МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Сумський державний університет
Факультет електроніки та інформаційних технологій
Кафедра інформаційних технологій

«До захисту допущено»
В.о. завідувача кафедри

_____ Світлана ВАЩЕНКО

____ _____ 2023 р.

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня магістр

зі спеціальності 122 «Комп'ютерні науки»,
освітньо-професійної програми «Інформаційні технології проектування»
на тему: Інтелектуальна інформаційна технологія обробка даних покупців e-commerce додатків

Здобувача групи ІТ.м-22 (шифр групи) Беккера Дмитра Олександровича (прізвище, ім'я, по батькові)

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

_____ (підпис) Дмитро БЕККЕР (Ім'я та ПРІЗВИЩЕ здобувача)

Керівник к.т.н., доц., Анна НЕНЯ (посада, науковий ступінь, вчене звання, Ім'я та ПРІЗВИЩЕ) _____ (підпис)

Суми – 2023

ABSTRACT

The topic of Master's Diploma: «Intelligent Information Technology for the Buyer's Data Processing of E-commerce Applications».

The explanatory note includes an introduction, 4 chapters, conclusions, 44 references, and appendices. The total volume of the work is 93 pages, including 80 pages of the main text, 5 pages of the list of references, and 8 pages of appendices.

The relevance of this work comes from many factors, but the main one is the global trend of most businesses transforming. Businesses are continuously moving and improving the ways they operate. One of the key transformations is moving most of the processes to the web. As a result, huge numbers of purchases and interactions are mainly done within the so-called e-commerce platforms. Those platforms represent the whole set of technologies intended to set up the business processes and make it possible to sell the products and all related to sub-processes like forming cart and order, processing the payment and event automating the delivery, and so on. So, as you can see, many things are done on such platforms and those can be used as valuable information for deep analyses of the customer's behavior. This leads to better sales processes, overall recommendations, and product logistic optimization.

The main goal of this work is to use data science and data mining techniques to extract and discover the valuable data that can help the business understand who the end customers are and to improve the business processes like sales and logistics based on this information.

Firstly, this work focuses on the overall understanding of the data science field and its related scientific works. In the scope of this work, the analysis and comparison of three papers was done. All those papers are related to e-commerce, data science, and data mining techniques to get valuable data. But the focus of such papers is, of course, the methods used to uncover those valuable pieces of information from the regular data that is collected and stored on the e-commerce platforms. This analysis and overall trends plus methods have shaped this work.

The comparison showed that those works mainly focus on the methods used to discover valuable data but need to mention how important it is to prepare the data. Even though the works include quite complicated methods to do the thing, this work uses simple methods to analyze the dataset and retrieve the results based on it. On top of that, the multiple methods used for analysis was compared.

Secondly, in the scope of this study, the tasks and requirements were formed. Having a specific set of goals is important when doing such work. As per essential tasks, the following were chosen:

Initial Data Processing. This part includes cleaning the data set, removing the null values, searching, and eliminating the duplicates, transforming some of the fields to appropriate for analysis data format that will be convenient to use for complicated data processing including the machine learning techniques which most of the time require score based or at least scalar representation of different data.

Augmented extending of data based on statistical analysis. This task includes searching for patterns and trends based on characteristics like frequency, monetary, and so on. And use those findings to classify the customers. Those classifications can be quite valuable for future predictive or classifying models.

Visualizing the results and uncovered insights. This part mainly focuses on bringing visual and understandable results, which later can be used by analysts and the marketing department to build strategies and patterns. But for the scope of this work, it helps to get a better understanding of the findings and start designing the models by simply peeking at the most valuable features of the dataset.

Building regression and classification models. At this stage, the models are built to predict the future behavior of the customers, such as understanding how often and based on what customer will return the products. Moreover, the classification models are used to build the groups of customers and relate them to these specific groups. This is where those selected features come into play, based on that one, those models can train to classify the customers. The last piece here is to compare the models with each other.

In addition, the practical part of the work can be separated into multiple processes. Each requires a set of methods or techniques to achieve the desired results. For example, cleaning the data set is essential to make it a valuable source of discoveries, and this may include removing null values or marking them into some other labels, like the invoices without any customer can be classified as anonymous customers. Those processes require some methods like normalization to keep the data clean and more accurate. But the main methods are used to build the machine learning models. Logistic regression was used. This method is similar to human decision-making based on whether one belongs to a class or not. In such cases, the model helps to classify and gain insight into whether this customer is likely to return the products. As for the other classifications for this work the next methods mentioned were: Logistic Regression, Decision Trees, Random Forrest, and k-nearest Neighbors. Each of the methods was used to classify the customers by RFM analysis. Recency Frequency Monetary (RFM) is quite a popular approach to splitting customers into categories based on certain purchase behaviors. Recency shows how a while ago the customer was buying anything, which leads to some patterns. If such patterns are observed it is likely that the customer will buy more. Frequency and monetary are quite self-explanatory, frequency is needed to see how often the customer buys something, and monetary is how big the total bills are. Overall, combining such different methods and approaches leads to a better understanding and unraveling of valuable information about the customers.

Thirdly, this work is focused on modeling the functional and structural parts of the work. The IDEF0 modeling helps to interpret the inputs, outputs, controls, and mechanisms to see the functionality of the entire analysis. Based on that, a more detailed decomposition diagram was drawn to represent the more detailed flows and nodes that are conducting the actual pieces of work like processing the data or building the classification models at the end of such processes, some outcomes used either by other processes or being the end, one representing the overall results of the working piece of technology.

The fourth chapter of this work represents the found results. The very ground of it is, of course, unraveling the most interesting information about the customers.

In scope of that, it was found that there are quite a lot of anomalies related to the negative quantities. This quite often represents the returns, and based on the visualization results the product selection and many other statistical results were built. On top of that, the logistic regression model was built to see how likely the person would return the products. As an additional part of the work, the RFM approach was used to classify the customers and see what those classes of customers can bring to the overall picture of understanding customer behavior. And it gave a lot like understanding the best customers in the category of champions and seeing what customers are at risk, which means they are likely not to come back and buy something. The classification models that were built on top of those discovered classes can have so many appliances, but for this work, it was used to identify and flag the customers later it can be used for specific propositions building trust strategies, and so on.

In conclusion, this work is focused on unraveling customer behavior from the transactional data. By using data science and data mining techniques and methods the discoveries towards the customers of the e-commerce platform were made.

Keywords: data science, data mining, customer analysis, customer behavior, machine learning.

ЗМІСТ

ВСТУП

У сучасну цифрову епоху галузь електронної комерції зазнала колосальної трансформації, яка змінила спосіб роботи бізнесу та способи здійснення покупок. Ця цифрова революція призвела до стрімкого зростання генерації даних у додатках електронної комерції - потоку інформації, наповненого цінними відомостями про поведінку, вподобання та взаємодію споживачів. Ця скарбниця даних здатна революціонізувати бізнес-стратегії, тому важливо повністю використати її потенціал.

У 21-му столітті дані стали найбільш важливим ресурсом для цифрової сфери, й це особливо актуально для підприємств електронної комерції. У складній павутині даних про клієнтів, що охоплює історії покупок, шаблони перегляду, демографічні профілі та відгуки, міститься сировина для інновацій та зростання. Ці дані не лише висвітлюють вподобання клієнтів, але й прокладають шлях до персоналізованих пропозицій, розширення ринку послуг та маркетингових кампаній.

Зростання обсягу даних створює як можливості, так і виклики, вимагаючи від бізнесу перетворення їх на дієві рішення. Саме в цій трансформації вступають у гру інтелектуальні системи та інтелектуальний аналіз даних, пропонуючи міст від необроблених даних до прийняття обґрунтованих рішень, оптимізації маркетингових стратегій і, насамкінець, підвищення прибутковості.

Інтелектуальний аналіз даних, як важливий компонент штучного інтелекту та машинного навчання, слугує фундаментальним інструментом для виявлення прихованих закономірностей: тенденцій та кореляцій у величезних масивах даних. Ці закономірності дають змогу ефективно сегментувати клієнтів, прогнозувати їхню майбутню поведінку та пропонувати персоналізовані рекомендації, таким чином підвищуючи рівень задоволеності клієнтів та ефективність маркетингу.

Поведінка споживача охоплює рішення та дії, які приймають люди, купуючи та використовуючи продукти. Коли клієнти взаємодіють з онлайн-платформами, вони надають детальну інформацію про свої вподобання та поведінку. Використання цих

даних може значно підвищити ефективність ведення бізнесу, розшифровуючи минулі та поточні моделі поведінки клієнтів, щоб спрогнозувати та зрозуміти потенційних майбутніх клієнтів та їхній вибір. Використовуючи можливості інтелектуального аналізу даних, профілювання та персоналізації клієнтів, компанії можуть розпізнавати тенденції на основі накопичених даних про вподобання та поведінку клієнтів, що дозволяє їм проводити перспективні дослідження, розробляти стратегію своєї діяльності та впроваджувати інновації.

Проте, інтелектуальний аналіз даних – це лише частина рівняння. Інтелектуальні системи, оснащені можливостями штучного інтелекту, виводять цей процес на новий рівень, інтегруючи обробку природної мови, комп’ютерний зір і глибоке навчання. Ці системи надають розширену аналітику, автоматизують процеси прийняття рішень і пропонують прогностичні висновки, передбачаючи потреби клієнтів у змінному середовищі e-commerce.

Кваліфікаційна робота магістра має на меті дослідження інтелектуальних інформаційних технологій обробки та аналізу даних клієнтів платформ електронної комерції для визначення дієвих методів та формування стратегій покращення бізнес процесів.

Основними задачами даної роботи є: визначення предметної області, розуміння значення та потенціалу інтелектуального аналізу даних у сфері електронної комерції; вивчення різних підходів до підготовки даних, забезпечуючи їхню готовність до подальшого аналізу; застосування популярних методів та алгоритмів інтелектуального аналізу даних до готових наборів даних, планування робіт виконання, ідентифікація та стратегії вирішенні ризиків.

1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Огляд останніх досліджень та публікацій

«Підхід інтелектуального аналізу даних до поведінки клієнтів в електронному середовищі», досліджений А. Топе-Оке, К. А. Афолалу та О. Омофаде, занурюється у використання складних алгоритмів кластеризації [1]. Це дослідження вирізняється своїм детальним фокусом на алгоритмічних методологіях, що забезпечує міцну основу для розуміння динаміки на платформах електронної комерції.

Дана робота розширює ці фундаментальні концепції, використовуючи більш комплексний процесно-орієнтований підхід, що охоплює ширший спектр діяльності в життєвому циклі інтелектуального аналізу даних. Це починається з ретельної підготовки набору даних, включаючи збір, очищення та виявлення аномалій, щоб забезпечити якість та надійність. Ці попередні кроки мають вирішальне значення для створення міцного фундаменту для подальшого аналізу.

Виходячи за рамки простого аналізу даних, дана робота акцентує увагу на всьому процесі інтелектуального аналізу даних. Вона включає не лише застосування простіших, але ефективних алгоритмів, таких як K-найближчих сусідів та базові статистичні методи, але й поширюється на завершальні етапи розгортання результатів. Розгортання включатиме створення інтуїтивно зрозумілих діаграм і зведень, які можуть бути безпосередньо використані для інформування та розробки конкретних стратегій електронної комерції.

З точки зору візуалізації та інтерпретованості, подібно до підходу Топе-Оке та ін., дана робота також буде зосереджена на отриманні результатів, які будуть легко зрозумілими та придатними для практичних дій. Важливу роль у цьому відіграватиме використання дерев рішень та видобування асоціативних правил, що дасть чітке уявлення про вподобання та поведінку споживачів.

Підсумовуючи, можна сказати, що хоча дослідження Топе-Оке та його колег забезпечує глибоке занурення в конкретні алгоритми аналізу взаємовідносин між клієнтом і продуктом, дана робота має на меті охопити весь спектр діяльності в галузі інтелектуального аналізу даних. Від початкових етапів підготовки даних до остаточного розгортання дієвих ідей у вигляді стратегій і візуальних узагальнень, проект покликаний відобразити цілісний шлях інтелектуального аналізу даних в контексті електронної комерції.

Дослідження Цзянь Чжана, Пеїхуанг Ліна та Алессандро Сімеоне на тему "Інформаційний аналіз вподобань споживачів для визначення специфікацій продукції з використанням великих даних про продажі" підкреслює критичну роль конкурентоспроможності продукції [2]. В їх роботі важливою є здатність впливати на проектні специфікації, використовуючи великі дані про продажі для розуміння вподобань споживачів. Це дослідження зосереджене на кількісній оцінці вподобань клієнтів за допомогою інформаційної ентропії та використовує методи кластеризації на основі щільності, зокрема DBSCAN, для прийняття рішень при визначенні специфікацій продукту.

Цей підхід особливо пристосований до дизайну продукту у виробничому контексті, використовуючи великі обсяги даних про продажі для узгодження специфікацій продукту з уподобаннями клієнтів. Відрізняючись від цього дослідження, дана робота використовує більш комплексний підхід в процесі інтелектуального аналізу даних. Вона охоплює діяльність від підготовки даних і виявлення аномалій до розгортання дієвих ідей за допомогою діаграм і резюме.

В той час як дослідження Чжана та ін. зосереджене на виробництві та дизайні продукції, використовуючи великі дані для обґрунтованих дизайнерських рішень, дана робота має на меті надати комплексне уявлення про шлях інтелектуального аналізу даних в електронній комерції. Це включає завершення візуальними узагальненнями та стратегічними висновками, які підтримують стратегії електронної комерції.

Ця різниця у фокусі та застосуванні підкреслює унікальні аспекти кожного дослідження. Робота Чжана та ін. пропонує цінну інформацію про специфікацію

продукції на виробництві, тоді як дана робота охоплює ширшу сферу аналізу поведінки клієнтів в електронній комерції. Вона використовує більш прості алгоритми, такі як K-найближчих сусідів та базові статистичні методи, на відміну від більш спеціалізованого підходу до кластеризації на основі щільності та інформаційної ентропії.

Складний метод та комплексний підхід, що задіяні в їхній роботі, не просто зрозуміти та візуалізувати. В свою чергу, дана робота фокусується на комплексному, але більш простому підході до аналізу даних користувачів. Хоча більш глибокі техніки можуть краще розкрити проблему, широкий аналіз показує варіативність підходів та їх ефективність у порівнянні між собою, враховуючи специфіку кожного підходу.

Стаття "Ефективний видобуток внутрішньо-періодичних частих послідовностей" висвітлює обмеження традиційних алгоритмів видобутку частих послідовностей (FSM) та видобутку періодичних частих шаблонів (PFPM) у виявленні шаблонів, які періодично з’являються в даних. У цій роботі розширюється підхід FSM для включення аспектів внутрішньої періодичності, що призвело до розробки нового алгоритму — Intra-Periodic Frequent Sequence Miner. Цей метод враховує як мінімальну, так і максимальну внутрішню періодичність послідовностей, а також їхню частоту у загальній базі даних [3].

Дане дослідження представляє відмінний підхід від ширшого процесу інтелектуального аналізу даних у аналізі поведінки клієнтів в електронній комерції. Хоча стаття фокусується на виявленні періодичних закономірностей, дане дослідження охоплює ширший діапазон дій, від підготовки даних до розробки ефективних стратегій. Методологія дослідження включає простіші алгоритми, такі як K-найближчих сусідів, і базові статистичні методи, що контрастує зі складнішими методами майнингу, використаними у статті.

У даному дослідженні акцентується на розгортанні результатів у формах, які підтримують стратегії електронної комерції, включаючи візуальні узагальнення та стратегічні висновки. З іншого боку, алгоритм у статті зосереджений на підвищенні здатності виявляти періодичні шаблони в різних сферах.

Загалом, внесок статті у видобуток внутрішньо-періодичних послідовностей у PFPM представляє певний аспект інтелектуального аналізу даних, тоді як дане дослідження охоплює більший спектр діяльності в сфері електронної комерції. Ця робота поєднує різні методології для всебічного аналізу даних про клієнтів, пропонуючи широкий огляд поведінки клієнтів в цілому.

Таблиця 1.1 – Порівняння результатів дослідження в оглянутих публікаціях та дослідження здобувача

Аспект	**Дане дослідження**	**Zhang et al. (Product Specifications)**	**Tope-Oke et al. (Customer Behavior)**	**Article on Intra-Periodic Frequent Sequences**
Спектр фокусування	Цілісний аналіз поведінки клієнтів електронної комерції	Конкурентоспроможність продукції через вподобання споживачів у специфікаціях продукції	Інтелектуальний аналіз даних для взаємодії клієнт-продукт в електронній комерції	Видобуток внутрішньоперіодичних частих послідовностей у даних
Методологія	K-Найближчі сусіди, основні статистичні методи, візуалізація даних	Кластеризація на основі щільності, інформаційна ентропія	Консенсусна кластеризація, графічний підхід, аналіз потоку кліків	Внутрішньоперіодичний видобуток, розгляд мінімальної та максимальної внутрішньоперіодичності
Використані дані	Різноманітні набори даних електронної комерції транзакційні дані	Великі дані про продажі для аналізу специфікацій продукції	Інформація про клієнтів та історія з веб-сайтів електронної комерції	Транзакційні дані, впорядковані за часом бази даних
Алгоритмічна складність	Помірний (простіші алгоритми для ширшого розуміння)	Високий (складні розрахунки кластеризації та ентропії)	Високий (передові методи кластеризації та інтеграції даних)	Високий (внутрішньоперіодичний видобуток послідовностей і просунуте розпізнавання образів)

Продовження таблиці 1.1 – Порівняння результатів дослідження в оглянутих публікаціях та дослідження здобувача

Аспект	**Дане дослідження**	**Zhang et al. (Product Specifications)**	**Tope-Oke et al. (Customer Behavior)**	**Article on Intra-Periodic Frequent Sequences**
Кінцева мета/застосування	Всебічне розуміння поведінки клієнтів для формування стратегій електронної комерції	Покращений дизайн продукту шляхом узгодження з уподобаннями клієнтів	Розуміння взаємовідносин між клієнтом і продуктом для стратегії електронної комерції	Виявлення періодичних закономірностей для різних застосувань (наприклад, аналіз ринку, управління запасами)
Виклики/обмеження	Збалансування обсягу з часовими обмеженнями, забезпечення якості та актуальності даних	Чутливість до збору та попередньої обробки даних, придатність для зрілих продуктів	Великі розміри даних, обчислювальна складність, обмеження локального хостингу	Усунення обмежень традиційних FSM і PFPM, складність вилучення шаблонів
Інновації/внесок	Широкомасштабний аналіз та дієві ідеї для стратегій електронної комерції	Новий метод визначення специфікації продукту на основі великих даних про продажі	Комплексний підхід до аналізу поведінки клієнтів з використанням консенсусної кластеризації	Новий алгоритм видобутку внутрішньоперіодичних частих послідовностей, що розширює традиційний FSM

Джерело: розроблено автором

У підсумку можна сказати що дана робота у порівняні з іншими робить більш обширний аналіз використовуючи простіші але дієві методи аналізу даних. Основною перевагою є саме описання процесів які передують ефективному аналізу. Тобто починаючи від очищення даних та пошуку аномалій до оптимізації набору даних для результативного побудування моделей.

