text] // Academic Press / Elsevier — 2015.

7. Ray Solomonoff «An Inductive Inference Machine» A privately circulated report from the 1956 Dartmouth Summer Research Conference on AI [Text]. — 2009.

8. Ray Solomonoff «An Inductive Inference Machine» A privately circulated report from the 1956 Dartmouth Summer Research Conference on AI [Text]. — 2009.

9. Teaching Machine Learning from Examples — видеолекция и презентация Isabelle Guyon, Clopinet. [Електронний ресурс]. – Режим доступу: www/ URL: http://videolectures.net/learning06_guyon_tmle/.

10. Баглаева, Е. А. Многодисциплинарная оптимизация, анализ данных и автоматизация инженерных расчетов с помощью программного комплекса pSeven [Текст] // CAD/CAM/CAE Observer #4 (88). — 2014.

11. Барковський В. В., Барковська И. В., Лопатхн О. К. Математика для економістів. Теорія ймовірностей та математична статистика [Текст] – 1987. – С. 726-740.

12. Гмурман В. Е. Руководство к решению задач по теории вероятностей и математнческой статистике [Текст]. — М.: Вьісш. шк., 1999.

13. Гмурман В. Е. Теория вероятностей н математическая статистика [Текст]. – 2006. – С. 262 – 271.

14. Горбань С. Ф., Снижко [Текст] //. В. Теория вероятностей и математическая статистика. – 1999. – С. 162 – 271.

15. Ефимов А.С. Решение задачи кластеризации методом конкурентного обучения при неполных статистических данных [Текст] // Вестник Нижегородского университета им. Н.И. Лобачевского. – 2010. – №1. – С. 220 – 225.

16. Жлуктенко В. /., Наконечний С. /. Практикум з математичної статистики [Текст]. — 1991.

17. Жчуктенко В. /., Наконечний С. І. Теорія ймовірностей із елементами математичної статистики [Текст]. — 1991. — С. 404-417.

18. Крємер Н. Ш. Теория вероятностей и математическая статистика.- [Текст]. — 2000.

19. Литтл Р.Дж.А., Рубин Д.Б. Статистический анализ данных с пропусками [Текст] // Финансы и статистика, 1991. – 430 с.

20. Снитюк В.Е. Эволюционный метод восстановления пропусков в данных [Текст] // Сборник трудов VI-й Международной конференции «Интеллектуальный анализ информации», г. Киев, 2006. – С. 262 – 271.

21. Google Analytics [Електронний ресурс]. – Режим доступу: https://analytics.google.com/analytics/web – 2021 (дата звернення: 10.10.2021)