1.2 Аналіз сучасного стану галузі

Аналіз клієнтських даних та аналітика цих даних – це процес, що включає в себе збір, організацію та аналіз інформації про цих клієнтів, з метою розуміння їх поведінки. Цей процес відіграє важливу роль в створенні стратегій бізнесу, управлінні продажами та маркетингових кампаній, які задовольняють очікування клієнтів. Для проведення аналізу даних необхідно мати певні інструменти для збору і управління даними, формування гіпотез та правильної інтерпретації отриманих даних [4].

Глибоке розуміння аналізу клієнтських даних передбачає занурення у складний світ клієнтської аналітики. Цей процес стає невід'ємною частиною бізнес-стратегії компанії, яка прагне не тільки зрозуміти, але й передбачити вподобання та поведінку своїх клієнтів. Завдяки систематичному збору і аналізу отриманих даних, компанії виявляють цінну інформацію, яка стає основою у прийнятті стратегічних рішень. Ця інформація використовується під час розробки продуктів і послуг, які відповідають потребами кінцевого споживача [5].

Аналіз клієнтських даних йде далі за простий збір інформації; це комплексний підхід, який включає управління даними та їх глибоку інтерпретацію. Для оптимізації цього процесу компанії використовують різноманітні технології та інструменти, від передового аналітичного ПЗ до систем управління взаємовідносинами з клієнтами (CRM). Ці інструменти дозволяють виявити тенденції та закономірності в даних, що сприяє кращому розумінню потреб клієнтів [6].

Одним з найважливіших аспектів аналізу даних про клієнтів є зосередження на прийнятті рішень, орієнтованих на отримані дані. Бізнес, який покладається на емпіричні дані замість інтуїції, здатен приймати більш обґрунтовані та ефективні рішення. Це не тільки покращує досвід клієнтів, але й сприяє удосконаленню продуктів та більш цілеспрямованим маркетинговим кампаніям [7].

Важливо зазначити, що аналіз даних про клієнтів є гнучкою системою, що адаптується до змін ринку та поведінки споживачів. У зв'язку з цим, компаніям

необхідно бути гнучкими та постійно оновлювати свої аналітичні підходи, щоб залишатися конкурентоспроможними.

Отже, аналіз даних про клієнтів є невід'ємною частиною сучасної стратегії бізнесу. Він забезпечує глибоке розуміння поведінки клієнтів, сприяє прийняттю більш ефективних рішень та зміцнює відносини з клієнтами. Завдяки ефективному використанню аналітичних інструментів та методів, компанії можуть повною мірою розкрити потенціал даних про своїх клієнтів, що стимулює стабільне зростання та успіх [8].

Важливість та застосування аналізу даних про клієнтів у сучасному бізнесі важко переоцінити. Цей аналітичний підхід відіграє ключову роль у розумінні поведінки клієнтів і їх взаємодії з брендом. Проводячи аналіз клієнтів та взаємодії з ними, бізнес знаходить фактори, що мають вплив на задоволеність клієнтів покупками та покращують їхній досвід шляхом вживання необхідних заходів. Адаптація досвіду покупців є одним з основних застосувань аналізу даних про клієнтів. Аналіз даних, отриманих шляхом різних взаємодій клієнтів і e-commerce систем може бути використаний бізнесом для створення персоналізованого досвіду, що задовольняє потреби клієнтів та покращує їх лояльність до бізнесу [9].

Аналіз даних також відіграє важливу роль у розробці маркетингових кампаній, що безпосередньо відповідають потребам аудиторії, та сприяє ефективному використанню маркетингових ресурсів. Водночас, ця інформація виявляє прогалини на ринку та вказує на можливості для вдосконалення продуктів.

Клієнтська аналітика також важлива для стратегій утримання клієнтів, допомагаючи розуміти фактори лояльності та розробляти цільові програми утримання для збереження найцінніших клієнтів.

Завдяки тому, що клієнтські дані безперервно зростають в обсягах та різноманітності платформ, наприклад з соціальних мереж та платформ онлайн продажів, компанії отримують все більше можливостей для вивчення уподобань їхніх клієнтів. Використовуючи актуальні технології, такі як машинне навчання та ШІ, бізнес може більш точно аналізувати величезні обсяги даних та виявляти приховані закономірності поведінки споживачів. Від загальнодоступного сайту до логістики

виконання замовлення - машинне навчання допомагає індустрії електронної комерції краще задовольняти потреби клієнтів. Платформи електронної комерції мають більше даних, ніж будь-коли раніше. Ці дані можна передавати алгоритму, який показує, що цікавить різних клієнтів або відвідувачів e-commerce платформи. Це дозволяє більш точно сегментувати клієнтів [10].

Це, в свою чергу, сприяє розробці інноваційних продуктів та послуг, які краще відповідають очікуванням клієнтів. Крім того, здатність швидко адаптуватися до змін на ринку та в поведінці споживачів є критично важливою для підтримання конкурентоспроможності. Завдяки аналітиці, компанії можуть оперативно реагувати на ці зміни, підтримуючи високий рівень задоволеності та лояльності своїх клієнтів.

Клієнтська аналітика охоплює такі важливі категорії, як аналіз клієнтської подорожі, когортний аналіз та оцінка життєвої цінності клієнта. Кожна категорія пропонує унікальне розуміння різних аспектів взаємодії та поведінки клієнтів [11].

Клієнтська аналітика, особливо в контексті електронної комерції, відіграє життєво важливу роль у розумінні та покращенні клієнтського досвіду.

У динамічному світі електронної комерції, клієнтська аналітика є ключовим інструментом, який допомагає компаніям розуміти та покращувати досвід своїх клієнтів. Вона включає аналіз різноманітних аспектів взаємодії клієнтів з брендом, що сприяє формуванню ефективних бізнес-стратегій.

Довічна цінність клієнта (CLV) вимірює загальну цінність клієнта для бізнесу протягом всього періоду відносин. Розуміння CLV допомагає компаніям зосередитись на утриманні цінних клієнтів та ідентифікувати нових потенційно цінних споживачів [12].

Збір та аналіз даних, включаючи історію транзакцій, демографічні дані та взаємодії з маркетинговими кампаніями, є фундаментом цих аналітичних зусиль. Вони дозволяють зробити висновки про поведінку та вподобання клієнтів, що є надзвичайно цінними для бізнесу.

Інтеграція цих підходів дає компаніям у сфері електронної комерції цілісне розуміння своїх клієнтів. Знання про клієнтську подорож і CLV дозволяють

розробляти більш ефективні та цілеспрямовані стратегії, покращуючи досвід клієнта та ефективність бізнесу в цілому [13].

Основними критеріями аналізу даних користувачів платформи електронної комерції є:

- Розуміння та попередня обробка даних
    - Оцінка, чи є набір даних повним і чи містить він усі необхідні записи та характеристики для аналізу.
    - Визначення точності та чистоти даних, включаючи наявність будь-яких аномалій, пропусків або неправильних значень.
    - Попередня обробка, робота з відсутніми значеннями, нормалізація та трансформація даних.
- Описова статистика:
    - Вивчення розподілу ключових ознак.
    - Пошук значущих кореляції між ознаками, які можуть вплинути на модель або надати бізнес-інформацію.
- Функціональна інженерія:
    - Створення нових функцій, таких як оцінка RFM.
    - Оцінка впливу нових характеристик на продуктивність моделі.
- Побудова та перевірка моделей:
    - Раціоналізувати вибір алгоритмів на основі характеру проблеми класифікації.
    - Використання таких методів, як перехресна перевірка, щоб забезпечити надійність моделі.
- Інтерпретація моделі та бізнес-відкриття:
    - Пояснення результатів аналізу та їх значення в бізнес-контексті.

2. ПОСТАНОВКА ЗАДАЧІ ТА МЕТОДИ ДОСЛІДЖЕННЯ

2.1 Постановка задачі

Виходячи з проаналізованих робіт та загального стану галузі є потреба у розробці за застосуванні інтелектуальної інформаційної технології. Основною метою є вирішення загальних потреб бізнесу у розумінні хто є кінцеві клієнти їх звички та вподобання. Це дослідження фокусується на всіх важливих процесах для побудови такої інтелектуальної системи. Тобто починаючи від процесів очищення даних та загальної обробки та оптимізації набору даних до побудови моделей які можуть класифікації клієнтів відповідно до їх характеристик.

Основними задачами до проекту є:

- Попередній аналіз та очищення набору даних;
- Аргументоване наповнення даних на основі статистичного аналізу;
- Побудова моделей для класифікації;
- Формування рекомендацій для покращення бізнесу.

Метою даного дослідження є розробка технології використання інтелектуального аналізу даних про клієнтів з платформ електронної комерції з метою отримання дієвих висновків та підвищення ефективності стратегій електронної комерції.

Ретельне вивчення предметної області дозволило виявити, що дані про клієнтів e-commerce платформ є джерелом невикористаних даних. Таким чином, було визначено мету проектування: описати та реалізувати основні процеси для комплексного аналізу даних, та побудувати моделі машинного навчання, які будуть формувати відкриття та прогнози. Основною метою є виокремлення з набору даних важливої інформації, аналіз якої, допомагає прийняти обґрунтовані рішення в сфері онлайн продажів.

Для вирішення поставленої задачі необхідно вирішити такі під задачі проектування:

- Аналіз предметної області: визначення актуальності роботи, порівняння з аналогами.
- Значення технологій для реалізації: визначення підходів та інструментів для вирішення задач.
- Збір інформації: пошук відповідних наборів даних про клієнтів електронної комерції з публічних або імітаційних середовищ;
- Підготовка наборів даних: очистка даних, усунення пропущених значень, усунення нерелевантної інформації, щоб забезпечити якість даних;
- Дослідження даних: провести початковий дослідницький аналіз даних, щоб зрозуміти основні властивості, розподіли та особливості набору даних;
- Відбір та інженерія ознак: визначення основних атрибутів даних і можливість створення нових функцій для подальшого вдосконалення аналізу;
- Застосування підходів інтелектуального аналізу даних:
    - Описовий аналіз: заглиблення у внутрішню структуру даних;
    - Кластеризація: використання алгоритмів, як K-середні, щоб розбити дані на ознакові кластери;
    - Предикативний аналіз: розгортання простих моделей машинного навчання, таких як лінійна регресія або дерево рішень для прогнозування;
    - Виявлення аномалій: впровадження алгоритмів або статистичних методів для виявлення незвичайних шаблонів або відхилень у даних;
- Оцінка: загальна оцінка продуктивності і точності застосованих алгоритмів;
- Візуалізація: підвищення рівня презентації результатів, використовуючи передові методи візуалізації, щоб зробити інтерпретацію даних більш інтуїтивно зрозумілою та цікавою;
- Документація: підготовка детального звіту, що містить застосовані методи, досягнуті результати та отримані висновки.

2.2 Функціональні та нефункціональні вимоги

Чітко визначені вимоги є важливими ознаками на шляху до успішного проекту. Вони встановлюють формальну домовленість між клієнтами та постачальниками про те, що вони обидва працюють над досягненням однієї і тієї ж мети. Якісні, детальні вимоги також допомагають зменшити фінансові ризики та дотримуватися графіку виконання проекту. Вони включають в себе чітке формулювання цілей, завдань та потреб. Функціональні та нефункціональні вимоги – це дві фундаментальні категорії вимог у розробці програмного забезпечення. Кожен тип відіграє життєво важливу роль у визначенні характеристик і роботи рішення [14].

Функціональні вимоги – це особливості продукту або функції, які розробники повинні реалізувати, щоб користувачі могли виконувати свої завдання. Це вимоги, які кінцевий користувач висуває як базові можливості, які повинна пропонувати система. Всі ці функціональні можливості повинні бути обов’язково включені в систему як частина контракту. Вони представлені або сформульовані у вигляді вхідних даних, які повинні бути надані системі, виконуваних операцій та очікуваного результату. На відміну від нефункціональних вимог, це, по суті, вимоги користувача, які можна побачити безпосередньо в кінцевому продукті [15]. До застосунку висунуто наступні функціональні вимоги:

- Очищення та формування адекватного та релевантного набору даних;
- Формування додаткових аналітичних ознак на основі статистичного аналізу;
- Візуалізацій для побудови висновків про користувачів e-commerce платформи;
- Побудова моделей для прогнозування та відкриттів: базисних, додаткових ознак та класифікації;
- Формування висновків та відкриттів для команди аналітиків та маркетологів.

Нефункціональні вимоги (NFR) – це набір специфікацій, які описують робочі можливості та обмеження системи і намагаються покращити її функціональність. По суті, це вимоги, які визначають, наскільки добре система буде працювати, включаючи

такі речі, як швидкість, безпека, надійність, цілісність даних і т.д [16]. Основними нефункціональними вимогами до цього проекту є:

- Безпека даних. Дотримання стандартів обробки даних;
- Масштабування. Можливість аналізувати більші об’єми даних;
- Оптимізація. Аналіз даних, візуалізації та тренування моделей повинні бути обмежені часовими рамками;
- Доступність. Можливість мати доступ до результатів у вигляді простого звіту без потреби проганяти код.

2.3 Технології та інструменти реалізації

Для виконання аналізу даних та побудови візуалізації необхідно обрати такі інструменти як: мова програмування, інструменти для обробки й візуалізації даних та побудови моделей.

Найбільш важливим критерієм при виборі технологій та інструментів розробки є вибір мови програмування. А потім, відповідно до обраної мови підбираються інші інструменти, які матимуть важливий вплив на результати роботи. З поміж різних мов програмування для даної роботи було обрано Python, адже вона є універсальною мовою програмування, яку можна використовувати як для розробки веб-додатків, так і мобільних додатків. ЇЇ можна застосовувати також і в розробці складних числових і наукових програм. Синтаксис мови Python є простим і зрозумілим, тому нею можуть користуватися програмісти всіх рівнів кваліфікації. Python має велику кількість бібліотек, призначених в першу чергу для аналізу даних. Існує велика та активна спільнота розробників Python, яка займається наукою про дані та розробкою програмного забезпечення. Підтримка спільноти має вирішальне значення для вирішення проблем, обміну передовим досвідом та отримання актуальної інформації про прориви в аналітиці даних [17].

Наступними на черзі є інструменти для роботи з даними. І тут ідеально підходить використання Pandas. Бібліотека зарекомендована для роботи з наборами даних різної складності та має в арсеналі велику кількість інструментів. Pandas - це бібліотека для маніпулювання даними на Python, яка спрощує обробку даних у багатьох IT– секторах та ролях, включаючи аналітиків, data scientists, ETL– інженерів та інших, для читання, обробки та запису даних. Більше того, як це зазвичай буває з Python, всі дії можна написати в 1-2 рядках. Крім того, при такій короткій довжині коду, бібліотека Pandas надає дуже зручні та зрозумілі механізми обробки даних. Основними структурами даних Pandas є серії та фрейми даних; тут ми розглянемо, як інструменти бібліотеки Pandas можуть допомогти вам в управлінні даними та аналізі даних [18].

Для виконання складних обчислень було обрано NumPy. NumPy (Numerical Python) – це бібліотека Python з відкритим вихідним кодом, яка використовується майже у всіх галузях науки та інженерії. Це універсальний стандарт для роботи з числовими даними на Python, який лежить в основі наукових екосистем Python та PyData. Користувачами NumPy є всі – від програмістів-початківців до досвідчених дослідників, які проводять найсучасніші наукові та промислові дослідження та розробки. API NumPy широко використовується в Pandas, SciPy, Matplotlib, scikit-learn, scikit-image та більшості інших пакунків для науки про дані та наукового Python. Бібліотека NumPy містить багатовимірні структури даних у вигляді масивів і матриць. NumPy можна використовувати для виконання широкого спектру математичних операцій над масивами. Він додає до Python потужні структури даних, які гарантують ефективні обчислення з масивами та матрицями, а також надає величезну бібліотеку високорівневих математичних функцій, які оперують з цими масивами та матрицями [19].

Scikit-learn – це популярна і надійна бібліотека машинного навчання, яка має широкий набір алгоритмів, а також інструменти для візуалізації, попередньої обробки, підгонки, вибору та оцінювання моделей. Побудована на основі NumPy, SciPy та matplotlib, Scikit-learn має низку ефективних алгоритмів для класифікації, регресії та кластеризації. До них відносяться машини опорних векторів, дощові ліси,

градієнтний бустинг, k-середні та DBSCAN. Scikit-learn може похвалитися відносною простотою розробки завдяки послідовному та ефективно спроектованому API, обширній документації для більшості алгоритмів та численним онлайн-урокам. Бібліотека машинного навчання (ML) для мови програмування Python, Scikit-learn має велику кількість алгоритмів, які можуть бути легко розгорнуті програмістами та дослідниками даних у моделях машинного навчання [20].

Отже, аналіз даних про клієнтів є невід'ємним елементом стратегії кожного сучасного бізнесу. Matplotlib використовується для візуалізації даних та формування важливих висновків на основі вхідних даних. Ця бібліотека надає API для побудови різних графіків для користувачів Python та NumPy. Matplotlib є достатньо простою у використанні і включає в себе перелік графічних інструментів, необхідних для отримання обширних візуалізацій [21].

Seaborn – це бібліотека для створення статистичної графіки на Python. Вона побудована на основі matplotlib і тісно інтегрується зі структурами даних pandas. Seaborn допомагає досліджувати та візуально сприймати дані. Його функції побудови графіків працюють з фреймами даних і масивами, що містять цілі набори даних, і внутрішньо виконують необхідне семантичне відображення і статистичну агрегацію для створення інформативних графіків. Декларативний API, орієнтований на набори даних, дозволяє зосередитися на тому, що означають різні елементи графіків, а не на деталях їх побудови [22].

Бібліотека plotly для Python – це інтерактивна бібліотека побудови графіків з відкритим кодом, яка підтримує понад 40 унікальних типів діаграм, що охоплюють широкий спектр статистичних, фінансових, географічних, наукових та тривимірних застосувань. Побудована на основі JavaScript бібліотеки Plotly (plotly.js), plotly дозволяє користувачам Python створювати красиві інтерактивні веб-візуалізації, які можна відображати в блокнотах Jupyter, зберігати в окремих HTML-файлах або використовувати як частину чистих веб-додатків, створених на Python за допомогою Dash [23].

Bokeh – це бібліотека Python, яка використовується для створення інтерактивних візуалізацій у веб-браузері. Вона надає потужні інструменти, які

забезпечують гнучкість, інтерактивність та масштабованість для дослідження різноманітних даних. Bokeh – це велика бібліотека, яка пропонує різноманітні графіки та діаграми для дослідження різноманітних даних. Можна створювати різні візуальні діаграми для представлення статистичної інформації з наборів даних [24].

Jupyter Notebook (раніше відомий як IPython Notebook) – це інтерактивний веб-додаток для створення та обміну обчислювальними документами. Це повністю відкритий продукт, і користувачі можуть використовувати всі доступні функції безкоштовно. Блокнот – це інтерактивне обчислювальне середовище, в якому користувачі можуть виконувати певний фрагмент коду і спостерігати за результатами, а також вносити зміни до коду, щоб привести його до бажаного результату або дослідити більше. Блокноти Jupyter широко використовуються для дослідження даних, оскільки вони передбачають багато повторень. Він також використовується в інших робочих процесах науки про дані, таких як експерименти та моделювання машинного навчання. Його також можна використовувати для документування зразків коду. Блокнот Jupyter має незалежні комірки з виконуваним кодом, які користувачі можуть запускати в будь-якому порядку. Документування можна здійснювати, чергуючи комірки з кодом та розміткою [25].

Colaboratory, або скорочено "Colab", – це продукт від Google Research. Colab дозволяє будь-кому писати і виконувати довільний код на python через браузер, і особливо добре підходить для машинного навчання, аналізу даних та освіти. З технічної точки зору, Colab – це хостинговий сервіс для ноутбуків Jupyter, який не потребує жодних налаштувань, але надає безкоштовний доступ до обчислювальних ресурсів, включаючи графічні процесори [26].

Для реалізації проекту буде задіяно наступний список технологій: Python з його набором бібліотек, таких як: Pandas, NumPy та Scikit-learn, стане наріжним каменем для маніпуляцій з даними, майнінгу та завдань машинного навчання. Окрім традиційних Matplotlib та Seaborn у Python, будуть використовуватись більш просунуті бібліотеки візуалізації, такі як Plotly або Bokeh. Ці бібліотеки дозволяють створювати інтерактивні та естетично привабливі візуалізації.

2.4 Методи дослідження

Розробка інтелектуальної технології для аналізу даних користувачів у сфері електронної комерції вимагає уважного підходу до вибору методів, які є ключовими для ідентифікації цінних відкриттів та максимального використання потенціалу зібраних даних.. Цей процес не лише вимагає глибокого розуміння поведінки клієнтів, яке має загальний та абстрактний характер, але й орієнтується на точний аналіз даних для формування конкретних припущень, що ведуть до визначення трендів, подібностей та інших важливих характеристик. За допомогою цих характеристик розробляються моделі класифікації, що становлять основу аналітичної роботи.

Переходячи від теоретичних основ до практичних аспектів, наступним важливим етапом є підготовка даних. Підготовка даних для аналізу в сфері електронної комерції включає в себе ряд ключових процедур, спрямованих на підвищення їх якості та аналітичної цінності. Ці процедури охоплюють різноманітні аспекти, від нормалізації та вибірки до оптимізації, очищення та виявлення аномалій. Особливу увагу приділяється усуненню нетипових закономірностей, таких як аномальні ціни, які можуть спотворювати аналіз. Ці процеси не тільки покращують якість даних, але й готують їх до більш глибокого аналітичного вивчення та моделювання, що є важливим для виявлення інсайтів та ідентифікації потенційних проблем із даними.

Розглянемо більш детально кожен із ключових кроків у процесі підготовки даних. Процес підготовки даних у контексті електронної комерції охоплює різноманітні кроки, які в сукупності покращують їх аналітичну цінність [27]. Від очищення даних, що включає видалення помилкових або неповних записів, до вибірки та оптимізації - кожен етап має свій внесок у загальну якість аналізу. Виявлення аномалій, наприклад, дозволяє виявити та усунути викиди, що можуть вплинути на точність результатів. Також значну увагу приділяється трансформації даних, включаючи нормалізацію та масштабування, щоб вони були готові до

подальшого аналітичного використання. Ці маніпуляції з даними спрямовані на забезпечення їх відповідності технічним вимогам моделей машинного навчання та підвищення ефективності аналітичних алгоритмів.

Окрім технічних аспектів підготовки даних, важливо врахувати їх вплив на результати машинного навчання та аналітичного моделювання. У сфері машинного навчання підготовка даних відіграє ключову роль. Проекти предиктивного моделювання, особливо ті, що включають структуровані або табличні дані, вимагають перетворення вихідних даних відповідно до вимог окремих алгоритмів машинного навчання. Вибір способу представлення даних має вирішальне значення для ефективного виявлення невідомої структури, що лежить в основі проблеми прогнозування [28].

Основні методи підготовки даних, такі як нормалізація, обробка відсутніх даних, трансформація категоріальних змінних та оптимізація ознак, є вирішальними для створення ефективних аналітичних моделей. Ці методи є невід'ємною складовою для того, щоб зробити дані більш керованими та інтерпретованими для алгоритмів, які використовуються в аналізі даних. Належним чином підготовлені дані призводять до більш надійних результатів, дозволяючи приймати більш обґрунтовані рішення, знижуючи витрати на управління даними та аналітику, і значно сприяють успіху проектів з інтелектуального аналізу даних та машинного навчання.

У контексті цієї роботи, яка зосереджена на даних клієнтів електронної комерції, підготовка даних буде наріжним каменем аналізу. Такі методи, як кластеризація, групування та нормалізація, матимуть вирішальне значення для перетворення даних у форму, яка буде зрозумілою та сприятливою для прогнозного моделювання. Цей процес узгоджується з основною концепцією підготовки даних, уточнюючи дані для більш точних і дієвих прогнозів у секторі електронної комерції.

Завдяки ефективній підготовці даних проект має на меті використовувати передові методи інтелектуального аналізу даних для збагачення наборів даних. Таке збагачення не лише готує дані для таких моделей прогнозування, як k-найближчих сусідів та дерева рішень, але й слугує для покращення стратегій утримання клієнтів та рекомендацій щодо продуктів. Отже, ці зусилля принесуть значну користь

платформі електронної комерції, сприяючи зростанню бізнесу та задоволеності клієнтів.

Підготовка даних для аналізу в сфері електронної комерції починається з ретельного очищення даних. Цей процес включає усунення недосконалостей, таких як неповні, дубльовані, нерелевантні або нульові значення, що забезпечує цілісність та точність даних. Очищення даних адаптується до конкретних потреб проекту та спрямоване на підготовку надійного фундаменту для аналізу. Під час цього процесу особлива увага приділяється забезпеченню відповідності даних регулятивним стандартам, що є ключовим для забезпечення їхньої надійності та валідності [29].

Далі звертається увага на ефективне управління відсутніми даними. Стандартні практики включають видалення записів з пропущеними даними або їх заповнення середніми значеннями або медіаною, що допомагає зберегти представництво та надійність даних. Цей крок важливий для запобігання викривленням в аналітичних моделях, що може виникати внаслідок неповноти даних.

Ключовим аспектом підготовки даних є також зменшення їх розмірності та інженерія ознак. Це включає усунення непотрібних змінних та створення нових ознак, які можуть допомогти виявити приховані закономірності у даних. Це сприяє більш ефективному аналізу та підвищує продуктивність моделей. Оптимізація ознак включає в себе використання алгоритмічних методів для виявлення та відбору найбільш інформативних ознак, що значно підсилює аналітичні можливості моделі.

Стратегічна вибірка та перетворення даних дозволяє краще керувати великими обсягами інформації. Нормалізація та масштабування даних використовуються для досягнення більшої універсальності та придатності для аналізу. Ефективне перетворення даних також допомагає уникнути спотворень, пов’язаних із масштабом та розподілом ознак.

Нарешті, важливо вирішити проблему незбалансованості даних. Використання методів, як-от надмірна або недостатня вибірка певних класів, забезпечує створення збалансованого набору даних, що сприяє точності та об’єктивності аналізу. Правильне управління незбалансованими даними дозволяє отримати більш точні та

представницькі результати, що є критично важливим для прийняття обґрунтованих бізнес-рішень.

У сфері електронної комерції аналіз даних клієнтів починається з класифікації – розподілу даних за певними класами на основі їх атрибутів. Це дозволяє розділити клієнтів на групи за купівельною поведінкою або демографічними характеристиками, що в свою чергу допомагає у персоналізації маркетингових стратегій та оптимізації клієнтського досвіду. Цей процес класифікації стає основою для розробки більш цілеспрямованих продуктів і послуг.

Далі йде регресійний аналіз, який використовується для прогнозування майбутніх продажів на основі наявних даних. Цей метод допомагає визначити, як змінні, такі як покупки клієнтів або маркетингові кампанії, впливають на продажі. Точні прогнози, отримані за допомогою регресійного аналізу, забезпечують цінну інформацію для стратегічного планування та управління запасами [30].

Кластеризація є наступним ключовим етапом, де схожі дані групуються разом на основі характеристик клієнтів або продуктів. У контексті електронної комерції, це дозволяє ідентифікувати природні групи товарів або клієнтів і розробляти цільові маркетингові ініціативи. Це сприяє підвищенню ефективності рекомендацій продуктів та підвищенню задоволеності клієнтів [30].

Виявлення аномалій відіграє важливу роль у захисті електронної комерції від шахрайства та контролю якості. Аналіз незвичайних даних допомагає виявити потенційно шкідливі або аномальні транзакції, що може бути ознакою шахрайства чи помилкових даних.

Аналіз часових рядів є ключовим інструментом у електронній комерції для розуміння динаміки ринку та споживацької поведінки. Він дозволяє ідентифікувати та прогнозувати сезонні варіації в продажах, а також визначити оптимальні часові рамки для проведення маркетингових кампаній і спеціальних акцій. Завдяки аналізу часових рядів, підприємства можуть адаптувати свої запаси та розподіл ресурсів, щоб відповідати очікуваному попиту в різні періоди року, тим самим підвищуючи ефективність своєї діяльності і задоволеність клієнтів [31].

2.5 Деталізований огляд методів

Для кращого розуміння кожного із складних процесів для комплексного аналізу даних буде описано кожен підхід та можливі аналоги, так як складність та призначення можуть сильно різнитися як від структури даних так і поставленої цілі яку потрібно отримати.

Очищення даних є критичним етапом у процесі аналізу даних, особливо в контексті електронної комерції, де якість та цілісність даних безпосередньо впливають на точність інсайтів та бізнес-рішень. Для ідентифікації та видалення дублікатів часто використовуються SQL-запити або функції обробки даних, такі як drop_duplicates() в Pandas, що допомагає усунути дублікати, які можуть спотворити аналіз.

Обробка відсутніх значень також є важливою частиною процесу. Відсутні дані можуть бути оброблені різними методами, залежно від контексту. Можливі підходи включають видалення записів із відсутніми даними або їх заповнення за допомогою середнього значення, медіани або найпоширенішого значення. Використовується формула середнього значення [32]:

$$\bar{x} = \frac{\sum x_i}{n},$$

де x_i – індивідуальні значення, а n – кількість значень.

Виправлення неточностей та помилок у даних є ще однією важливою частиною. Це може включати виправлення помилок у введенні даних, використання регулярних виразів для стандартизації текстових значень, а також заміну нестандартних або неоднозначних термінів.

Нормалізація даних є важливим кроком у підготовці даних для аналізу, особливо в контексті електронної комерції, де дані часто є різноманітними та багатоатрибутними. Основна мета нормалізації полягає у приведенні всіх даних до

спільного масштабу, не втрачаючи при цьому інформацію про відносини між значеннями. Це дозволяє алгоритмам машинного навчання більш ефективно вчитися на даних, оскільки усувається проблема, коли певні атрибути через свою масштабність надмірно впливають на модель.

Нормалізація забезпечує однакове важливість всім атрибутам, допомагаючи уникнути спотворення у моделях, спричиненого великими або маленькими масштабами даних. Це особливо важливо в електронній комерції, де дані можуть варіюватися від цінових показників до оцінок користувачів [33]:

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}}.$$

Таке масштабування дозволяє моделям обробляти різні типи атрибутів однаково, сприяючи виявленню реальних закономірностей та взаємозв'язків у датасеті.

Завдяки нормалізації можна більш ефективно аналізувати поведінку клієнтів, адже дані представлені у форматі, який відображає їх справжні характеристики. Це дає змогу проводити більш точне сегментування та персоналізацію, враховуючи різні атрибути клієнтських даних. Отже, нормалізація є ключовим кроком у підготовці даних, який забезпечує точність та ефективність аналітичних моделей в електронній комерції.

Обробка категоріальних даних також важлива. Категоріальні дані часто перетворюються у числовий формат за допомогою методів, таких як one-hot encoding або label encoding, для забезпечення сумісності з алгоритмами машинного навчання.

Виявлення та видалення аномалій є ще одним важливим кроком. Виявлення аномалій, наприклад за допомогою методів Z-оцінки або IQR, допомагає ідентифікувати та вилучити помилкові дані. Формула Z-оцінки [34]:

$$Z = \frac{(x-\mu)}{\sigma},$$

де μ – середнє значення, а σ – стандартне відхилення, використовується для ідентифікації аномалій у даних.

Кожен з цих етапів відіграє важливу роль у подальшому аналізі даних та забезпечення точності отриманих результатів. Впровадження систематичного та добре структурованого процесу очищення даних є критично важливим для якісного аналізу даних.

Виділення ознак у дата аналітиці, особливо в контексті електронної комерції, включає ретельний вибір і аналіз даних, які найкраще відображають поведінку та вподобання клієнтів. Цей процес допомагає ідентифікувати ключові ознаки, що впливають на рішення клієнтів і відкривають нові відкриття для бізнесу.

"RFM" в RFM-аналізі означає релевантність, частоту та грошову цінність. RFM-аналіз - це спосіб використання даних, заснованих на поведінці існуючих клієнтів, для прогнозування того, як новий клієнт, ймовірно, буде діяти в майбутньому. Модель RFM будується з використанням трьох ключових факторів:

- як давно клієнт здійснював транзакцію з брендом
- як часто він взаємодіяв з брендом
- скільки грошей він витратив на продукти та послуги бренду.

Замість того, щоб сегментувати клієнтів лише на основі демографічних та психографічних даних, маркетологи можуть створювати сегменти на основі реальної поведінки людей, включаючи історію покупок через будь-який канал (онлайн чи офлайн), історію переглядів, відповіді на попередні кампанії тощо. Не дивно, що такий тип сегментації називається поведінковою сегментацією.

І навіть базова CRM-система може виконувати елементарне відстеження трьох легко вимірюваних характеристик, які сприяють RFM-аналізу:

Значення рецедентності: Це кількість часу, що минув з моменту останньої взаємодії клієнта з брендом, яка може включати в себе останню покупку, відвідування веб-сайту, використання мобільного додатку, "лайк" у соціальних мережах тощо. Повторюваність є ключовим показником, оскільки клієнти, які нещодавно взаємодіяли з вашим брендом, з більшою ймовірністю відгукнуться на нові маркетингові зусилля.

Значення частоти: Це кількість разів, коли клієнт здійснив покупку або іншим чином взаємодіяв з вашим брендом протягом певного періоду часу. Частота є ключовим показником, оскільки вона показує, наскільки глибоко клієнт взаємодіє з вашим брендом. Більша частота вказує на вищий ступінь лояльності клієнта.

Грошове вираження: Це загальна сума, яку клієнт витратив на придбання продуктів і послуг вашого бренду за певний період часу. Грошова цінність є ключовим показником, оскільки клієнти, які витратили найбільше в минулому, з більшою ймовірністю витрачатимуть більше в майбутньому.

Обчислення RFM для реального застосування, як правило, вимагає спеціальних аналітичних знань або просунутих математичних навичок. І, як і будь-яка модель, моделі RFM можуть варіюватися за складністю від простих до складних. Сегментація RFM починається з ранжування клієнтів у кожній з трьох категорій: за частотою, частотою та грошовою оцінкою. Зазвичай це робиться за шкалою від 1 до 10. Десятка позначає 10% найкращих у кожній категорії (тобто тих, хто нещодавно здійснював транзакції, найчастіше здійснював транзакції та купував найбільше), дев’ятка - наступні 10% і так далі. Використовуючи таку систему оцінювання RFM, ви можете побудувати ефективну маркетингову стратегію, зокрема, шляхом створення сегментів RFM клієнтів:

Найкращі клієнти: Це клієнти, які отримують найвищі бали в кожній категорії. Вони лояльні, готові щедро витрачати і, швидше за все, незабаром зроблять ще одну покупку. Такі клієнти добре реагують на програми лояльності. Вони з більшою ймовірністю зацікавляться новими продуктами, які ви запускаєте. А оскільки вони прихильні до вашого бренду та його продуктів, то, ймовірно, не має сенсу пропонувати їм знижки з точки зору бізнесу. Натомість збільшуйте CLTV, пропонуючи товари з великими знижками та рекомендуючи товари на основі минулих покупок.

Клієнти з групи ризику: Клієнти, які в минулому входили до вашого топ-рівня (найкращі, великі покупці та/або лояльні), але зараз мають низькі показники частоти та повторюваності покупок, представляють особливу можливість. Маркетологи повинні розглянути можливість націлювання на них повідомлень, спрямованих на

утримання, таких як знижки, ексклюзивні пропозиції та запуск нових продуктів. За допомогою CDP ви навіть можете створювати спеціальні клієнтські подорожі, спрямовані на повторне залучення та утримання клієнтів з групи ризику [35].

SMOTE – це метод синтетичної вибірки, що генеруються для класу меншин. Цей алгоритм допомагає подолати проблему надмірної пристосованості, що виникає при випадковій передискретизації. Він фокусується на просторі ознак, щоб генерувати нові екземпляри за допомогою інтерполяції між позитивними екземплярами, які лежать поруч. Спочатку встановлюється загальна кількість спостережень попередньої обробки, N. Як правило, вона вибирається таким чином, щоб розподіл бінарних класів був 1:1. Але це значення може бути змінено за потреби. Потім ітерація починається з випадкового вибору позитивного екземпляра класу. Потім для цього екземпляра отримується КНФ (за замовчуванням 5). Нарешті, вибирається N з цих K екземплярів для інтерполяції нових синтетичних екземплярів. Для цього, використовуючи будь-яку метрику відстані, обчислюється різниця у відстані між вектором ознак та його сусідами. Тепер ця різниця множиться на будь-яке випадкове значення з діапазону (0,1) і додається до попереднього вектора ознак [36]. Це наочно показано нижче:

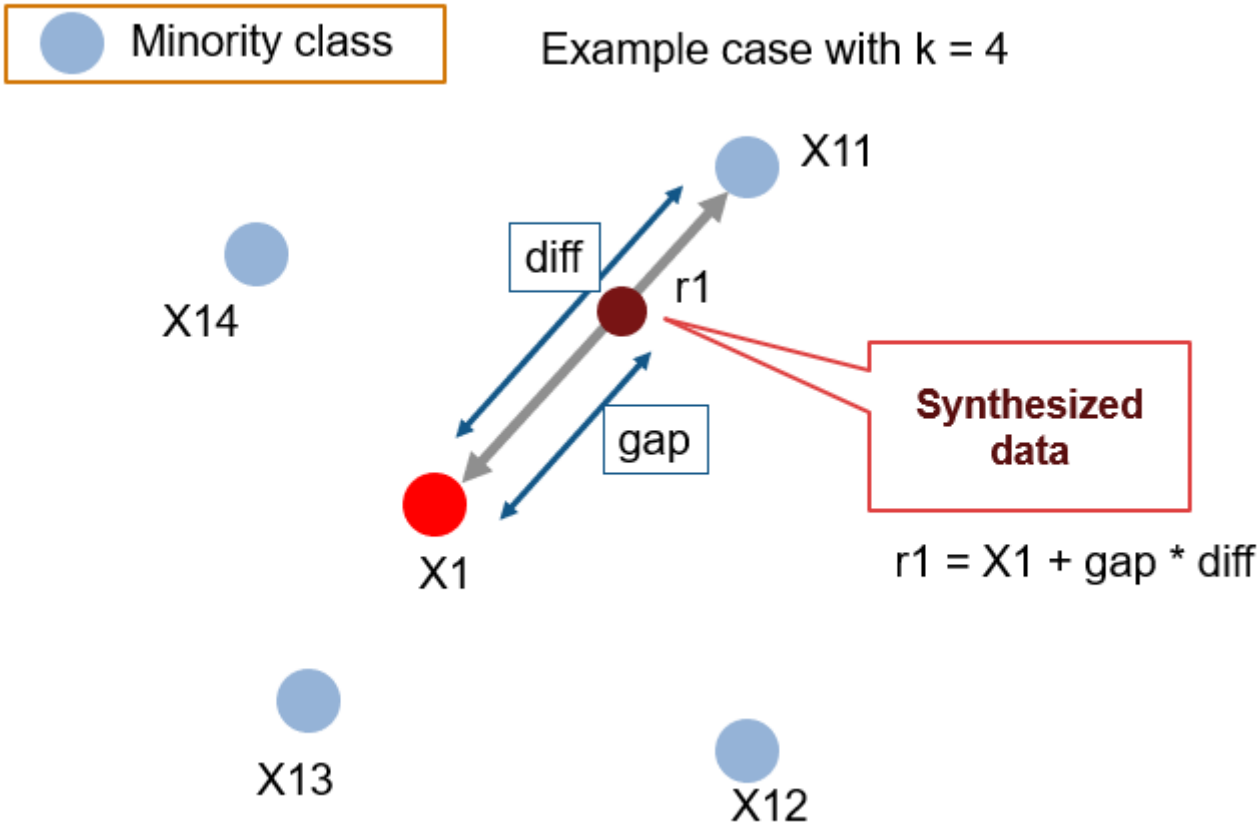


Рисунок 2.1 – Приклад роботи алгоритму SMOTE

Джерело: SMOTE for Imbalanced Classification with Python [36]

Мінорний клас (сині точки): Ці точки представляють існуючі екземпляри незбалансованого класу в наборі даних. Клас меншості - це клас у наборі даних, який має менше екземплярів і тому може потребувати надмірної вибірки, щоб збалансувати розподіл класів.

X1 (червона точка): Це конкретний екземпляр з мінорного класу, який було обрано як основу для генерації нових синтетичних екземплярів.

X11, X12, X13, X14 (сині точки): Це k найближчих сусідів X1, знайдених у класі меншості. Кількість найближчих сусідів, k, є параметром алгоритму SMOTE; у цьому випадку k=4.

r1 (точка на прямій між X1 та одним з її найближчих сусідів): Це синтетична точка даних, що генерується алгоритмом SMOTE. Вона створюється шляхом вибору одного з k найближчих сусідів (наприклад, X11), а потім створення нової вибірки в точці на відрізку між X1 і X11.

diff (векторна різниця між X1 та обраним сусідом): Це вектор, отриманий шляхом віднімання вектора ознак X1 від вектора ознак вибраного найближчого сусіда (наприклад, X11 - X1). Цей вектор використовується для створення нової точки вздовж напрямку лінії, що з'єднує точку X1 з її сусідом.

проміжок (скаляр): Випадкове число від 0 до 1. Цей скаляр множиться на вектор "diff", щоб визначити положення синтезованого зразка на відрізку між X1 і його сусідом.

Синтезовані дані (червона лінія і точка r1): Остаточна синтетична вибірка створюється шляхом додавання масштабованого вектора "diff" до вихідного вектора ознак X1 (тобто, r1 = X1 + gap * diff). Потім ця вибірка додається до набору даних, щоб збільшити кількість екземплярів у класі меншини.

Логістична регресія - це особливий випадок регресійного аналізу, який використовується, коли залежна змінна є номінально масштабованою. Це стосується, прилежної до класу користувачів на основі їх ознак. Тобто належить клієнт до якогось класу чи ні. Таким чином, логістичний регресійний аналіз є аналогом лінійної регресії, в якому залежна змінна регресійної моделі повинна бути принаймні інтервально масштабованою [37]. Загальна концепція класифікації через логістичну регресію представлено на рисунку 2.1.

Ймовірність того, що для заданих значень незалежної змінної дихотомічна залежна змінна у дорівнює 0 або 1, задається через [37]:

$$\mathrm{P}(y = 1|x_1 \dots x_n) = \frac{1}{1+e^{-(b_1 \cdot x_1 + \dots + b_k \cdot x_k + a)}},$$

$$\mathrm{P}(y = 0|x_1 \dots x_n) = 1 - \frac{1}{1+e^{-(b_1 \cdot x_1 + \dots + b_k \cdot x_k + a)}}.$$

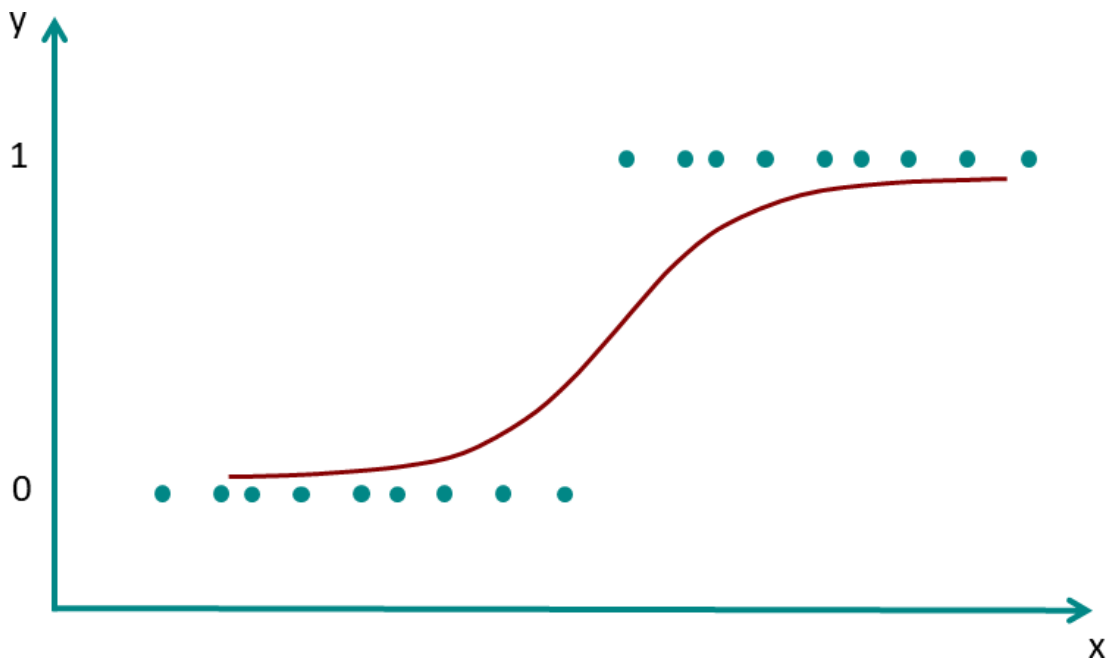


Рисунок 2.1 – Вигляд концепції логістичної регресії

Джерело: Logistic Regression [37]

Щоб обчислити ймовірність того, що клієнт належить або не належить до класу, використовуючи логістичну регресію для наведеного вище прикладу, спочатку необхідно визначити параметри моделі b1, b2, b3 і a. Логістична регресія має ряд важливих практичних застосувань у сфері електронної комерції. Одним з ключових використань є прогнозування покупок, де на основі попередніх поведінкових даних і демографічних характеристик клієнта визначається ймовірність придбання певного товару. Це дозволяє компаніям точніше адаптувати свої маркетингові стратегії та пропозиції до конкретних потреб і інтересів клієнтів. Іншим значущим застосуванням логістичної регресії є визначення ймовірності відтоку клієнтів. Використовуючи дані про історію покупок та взаємодії клієнтів, можна оцінити ризик відтоку і вжити відповідних заходів для утримання клієнтів. Цей аналіз допомагає бізнесу більш ефективно реагувати на потреби клієнтів, підвищуючи їх лояльність та загальну задоволеність послугами компанії [37].

Дерево рішень - це непараметризований алгоритм керованого навчання, який використовується як для задач класифікації, так і для задач регресії. Він має ієрархічну, деревовидну структуру, яка складається з кореневого вузла, гілок,

внутрішніх вузлів і листових вузлів. Дерева рішень є популярними, оскільки вони ближче, ніж інші класифікатори, імітують процес прийняття рішень людиною і є простими для розуміння та інтерпретації [38]. Візуальну репрезентацію алгоритму дерева рішень можна побачити на рисунку 2.2.

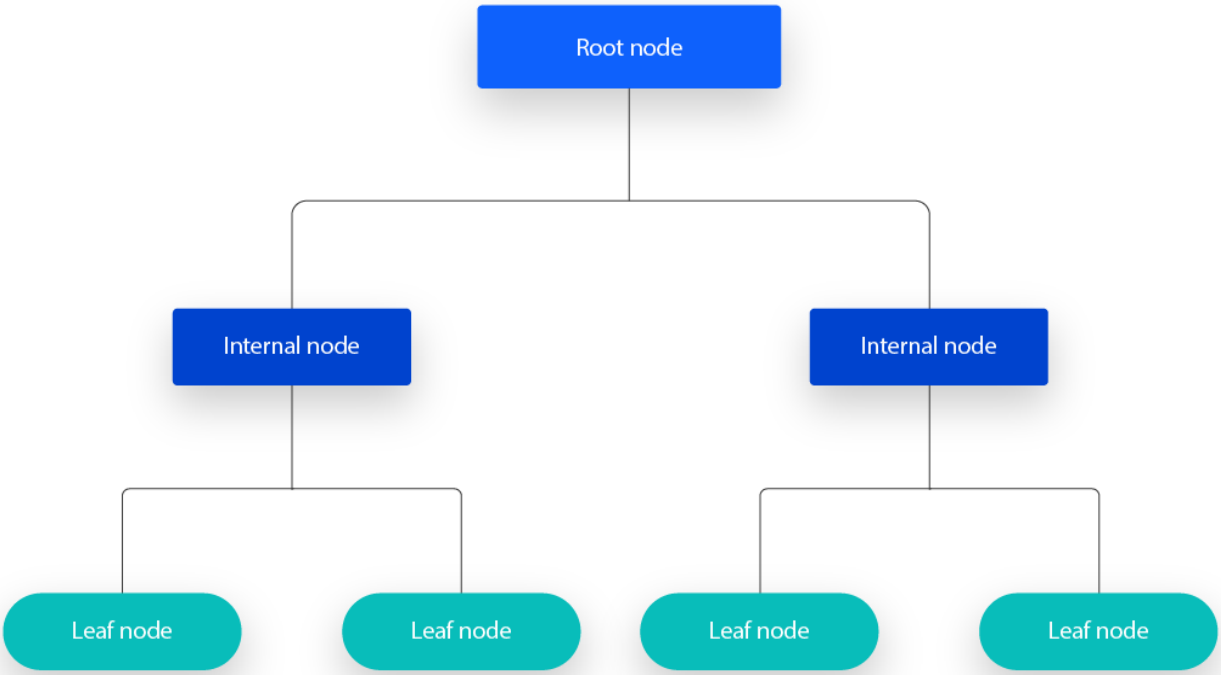


Рисунок – 2.3 Загальна концепція дерева рішень

Джерело: What is a Decision Tree? [38]

Як видно з наведеної вище схеми, дерево рішень починається з кореневого вузла, який не має жодних вхідних гілок. Вихідні гілки з кореневого вузла потім потрапляють у внутрішні вузли, також відомі як вузли прийняття рішень. На основі наявних можливостей обидва типи вузлів проводять оцінки для формування однорідних підмножин, які позначаються листовими вузлами або термінальними вузлами. Листяні вузли представляють всі можливі результати в наборі даних [38].

Задано навчальні вектори $x_i \in R^n$, i=1,..., l та вектор міток $y \in R^l$ дерево рішень рекурсивно розбиває простір ознак таким чином, щоб зразки з однаковими мітками або схожими цільовими значеннями були згруповані разом.

Нехай дані у вузлі m подано у вигляді $\mathrm{Q_m}$ з $\mathrm{n_m}$ вибірок. Для кожного кандидата розбиття $\theta = (\mathrm{j}, \mathrm{t_m})$ що складається з ознаки j та порогу $\mathrm{t_m}$ розбиваємо дані на $\mathrm{Q_m^{left}}(\theta)$ та $\mathrm{Q_m^{right}}(\theta)$ підмножини [39]

$$Q_m^{left}(\theta) = \{(x, y)|x_j \leq t_m\},$$
$$Q_m^{right}(\theta) = \frac{Q_m}{Q_m^{left}(\theta)}.$$

Якість розбиття вузла-кандидата обчислюється за допомогою функції домішок або функції втрат вибір якої залежить від розв'язуваної задачі (класифікація або регресія)

$$G(Q_m, \theta) = \frac{n_m^{left}}{n_m} H(Q_m^{left}(\theta)) + \frac{n_m^{right}}{n_m} H(Q_m^{right}(\theta)).$$

Вибір параметрів, які мінімізують домішки

$$\theta^* = argmin_\theta G(Q_m, \theta).$$

Рекурсив для підмножин $\mathrm{Q_m^{left}}(\theta^*)$ та $\mathrm{Q_m^{right}}(\theta^*)$ доки не буде досягнуто максимально допустимої глибини, $\mathrm{n_m} < \mathrm{min_{samples}}$ або $\mathrm{n_m} = 1$.

Рішення в кожному вузлі приймається на основі значення однієї з вхідних ознак. Ознака, яка використовується для прийняття рішення, і тип порівняння (більше, менше, дорівнює, не дорівнює) вибираються на основі їх здатності зменшувати невизначеність. Ця невизначеність часто оцінюється кількісно за допомогою таких метрик, як домішка Gini або ентропія.

Якщо ціль - це результат класифікації, що приймає значення 0,1,...,K-1 для вузла [39]. Нехай

$$p_{mk} = \frac{1}{n_m} \sum_{y \in Q_m} I(y = k)$$

частка спостережень класу k у вузлі m. Якщо є термінальною вершиною, то предикативна проба для цієї області дорівнює p_{mk}. Поширеними мірами домішки є наступні [39].

Gini:

$$H(Q_m) = \sum_k p_{mk}(1 - p_{mk})$$

Entropy:

$$H(Q_m) = -\sum_k p_{mk} \log p_{mk}$$

У цих формулах підсумовються всі класи, присутні у вузлі дерева рішень. Алгоритм дерева рішень намагатиметься максимізувати приріст інформації при кожному розбитті, що базується на зменшенні ентропії або домішки Джині після розбиття набору даних на атрибути [39].

Випадковий ліс (Random Forest) – це потужний метод машинного навчання, який використовує ансамбль дерев рішень для вирішення задач класифікації та регресії. Він працює шляхом створення множини дерев рішень під час тренування та виведення класу або середнього прогнозу від усіх дерев. Алгоритм випадкового лісу є розширенням методу мішків, оскільки він використовує як мішки, так і випадковість ознак для створення некорельованого лісу дерев рішень. Випадковість ознак, також відома як "пакування ознак" або "метод випадкового підпростору", генерує випадкову підмножину ознак, що забезпечує низьку кореляцію між деревами рішень. Це ключова відмінність між деревами рішень і випадковими лісами. В той час як дерева рішень розглядають всі можливі розбиття ознак, випадкові ліси вибирають лише підмножину цих ознак. Алгоритми випадкових лісів мають три основні параметри, які необхідно встановити перед початком навчання. До них відносяться розмір вузла,

кількість дерев і кількість вибіркових ознак. Після цього класифікатор випадкового лісу можна використовувати для розв’язання задач регресії або класифікації. Алгоритм випадкового лісу складається з набору дерев рішень, і кожне дерево в ансамблі складається з вибірки даних, взятої з навчального набору із заміною, яка називається бутстреп-вибіркою. З цієї навчальної вибірки третина відкладається як тестові дані, відомі як вибірка "поза пакетом" (out-of-bag, oob). Потім додається ще один екземпляр випадковості за допомогою пакування ознак, що додає більше різноманітності до набору даних і зменшує кореляцію між деревами рішень. Залежно від типу задачі, визначення прогнозу буде відрізнятися. Для задачі регресії окремі дерева рішень будуть усереднені, а для задачі класифікації передбачуваний клас буде визначатися більшістю голосів - тобто найбільш частою категоріальною змінною. Нарешті, вибірка oob використовується для перехресної перевірки, щоб остаточно підтвердити прогноз [40]. Загальний вигляд випадкового лісу можна побачити на рисунку 2.3.

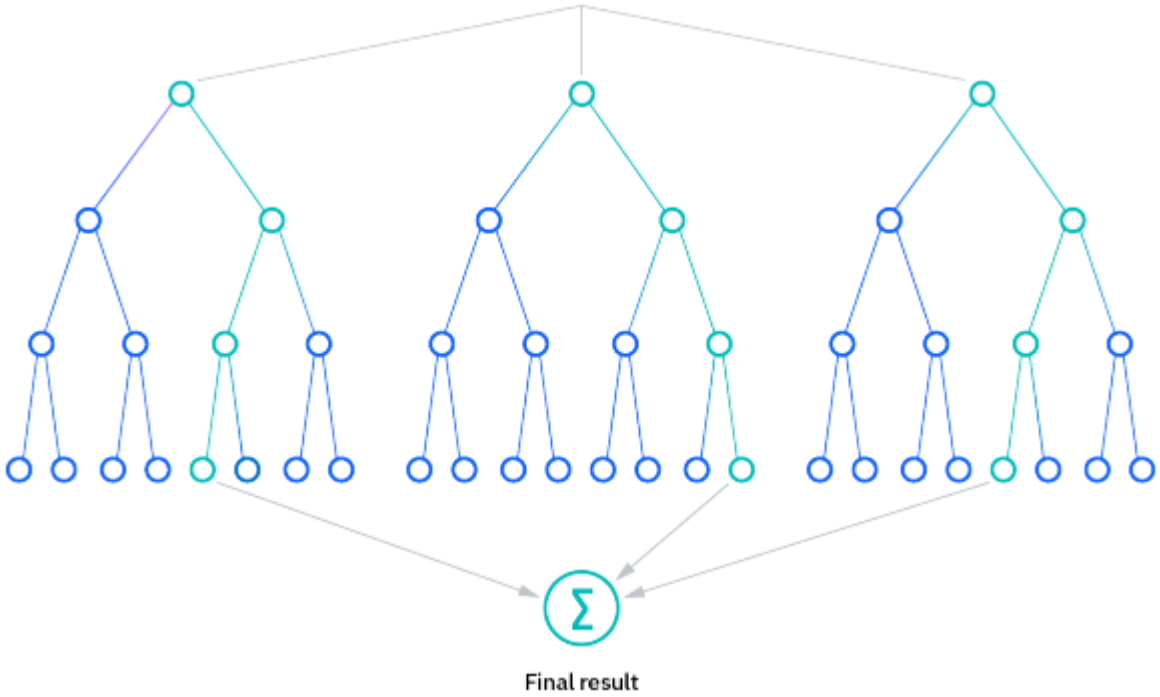


Рисунок 2.4 – Вигляд структури алгоритму роботи випадкового лісу

Джерело: What is random forest? [40]

3. МОДЕЛЮВАННЯ ІНТЕЛЕКТУАЛЬНОЇ ТЕХНОЛОГІЇ

3.1 Функціональне моделювання інтелектуальної технології в IDEF0

Діаграма IDEF0 є методологією, яка використовується для моделювання рішень, функцій та процесів і часто застосовується у системному аналізі та інженерії для зображення робочого процесу або системи. Основна мета цієї діаграми полягає в тому, щоб чітко визначити, що входить у систему (вхід), що виходить (вихід), які ресурси використовуються для досягнення результату (механізми) та які умови або правила впливають на процес (керування). На діаграмі IDEF0, кожен процес або функція зображується у вигляді блоку або "коробки", з якої ведуть стрілки, що представляють входи, виходи, механізми та керування. Ваш проект аналізу даних в електронній комерції включає збір даних з e-commerce платформи, їх обробку та аналіз з використанням різних аналітичних інструментів, і врешті-решт, формування інсайтів, прогнозів та візуалізацій, які можуть бути використані для підтримки бізнес-рішень [37]. Діаграму IDEF0 проілюстровано на рисунку 3.1.

Входи:

- Транзакційні дані: дані, які надходять від електронних транзакцій.
- Логи e-commerce додатку: інформація від активності користувачів на платформі e-commerce.

Виходи:

- Звіти: сформовані документи, які підсумовують результати аналізу.
- Інсайти: важливі відомості та знання, отримані з даних.
- Візуалізації: графічне представлення даних.
- Прогнози: передбачення майбутніх тенденцій на основі аналізу даних.

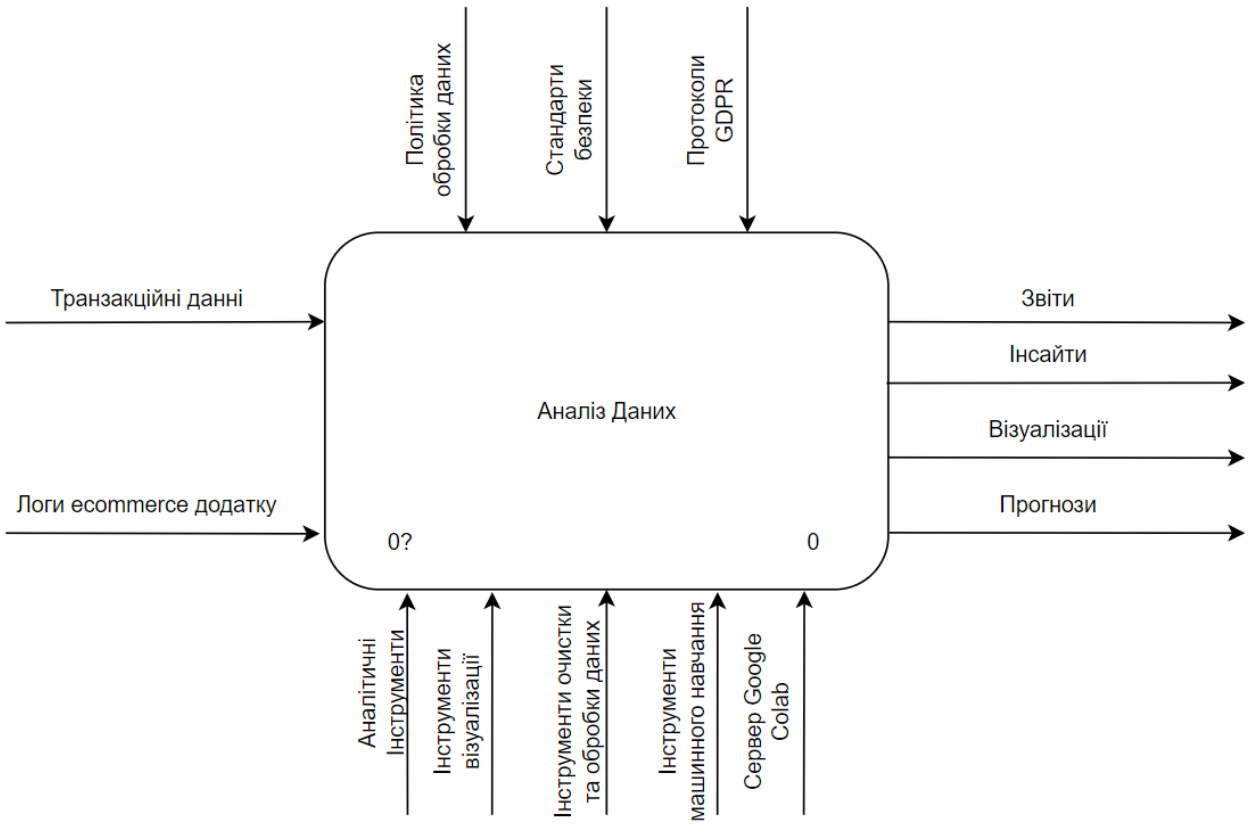


Рисунок 3.1 – IDEF0 для інформаційної технології аналізу даних користувачів e-commerce платформи

Джерело: розроблено автором

Механізми:

- Аналітичні інструменти: програми та інструменти, які використовуються для обробки та аналізу даних.
- Інструменти візуалізації: засоби для створення графічних зображень даних.
- Інструменти прогнозування: програмні рішення для створення прогнозів на основі аналізу.
- Інтеграція з машинним навчанням через Google Colab: використання хмарних сервісів для застосування моделей машинного навчання.

Керування:

- Політика обробки даних: набір правил та принципів, які керують збором та обробкою даних.

- Стандарти безпеки: вимоги, які забезпечують захист даних.
- Протоколи GDPR: регуляції, пов’язані з захистом персональних даних та приватності.

Для детального розгляду внутрішніх процесів системи було створено діаграму декомпозиції, яка розширює головний вузол IDEF0 діаграми. Ця діаграма декомпозиції виокремлює кожен із підпроцесів і деталізує їх внутрішню структуру. Зокрема, вона демонструє, як сирі дані перетворюються в структуровану інформацію через послідовність кроків, що включають їх збір, обробку, аналіз та врешті-решт використання аналітичних результатів для побудови стратегій бізнесу [38].

Такий підхід дозволяє зрозуміти як працює система зсередини та як це впливає на результати, які отримує бізнес. Кожна частина системи взаємодіє з іншими, утворюючи інтегрований потік даних, на основі якого будуються стратегії розвитку e-commerce. Як підсумок, поєднання частин системи сприяє підвищенню гнучкості й адаптивності системи аналізу даних, що є важливим для забезпечення потреб ринку електронної комерції. В кінцевому підсумку, ця інтеграція функціональних модулів забезпечує гнучкість та адаптивність системи аналізу даних у відповідь на постійно змінні потреби ринку електронної комерції. Діаграму наведено на рисунку 3.2.

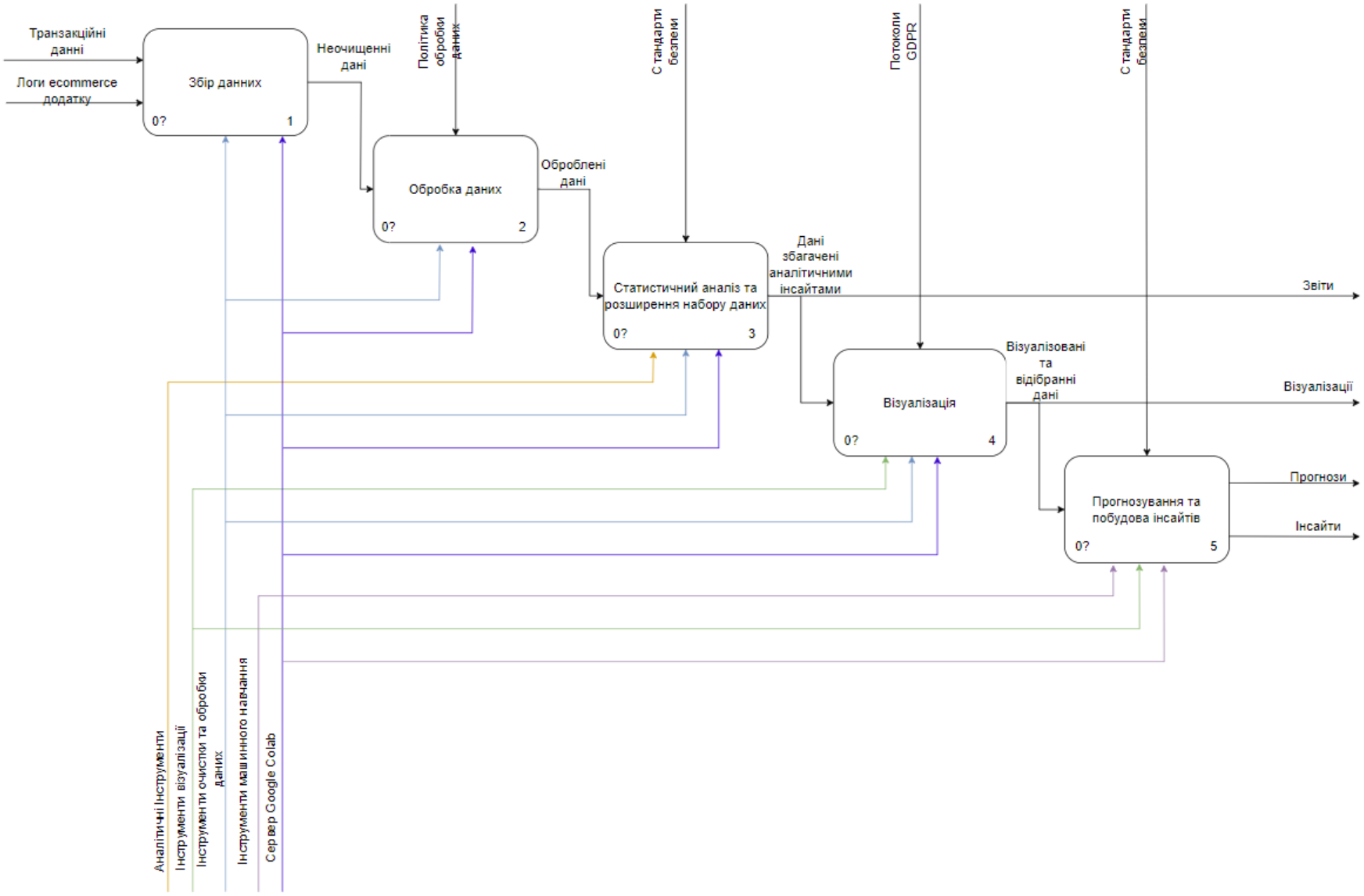


Рисунок 3.2 – Декомпозиція аналізу даних

Джерело: розроблено автором

Збір даних: це перший крок, де система збирає необроблені транзакційні дані з e-commerce платформи. Входи включають транзакційні дані та потік e-commerce додатків. Процес контролюється політикою збору даних і стандартами, з механізмами, що можуть включати засоби автоматизації.

Очищення даних: на цьому етапі дані фільтруються та очищаються для видалення помилок і дублікатів. Виходи з цього процесу — це очищені та чисті дані, які готові до подальшого аналізу.

Аналіз та розширення даних: тут відбувається статистичний аналіз і розширення набору даних через додавання проінформованих інсайтів. Аналітичні результати з цього процесу можуть включати звіти і прогнози.

Прогнозування та побудова інсайтів: використання аналітичних даних для побудови інсайтів та розвитку стратегій на основі аналізу. Виходами є інсайти та прогнози, які можуть бути використані для інформування бізнес-рішень.

Формування висновків та стратегій: кінцевий етап, де аналіз та прогнози використовуються для формування бізнес-стратегій та висновків, які направлятимуть майбутні дії e-commerce платформи.

Кожен з цих вузлів взаємодіє з іншими через входи і виходи, що дозволяє інформації текти через систему аналізу даних і забезпечує послідовність процесів від збору даних до реалізації бізнес-стратегій.

3.2 Проектування інформаційної системи

Діаграма варіантів використання (Use Case Diagram) є стандартним інструментом в UML (Unified Modeling Language) для визначення функцій системи з точки зору її кінцевих користувачів. Діаграма зображає основні взаємодії між акторами та системою, ілюструючи, як користувачі можуть використовувати систему для досягнення конкретних цілей [39].

На представленій діаграмі варіантів використання в контексті аналізу даних на e-commerce платформі визначені основні сценарії взаємодії з системою. Діаграма варіантів використання представлена на рисунку 3.3.

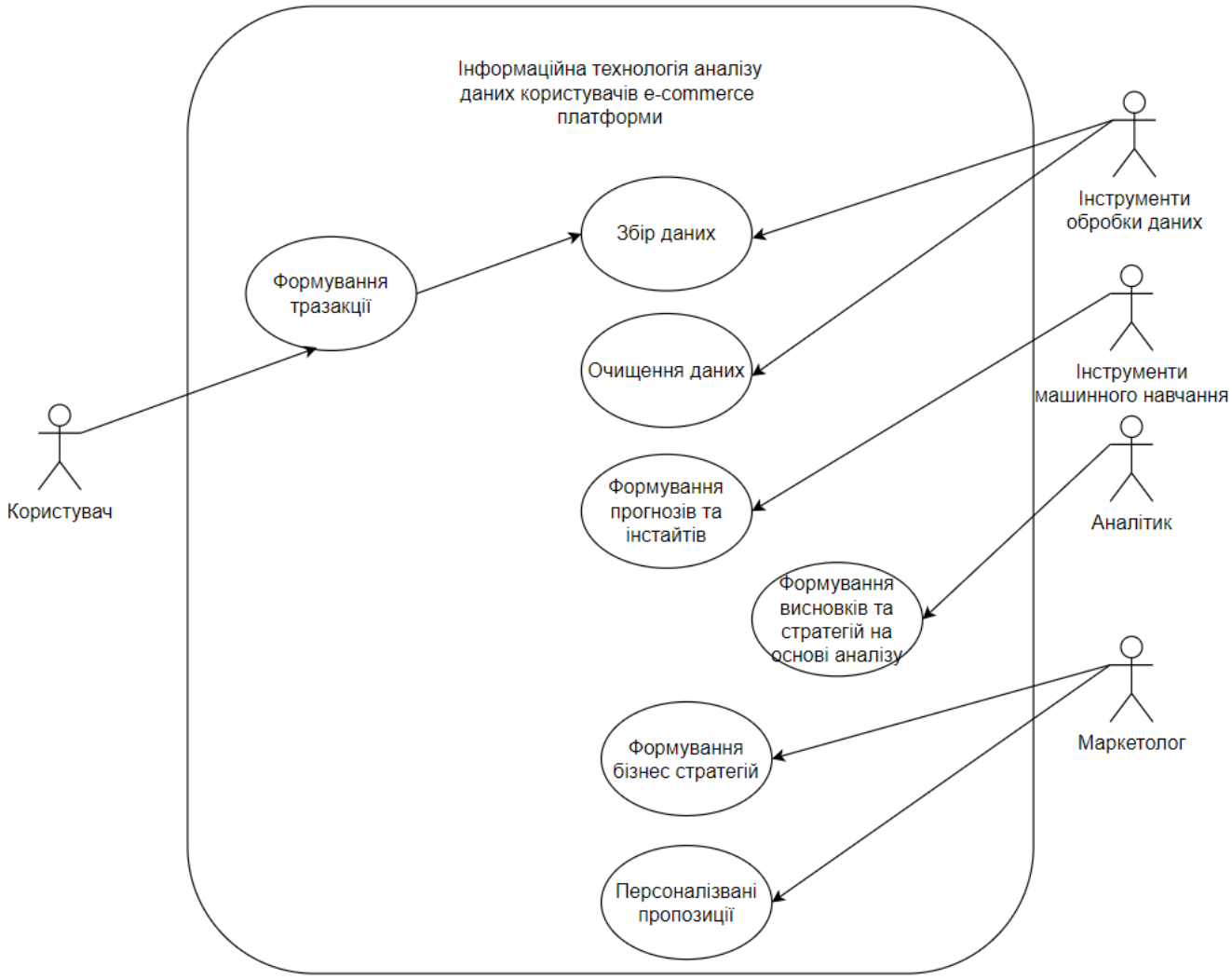


Рисунок 3.3 – Діаграма варіантів використання

Джерело: розроблено автором

Формування транзакцій: користувачі виконують покупки, результатом яких є транзакційні дані, що надходять до системи.

Збір та очищення даних: система збирає сирі дані, які підлягають процесам очищення для подальшої обробки.

Формування прогнозів та інсайтів: аналітики використовують систему для створення прогнозів на основі аналізу даних, виявлення трендів та інсайтів, які можуть вплинути на стратегічні рішення.

Формування бізнес стратегій: маркетологи використовують інформацію та аналітику для розробки та адаптації бізнес стратегій.

Персоналізація пропозицій: система дозволяє користувачам отримувати персоналізовані пропозиції на основі їхньої поведінки та переваг.

Інформаційна технологія аналізу даних користувачів e-commerce платформи є центральним елементом, навколо якого організовані всі взаємодії. Актори на діаграмі включають не тільки користувачів та аналітиків, але й маркетологів, кожен з яких використовує систему відповідно до своїх ролей і потреб.

Діаграма потоків даних (Data Flow Diagram, DFD) ілюструє потік даних між різними процесами та суб'єктами в системі. Вона використовується для моделювання інформаційної системи з акцентом на переміщення даних, їх обробку та висновки, які з цього випливають [40].

Рисунок 3.4 – Діаграма потоків даних DFD нульового рівня

Джерело: розроблено автором

На представленій DFD діаграмі для e-commerce платформи основні компоненти та їх взаємодія включають:

Користувачі: Це вихідний пункт, де здійснюються всі транзакції користувачів та збираються дані, які потрібні для аналізу.

Технологія обробки даних користувачів: Це центральний процес, де відбувається основна обробка даних. Він приймає вхідні дані від користувачів e-commerce платформи і перетворює їх у цінні відкриття.

Аналітик: Занурюється в аналітичні результати, щоб зрозуміти поведінку користувачів та тенденції ринку, забезпечуючи бізнес цінною інформацією для стратегічного планування.

DFD дозволяє візуалізувати, як дані потрапляють у систему, обробляються в ній і які результати це дає. Ця конкретна діаграма допомагає розумінню взаємодії між елементами системи і може використовуватися для поліпшення процесів та розробки нових функцій платформи.

Деталізований Data Flow Diagram (DFD) для процесу аналізу даних на e-commerce платформі (рис. 3.2) ілюструючи як інформація та команди переміщаються в системі. Вона підкреслює важливість кожного етапу обробки даних і як це впливає на прийняття рішень у бізнесі.

Збір даних: первинний етап, де збираються транзакційні дані від користувачів e-commerce платформи.

Обробка даних: дані очищаються та готуються для детального аналізу. Даний процес керується аналітиком для перевірки отриманих результатів процесу після оброки даних.

Аналіз та розширення даних: виконується додаткова обробка, включаючи статистичний аналіз та групування для визначення патернів та тенденцій.

Прогнозування та побудова інсайтів: на основі аналізованих даних створюються прогнози, які можуть бути використані для стратегічного планування.

Формування висновків та стратегій: результати аналізу використовуються для формування бізнес-стратегій.

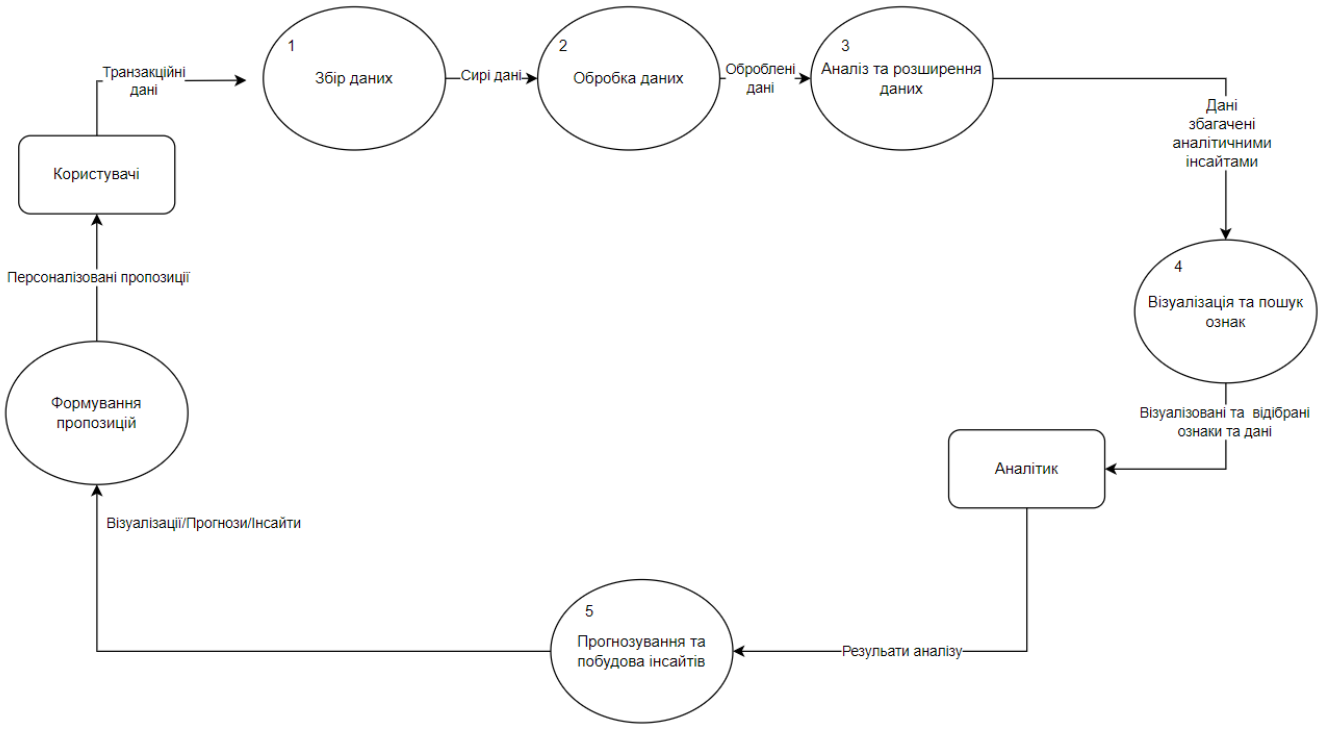


Рисунок 3.5 – Деталізована діаграма потоків даних DFD

Джерело: розроблено автором

Побудова стратегій покращення бізнес процесів: останній етап, де стратегії реалізуються для поліпшення загальної діяльності бізнесу.

Кожен етап взаємодіє з іншими, формуючи циклічний процес, що дозволяє систематично покращувати бізнес-операції на основі аналізу даних. Ключовими учасниками є маркетологи, які використовують інсайти для покращення маркетингових кампаній, та аналітики, які виконують глибинний аналіз даних для виявлення прихованих можливостей.

4. РЕАЛІЗАЦІЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ АНАЛІЗУ ДАНИХ КОРИСТУВАЧІВ

4.1 Опис вхідних даних

Вигляд набору даних перед обробкою. Загально це набір транзакційних даних. Цей набір містить 5 ознак які представлено заголовками колонок таблиці. Кожен рядок таблиці містить дані які характеризують одну унікальну покупку в рамках однієї загальної покупки. На рисунку 4.1 можна побачити вигляд набору даних.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

Рисунок 4.1 – Структура набору трансакційних даних користувачів e-commerce платформи

Джерело: розроблено автором

Перед тим як почати роботу важливо зрозуміти загальні параметри обраного набору даних. Цей набір містить 541909 рядків таблиці. Також кожна ознака має свій тип. Відомості про набір даних можна побачити на рисунку 4.2.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  object
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: float64(2), int64(1), object(5)
```

Рисунок 4.2 – Структура набору даних

Джерело: розроблено автором

4.2 Реалізація ідентифікації аномалій та їх обробка

Початковою метою є отримання загальних відомостей про набір даних. Це простий статистичний прогін який показує кількість середнє значення, мінімальні та максимальні значення в наборі. Результат перевірки ознак кількості та ціна показано на рисунку 4.3.

```
            Quantity      UnitPrice
count  535187.000000  535187.000000
mean        9.671593       4.645242
std       219.059056      97.364810
min    -80995.000000  -11062.060000
25%         1.000000       1.250000
50%         3.000000       2.080000
75%        10.000000       4.130000
max     80995.000000   38970.000000)
```

Рисунок 4.3 – Перевірка ознак ціни та кількості в наборі даних

Джерело: розроблено автором

Вже на цьому етапі помітно що існують якісь незвичайні дані що описують ціну та кількість. Тобто помітно що набір містить негативну кількість а ціну. Для

подальшого розуміння важливо провести додатковий підрахунок таких негативних значень. Результат підрахунку негативних значень представлено на рисунку 4.4.

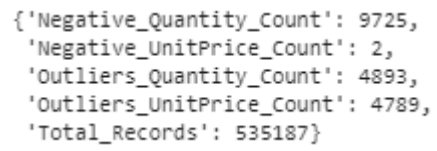


Рисунок 4.4 – Результати пошуку негативних значень по ціні та кількості

Джерело: розроблено автором

Опис знаходжень аномалій:

- Негативна кількість (Negative Quantity): Було знайдено 9,725 рядків що містять негативну кількість.
- Негативна ціна (Negative Unit Price): Було знайдено усього 2 з негативною ціною.
- Додаткові аномалії по кількіості (Outliers in Quantity): Було знайдено 4,893 рядків у яких кількість екстремально висока (понад 99 перцентиль).
- Додаткові аномалії по ціні (Outliers in Unit Price): Було знайдено 4,789 рядків з незвично високою ціною (понад 99 перцентиль).

На цьому етапі можна зробити проміжні висновки. По-перше існує доволі велика кількість рядків з негативною кількістю. Є явна потреба у перевірці таких рядків для розуміння походження негативних чисел. По-друге негативна ціна явно схожа на якісь помилкові дані адже всього дві негативні ціни за товар. По-третє є певні незбалансованості у наборі даних так як є доволі не мала кількість екстримально високих цін та кількості. На цьому етапі можна зробити наступне припущення – негативна ціна може характеризувати повернення товару. Тому з такою гіпотезою можна будувати подальший аналіз.

Тепер можна заглибитись в цю аномалію та зрозуміти що за товари так часто повертають та хто ці клієнти що їх повертають. На рисунках 4.5 – 4.6 зображено гістограму топ 10 товарів які повертають.

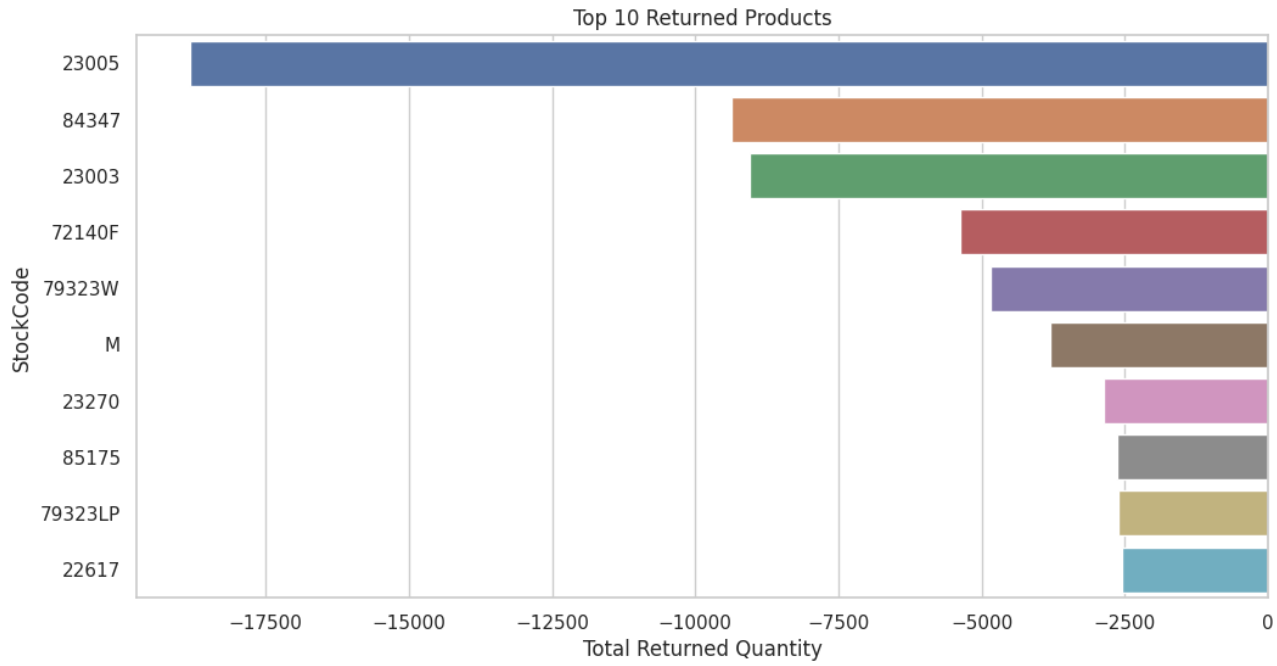


Рисунок 4.5 – Гістограма топ 10 товарів, які повертають

Джерело: розроблено автором

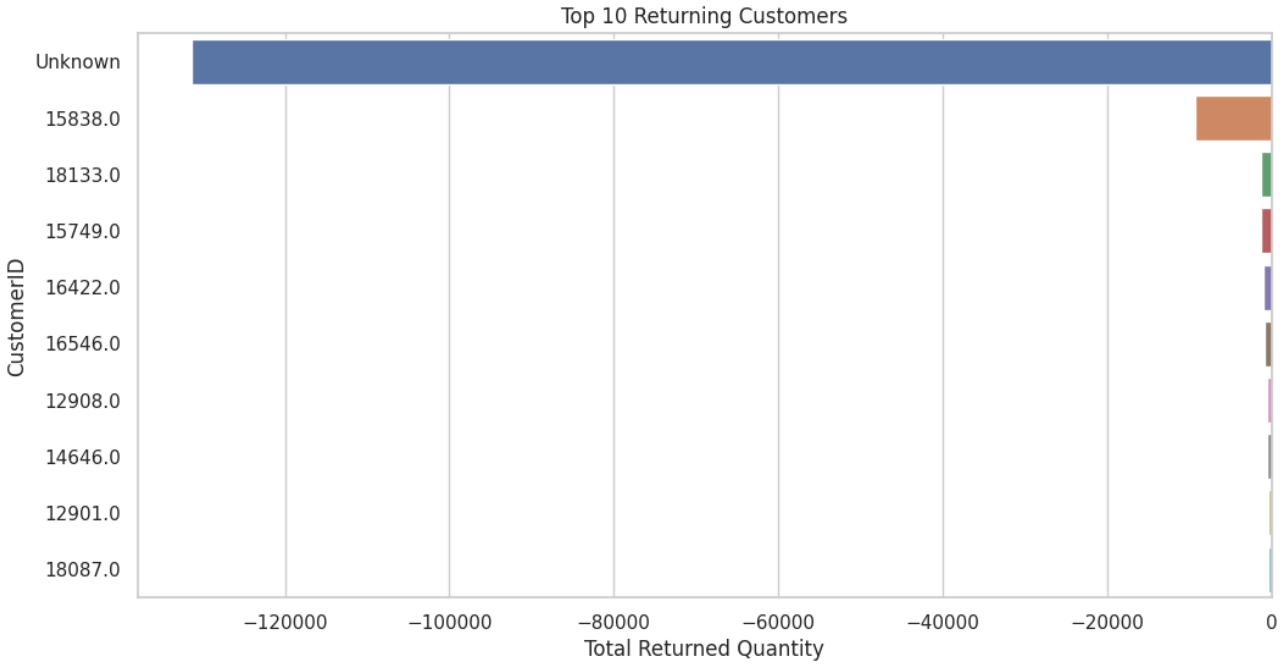


Рисунок 4.6 – Гістограма топ 10 клієнтів, які повертають товари

Джерело: розроблено автором

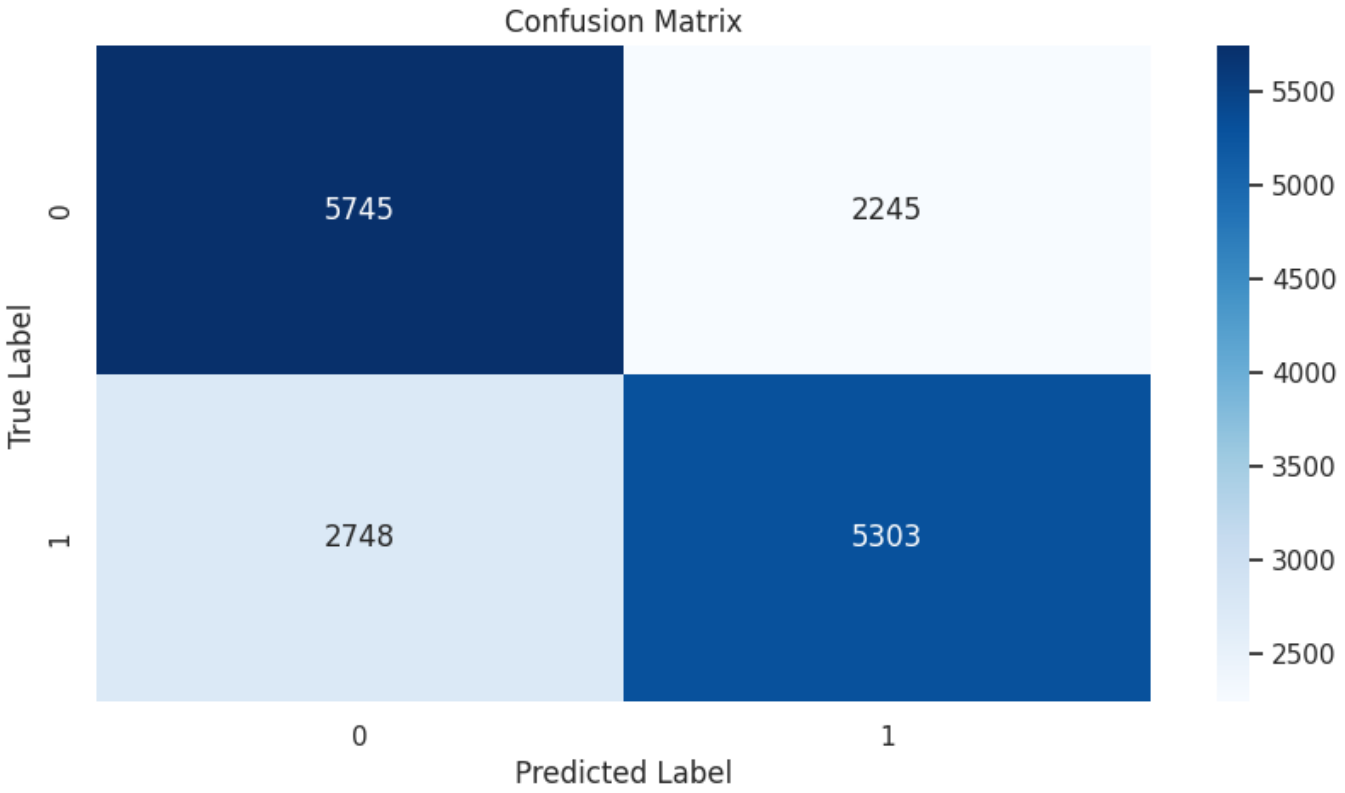


Рисунок 4.7 – Матриця точності моделі

Джерело: розроблено автором

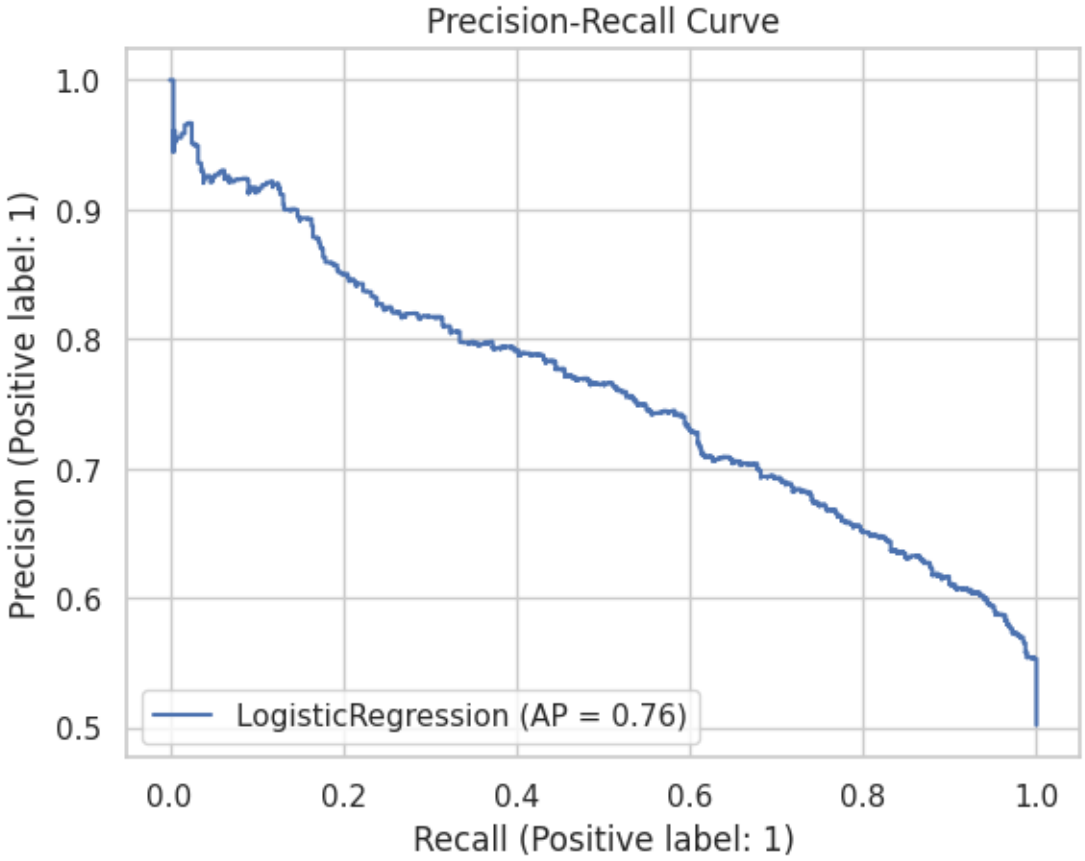


Рисунок 4.8 – Графік кривої точності відтворення

Джерело: розроблено автором

На основі регресійної моделі було проведено додатковий аналіз для визначення ознак повернення, це дає фактичні коефіцієнти на основі яких можна зрозуміти що це повернення. Розуміння того, які особливості є найбільш впливовими у прогнозуванні повернення, може дати цінну інформацію. Таким чином можна поглянути на коефіцієнти логістичної регресійної моделі, щоб побачити, які характеристики (наприклад, конкретні продукти, ціни або час покупки) найбільше пов’язані з вищою або нижчою ймовірністю повернення. На рисунку 4.9 зображені топ 10 ознак які вплинули на прогнозування повернення регресійної моделі.

	Feature	Coefficient	Absolute Coefficient
10	StockCode_22423	16.499578	16.499578
13	StockCode_22720	4.636924	4.636924
3	StockCode_21212	-3.596326	3.596326
15	StockCode_47566	2.968232	2.968232
17	StockCode_84879	-2.223053	2.223053
6	StockCode_22197	-2.113117	2.113117
12	StockCode_22469	-1.990587	1.990587
7	StockCode_22383	-1.521289	1.521289
20	UnitPrice	-1.516213	1.516213
1	StockCode_20727	-1.499831	1.499831

Рисунок 4.9 – Результат пошуку коефіцієнтів впливу на регресійну модель прогнозування повернень

Джерело: розроблено автором

Як можна побачити з рисунку 4.9 основний вплив на повернення товару є саме товар з номером 22423. Подальше його вивчення може привезти до загального розуміння поведінки повернення. Наступними кроками для вивчення цього товару будуть: базова статистка транзакцій, часовий аналіз та сегментація покупців. Результати проведеного аналізу можна побачити на рисунку 4.10.

```
YearMonth
2010-12    13
2011-01    18
2011-02    14
2011-03    28
2011-04    15
2011-05    10
2011-06    10
2011-07     8
2011-08     6
2011-09     7
2011-10     5
2011-11    12
2011-12     4
Freq: M, dtype: int64,
            Total_Purchases  Return_Count  Purchase_Count
CustomerID
13113.0                  48             7               4
15465.0                 317             6              14
17865.0                  87             5               7
14299.0                  -1             5               4
13089.0                 220             5              12
17511.0                  26             4               2
15189.0                 278             4              14
15482.0                  38             3               5
Unknown                1375             3             285
13767.0                  97             3              17)
```

Рисунок 4.10 – Результати додаткової аналітики по товару за номером 22423

Джерело: розроблено автором

Базова статистика транзакцій:

- Товар має широкий діапазон кількості за одну транзакцію, від -150 (повернення) до 272 (покупки).
- Середня ціна за одиницю товару становить близько 14,07, з деякими коливаннями.
- Близько 8% транзакцій з цим продуктом є поверненнями.

Часовий аналіз:

- Кількість повернень на місяць варіюється, причому в деякі місяці, такі як березень 2011 року та січень 2011 року, кількість повернень була вищою.
- На цю тимчасову закономірність можуть впливати такі фактори, як сезонний попит, рекламні акції або проблеми з запасами.

Сегментація покупців для ‘StockCode_22423’:

- Деякі покупці мають вищу частоту повернень для цього товару. Наприклад, CustomerID 13113.0 повертав товар 7 разів.
- Дані також показують суміш клієнтів, які часто купують товар, але також повертають його кілька разів.

Час між покупкою та поверненням

Підрахунок: Виявлено 621 парну транзакцію купівлі та подальшого повернення.

Середній час: В середньому між покупкою та поверненням цього товару проходить 78,8 днів.

Стандартне відхилення: Існує значна варіабельність (72,3 дні) у часі між покупкою та поверненням.

Діапазон: Розрив у часі коливається від одного дня (0 днів) до майже року (343 дні).

Покупці, які часто купують і повертають товари. Покупець, позначений як "Невідомий" (ймовірно, це незареєстрований або гостьовий покупець), демонструє найвищу частоту повторних покупок. Інші топ-покупці (наприклад, з ідентифікаторами 15465.0, 13767.0) також демонструють часті повторні покупки.

Міжпродуктовий аналіз наступний етап у вивченні цієї аномалії. Мета полягає в тому, щоб зрозуміти, чи моделі повернення, які спостерігаються для ‘StockCode_22423’, є унікальними для цього товару, або чи існують подібні моделі для інших товарів. Цей аналіз може допомогти виявити ширші проблеми, пов’язані з поверненнями, такі як обслуговування клієнтів, якість продукції або цінові стратегії.

Додатковим вивченням цієї аномалії буде пошук інших товарів зі схожою частотою повернення, топ 5 товарів представлено на рисунку 4.11.

StockCode	Total_Transactions	Total_Returns	Return_Rate
22423	1852	150	0.080994
22960	1070	70	0.065421
D	74	74	1.000000
M	473	196	0.414376
POST	140	88	0.628571

Рисунок 4.11 – Схожі товари до аномального, які також мають високу частоту повернення

Джерело: розроблено автором

На цьому етапі вже багато стало зрозуміло про цей аномальний товар, але для більш детального та обґрунтованого результату краще отримати додаткову інформацію таку як детальний опис відгуки та можливо додаткові характеристики. Таким чином можна буде більш детально зрозуміти такі анормальні повернення та сприяти покращенню продажів або логістики щодо цього та інших схожих товарів.

4.3 Реалізація загального аналізу клієнтської поведінки

Наступним етапом буде загальне вивчення поведінки клієнтів на основі частоти та обсягу покупок, середньої вартості транзакції та визначення ключових сегментів клієнтів. Це дасть чітке уявлення про купівельну поведінку різних груп клієнтів. Топ 10 результатів представлено на рисунку 4.12.

CustomerID	Frequency	Total_Quantity	Average_Transaction_Value
Unknown	1545	431745	514.680421
12748.0	210	25288	56.134810
17841.0	124	22834	157.029677
13089.0	97	31025	51.202577
14606.0	93	6187	81.023548
15311.0	91	38147	65.266044
12971.0	86	9289	8.193605
16029.0	63	40108	138.423016
13408.0	62	16232	18.804677
18102.0	60	64124	32.348667

Рисунок 4.12 – Загальне вивчення купівельної поведінки клієнтів

Джерело: розроблено автором

Було відібрано топ 3 клієнта з вибірки та зроблено проміжні підсумки.

Клієнт "Невідомий":

- Частота покупок: 1545 транзакцій
- Загальна кількість придбаних товарів 431 745 товарів
- Середня вартість транзакції: 514.68

Клієнт 12748.0

- Частота покупок: 210 транзакцій
- Загальна кількість придбаних товарів: 25 288 одиниць
- Середня вартість транзакції: 56.13

Клієнт 17841.0:

- Частота покупок: 124 транзакції
- Загальна кількість придбаних товарів: 22 834 одиниці
- Середня вартість транзакції: 157.03

Далі було проведено візуалізації для загального розуміння частоти покупок по кожному клієнту у список потрапили топ 10. Гістограма показана на рисунку 4.13.

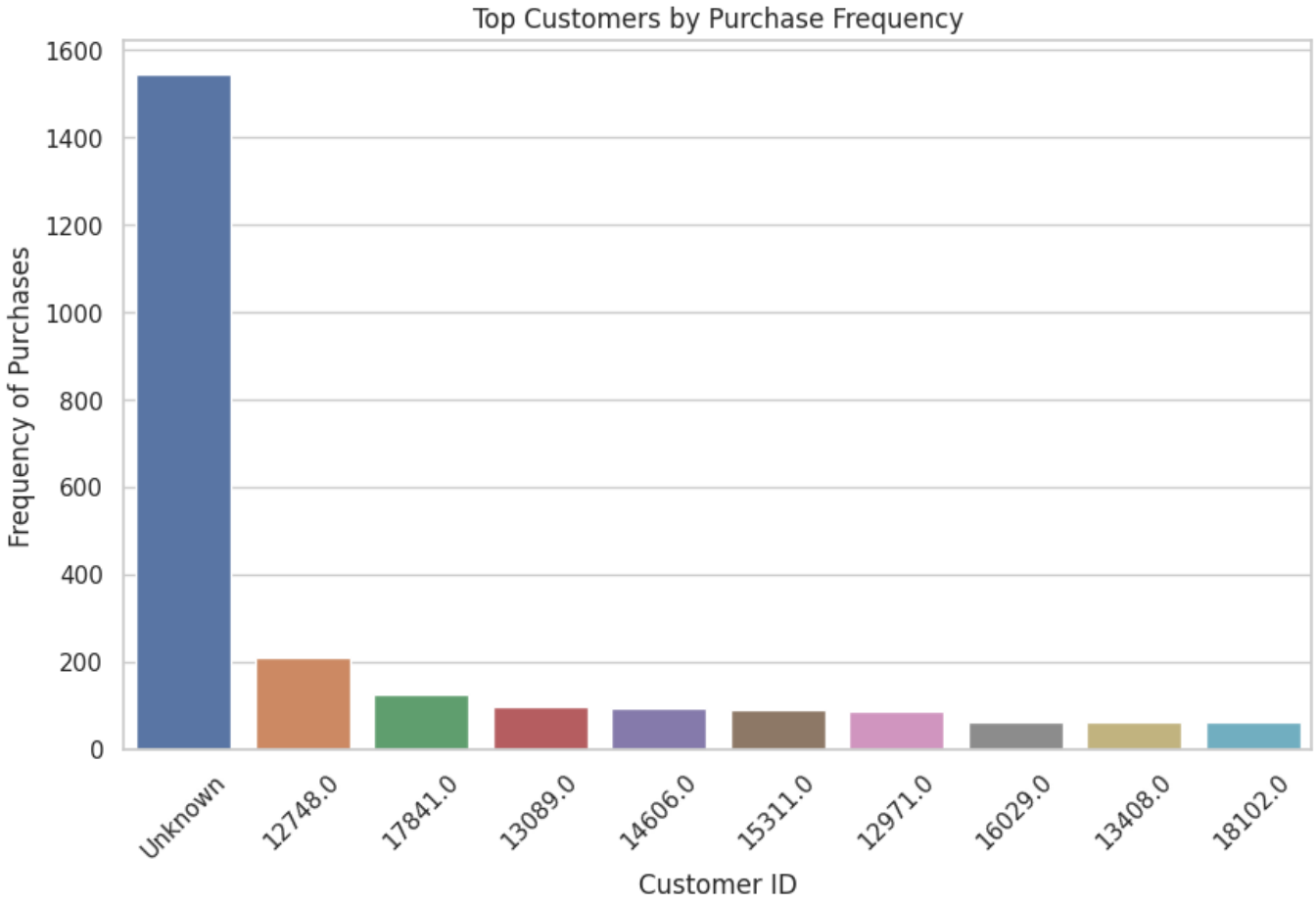


Рисунок 4.13 – Розподіл частоти покупок на топ 10 клієнтів платформи

Джерело: розроблено автором

Діаграма розсіювання відображає зв’язок між загальною кількістю придбаних товарів і середньою вартістю транзакції для кожного покупця.

Ця візуалізація допомагає зрозуміти, як обсяг покупок співвідноситься з середніми витратами на транзакцію. Це корисно для виявлення закономірностей, наприклад, чи клієнти, які купують більше, в середньому витрачають більше. Вигляд діаграми можна побачити на рисунку 4.14.

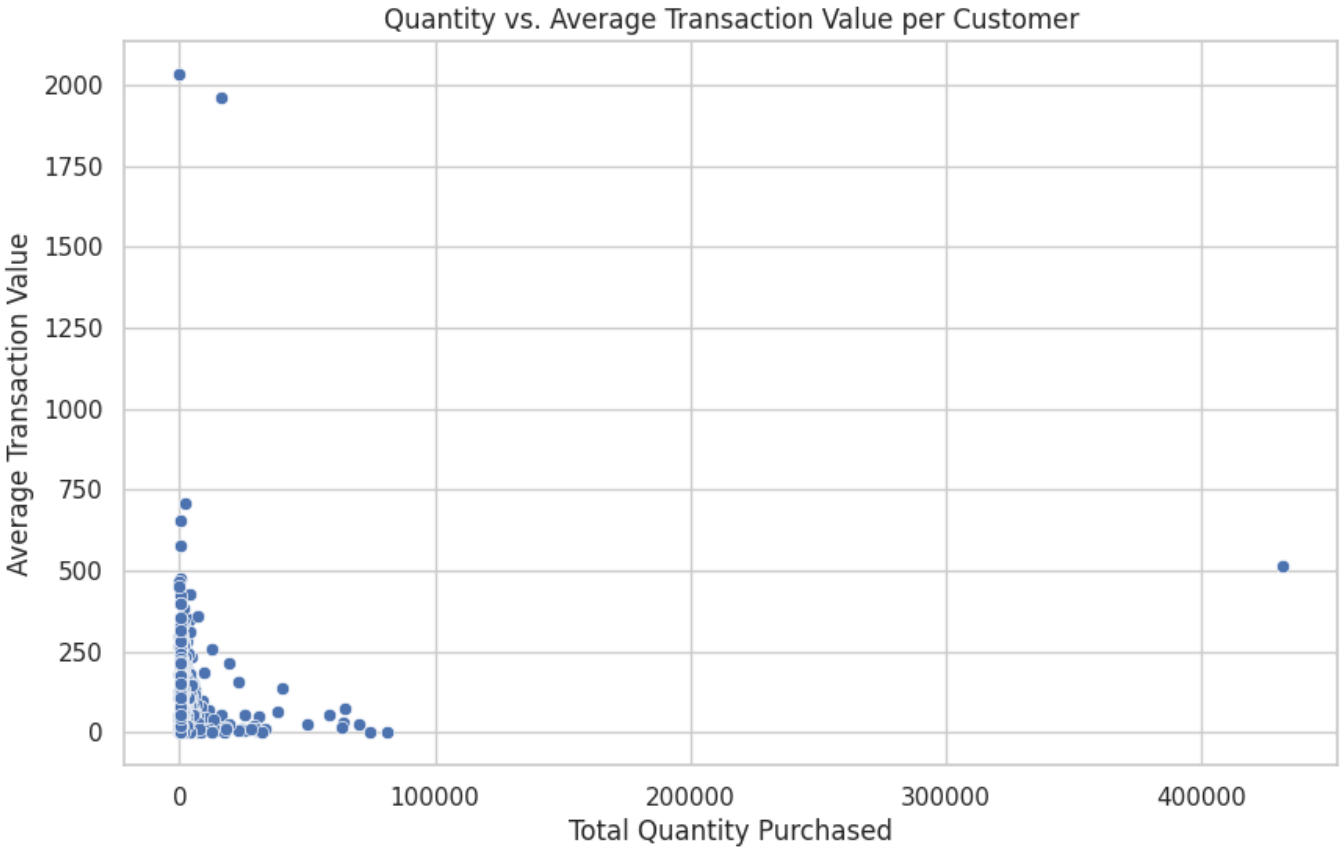


Рисунок 4.14 – Діаграма розсіювання між середнім вартістю транзакції та загально. Кількістю придбаних товарів

Джерело: розроблено автором

Наступним етапом є детальний аналізу високочастотних клієнтів, оскільки вони часто становлять основний сегмент клієнтської бази. Було розглянуто їхні моделі покупок, вподобання та будь-які інші доступні дані, щоб отримати уявлення про цей сегмент. Результати аналізу можна побачити на рисунках 4.15 – 4.16.

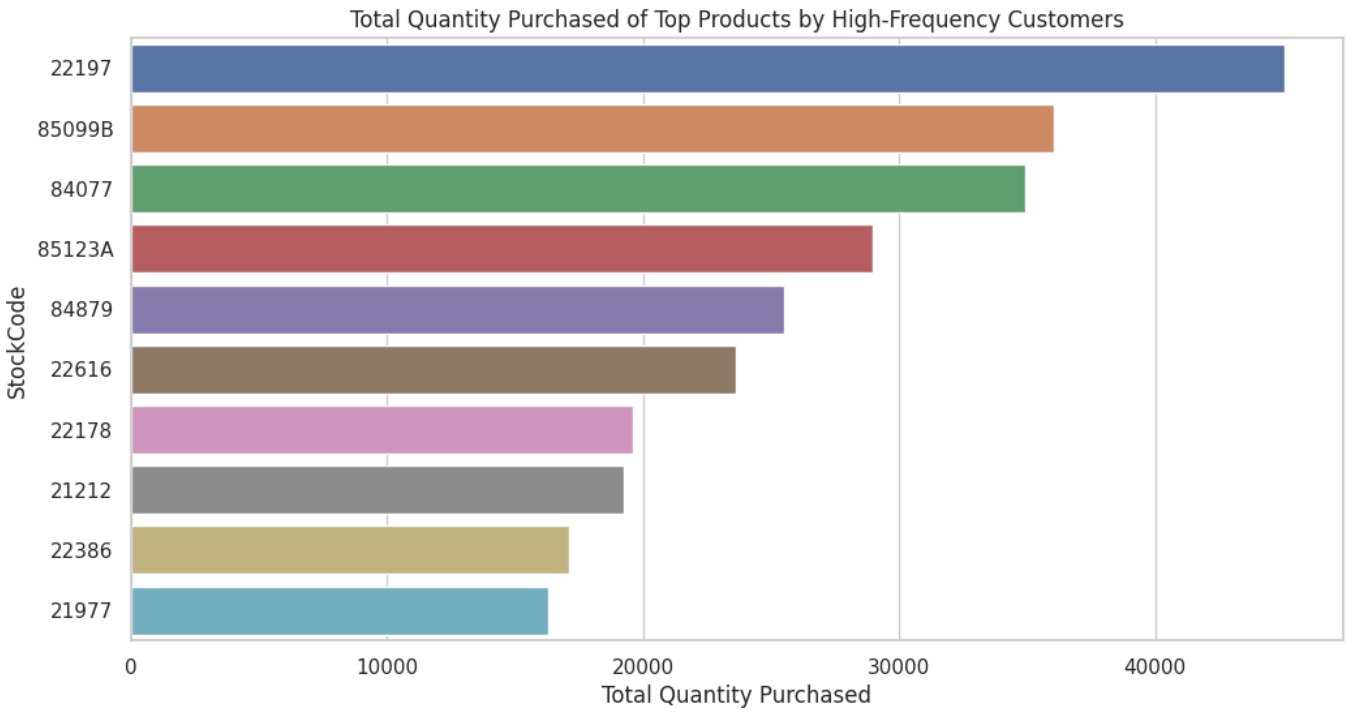


Рисунок 4.15 – Гістограма кількості покупок продуктів сегментом клієнтів, які купують з високою частотою

Джерело: розроблено автором

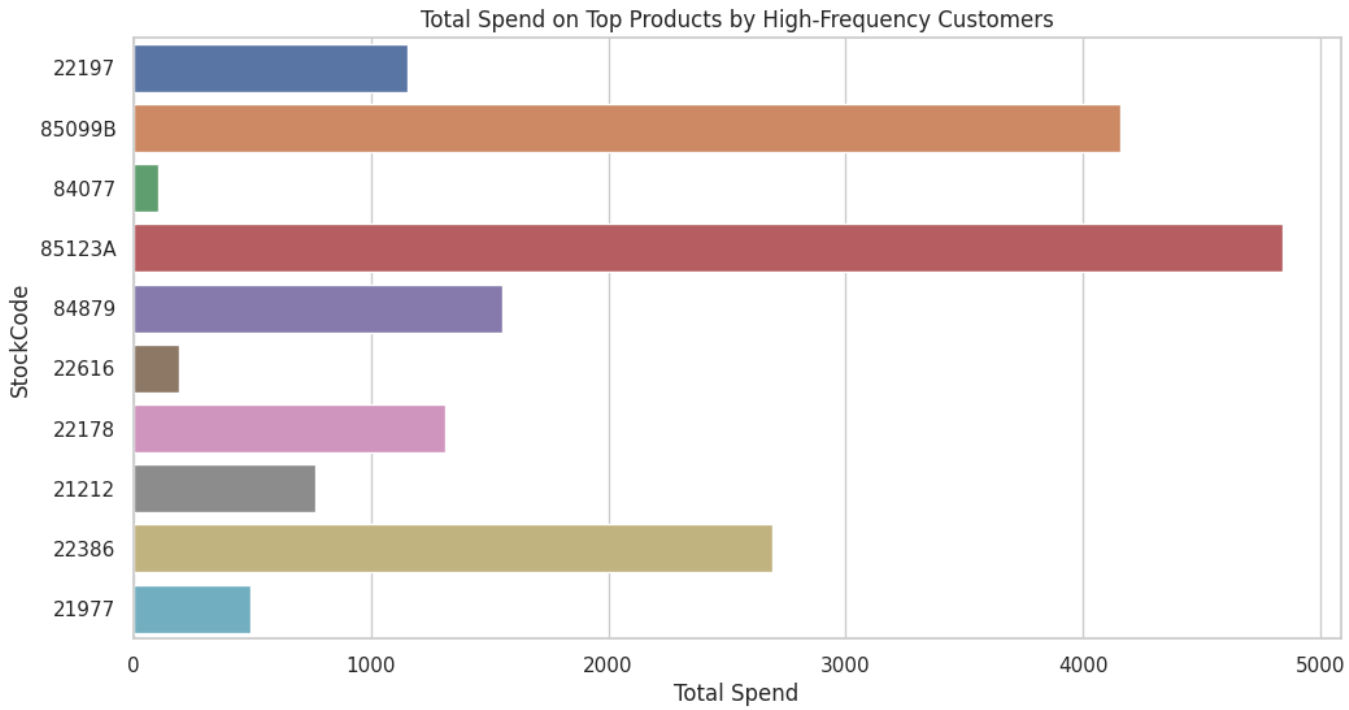


Рисунок 4.16 – Загальна кількість витрат на топ продукти високочастотним сегментом клієнтів

Джерело: розроблено автором

Високочастотні клієнти, схоже, надають перевагу певним товарам, на що вказує велика кількість придбаних товарів. Це можуть бути товари першої необхідності або популярні товари. Загальна сума витрат на ці товари варіюється, що свідчить про відмінності в цінах або купівельній спроможності клієнтів.

Вивчення інших клієнтських сегментів, зокрема, великих і цінних клієнтів, може дати цінну інформацію про різні аспекти поведінки та вподобань клієнтів.

Постійні клієнти – які купують велику кількість товарів, незалежно від частоти або загальної суми витрат. Було визначено таких клієнтів і проаналізовано їхні купівельні моделі, щоб зрозуміти, що спонукає їх до великих покупок.

Високоцінні клієнти – мають високу середню вартість транзакції, що свідчить про те, що вони або купують дорогі товари, або витрачають багато за одну транзакцію. Візуалізації аналізу можна побачити на рисунках 4.17-4.18.

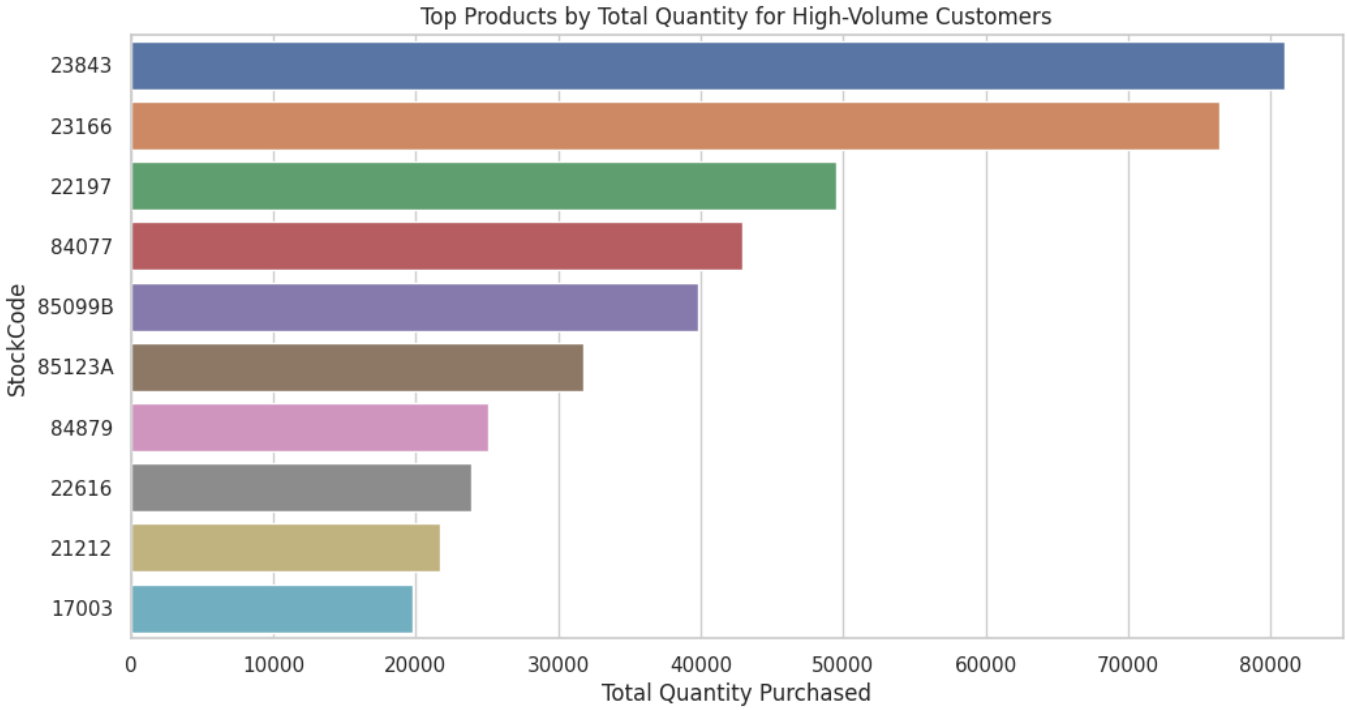


Рисунок 4.17 – Гістограма загальної кількості куплених товарів серед сегменту клієнтів, які купують у великих об’ємах

Джерело: розроблено автором

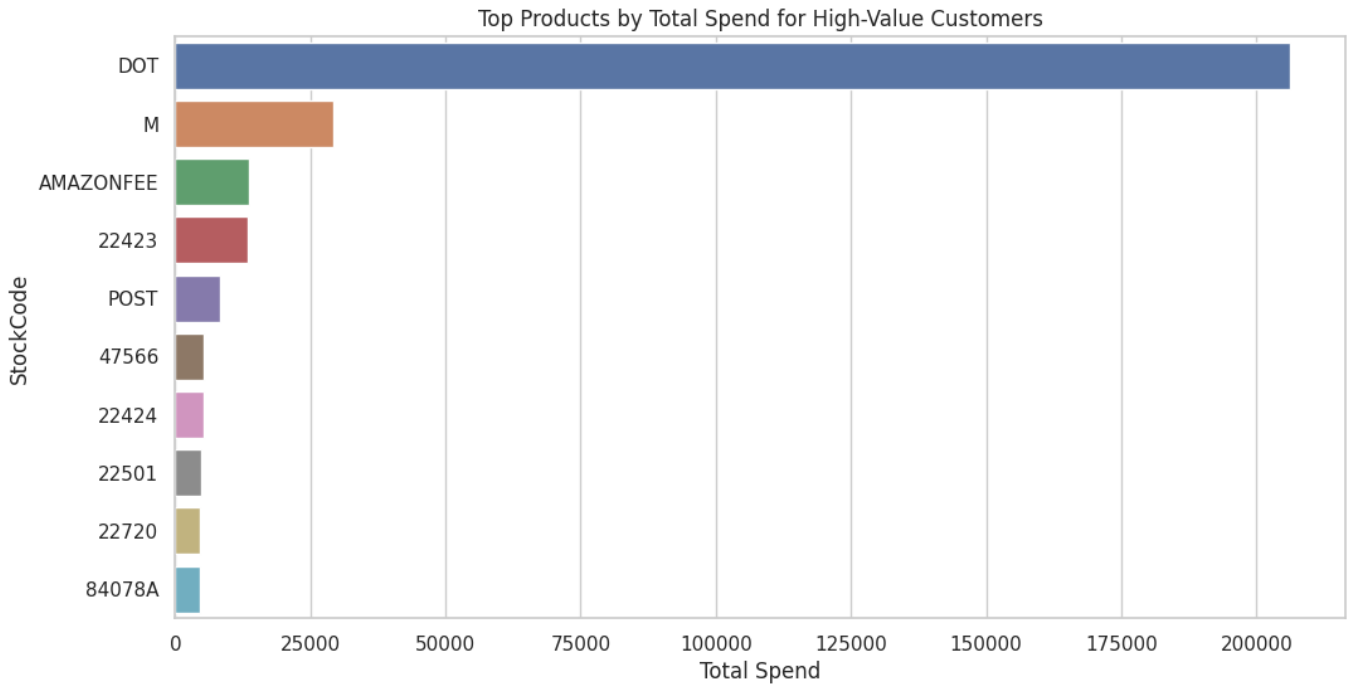


Рисунок 4.18 – Гістограма загальної ціни куплених товарів серед сегменту клієнтів, які купують за високими цінами

Джерело: розроблено автором

Оптові клієнти – купують великі партії товарів, часто за нижчими цінами. Це може свідчити про оптову купівельну поведінку або надання переваги певним товарам першої необхідності.

Цінні клієнти – їхні покупки більше зосереджені на загальній сумі витрат, часто включають менше, але дорожчі товари або більші транзакції.

Теплова карта – візуалізація розподілу покупок за часом може дати більш чітке уявлення про пікові моменти покупок. Вона представлена на рисунку 4.19.

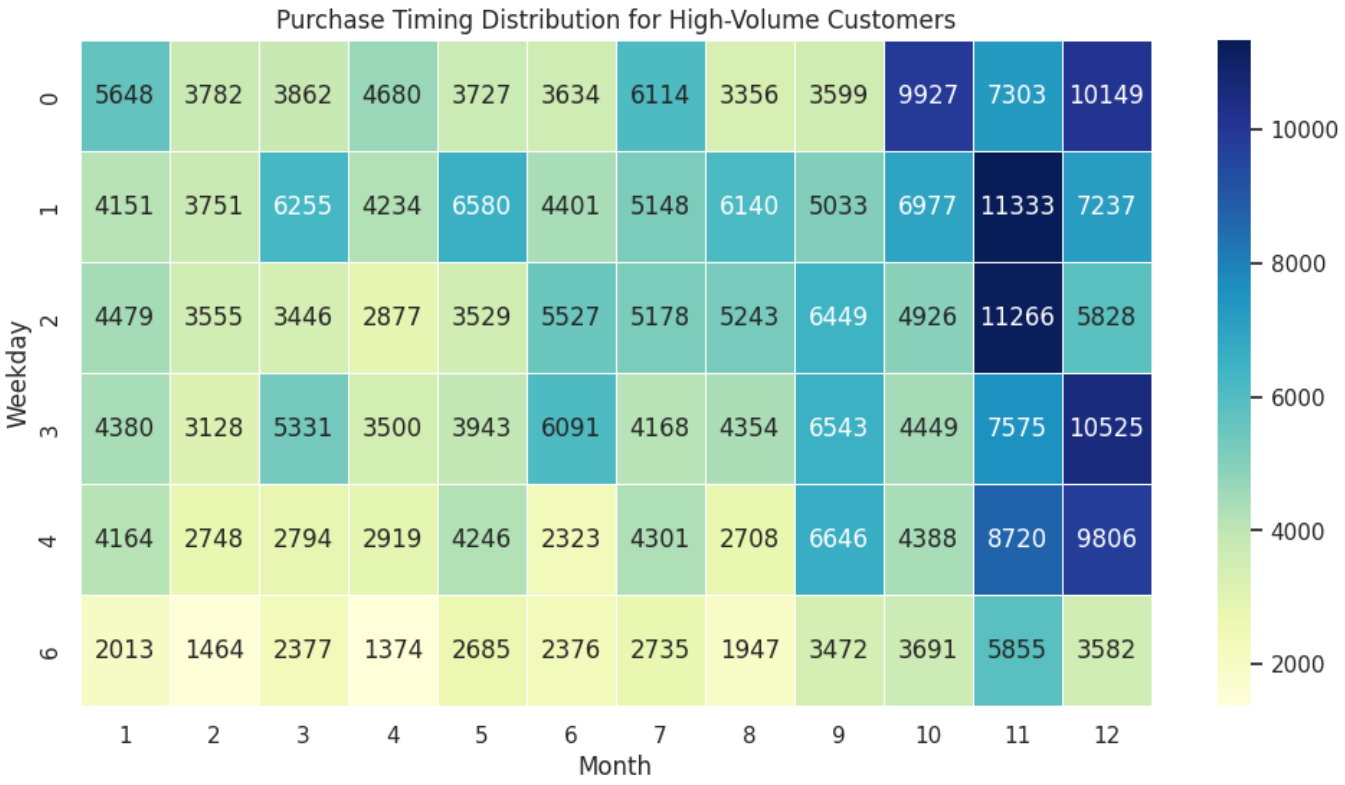


Рисунок 4.19 – Теплова карта розподілу покупок по дням місяця, для клієнтів, які закупають оптом

Джерело: розроблено автором

Теплова карта дозволяє виявити закономірності в часі покупок, наприклад, певні дні тижня або місяці, коли покупки відбуваються частіше.

Такі закономірності можуть бути пов’язані зі звичками покупців, сезонними тенденціями або реакцією на рекламні акції та маркетингові заходи.

Для просторового розуміння трьох базисних ознак (середнє значення транзакції, частота покупки, загальна кількість покупки) було побудовано просторову візуалізацію, яку можна побачити на рисунку 4.20.

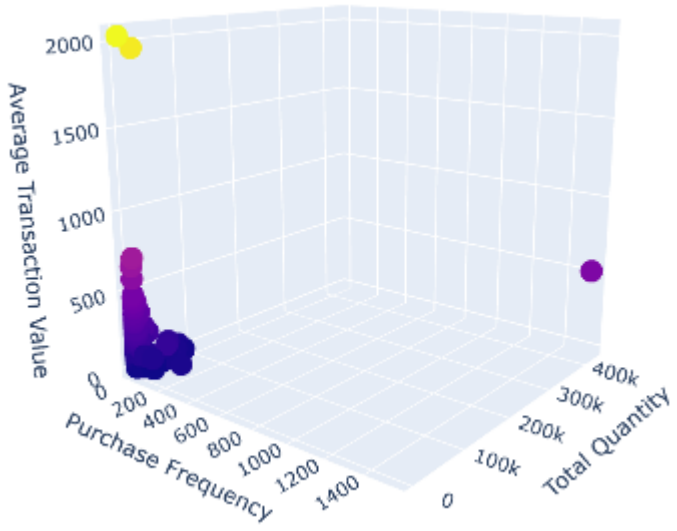


Рисунок 4.20 – Просторова діаграма за трьома базисними ознаками

Джерело: розроблено автором

4.4 Реалізація моделей класифікації за RFM методом

Наступним кроком буде застосування методу RFM для формування сегментів клієнтів. За допомогою цих показників було сегментувано клієнтів на групи. Результати операцій з агрегацією групуванням та знаходженням показників RFM показано на рисунках

CustomerID	Recency	Frequency	Monetary
12346.0	300	1	77183.60
12347.0	14	6	4085.18
12348.0	50	4	1797.24
12350.0	285	1	334.40
12352.0	11	8	2506.04

Рисунок 4.21 – Формування трьох класифікуючих ознак для клієнтів

Джерело: розроблено автором

CustomerID	Recency	Frequency	Monetary	R_Score	M_Score	F_Score	RFM_Score	Segment
12346.0	300	1	77183.60	1	4	1	114	Other
12347.0	14	6	4085.18	4	4	4	444	Champions
12348.0	50	4	1797.24	3	4	3	334	Other
12350.0	285	1	334.40	1	2	1	112	Other
12352.0	11	8	2506.04	4	4	4	444	Champions

Рисунок 4.22 – Розширений набір даних з ознаками балів по RFM методу

Джерело: розроблено автором

Візуалізації RFM методу застосовного до набору даних проілюстровано на рисунках 4.23 – 4.24.

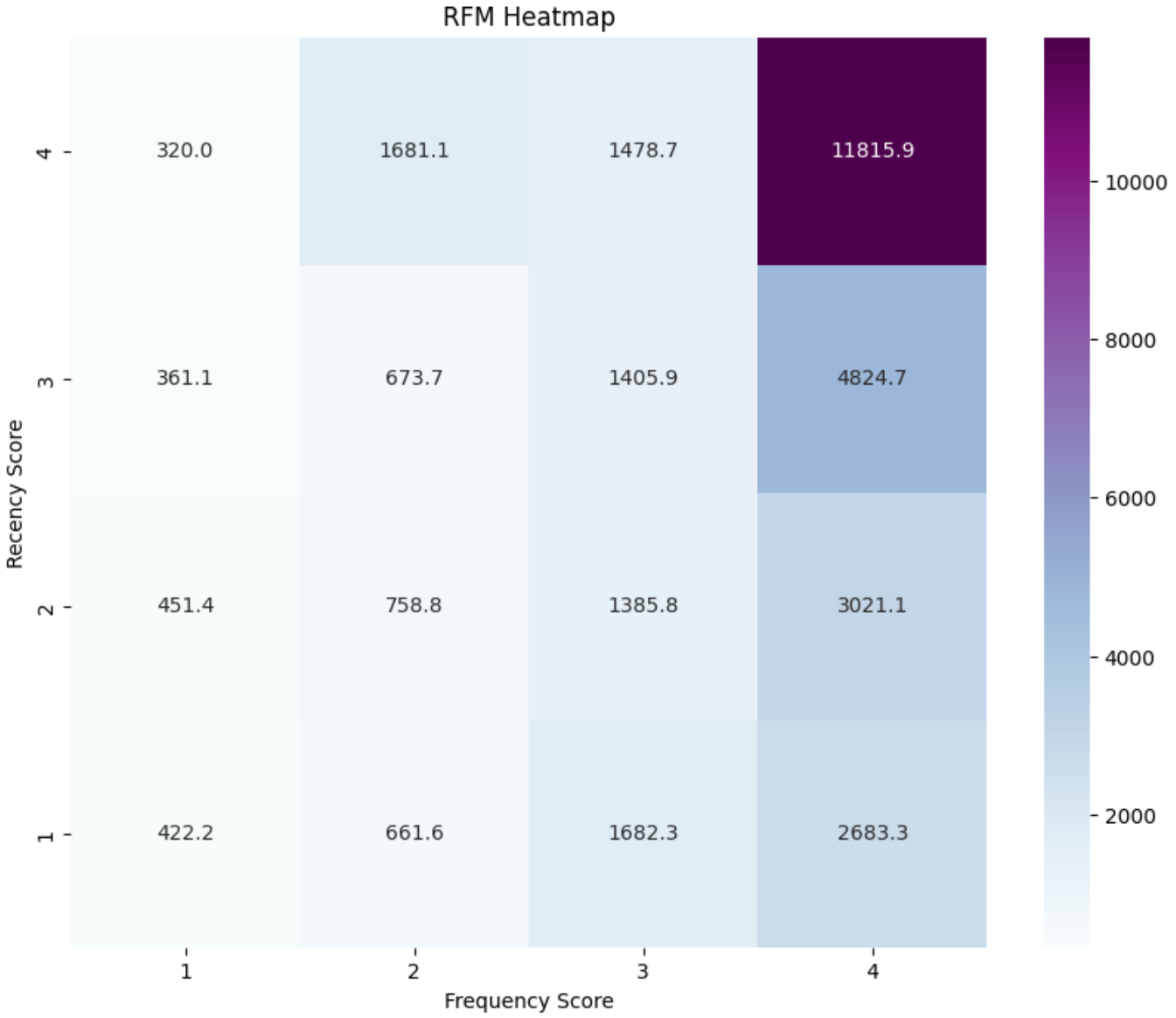


Рисунок 4.23 – Мапа RFM за двома ознаками та їх впливом на набір даних

Джерело: розроблено автором

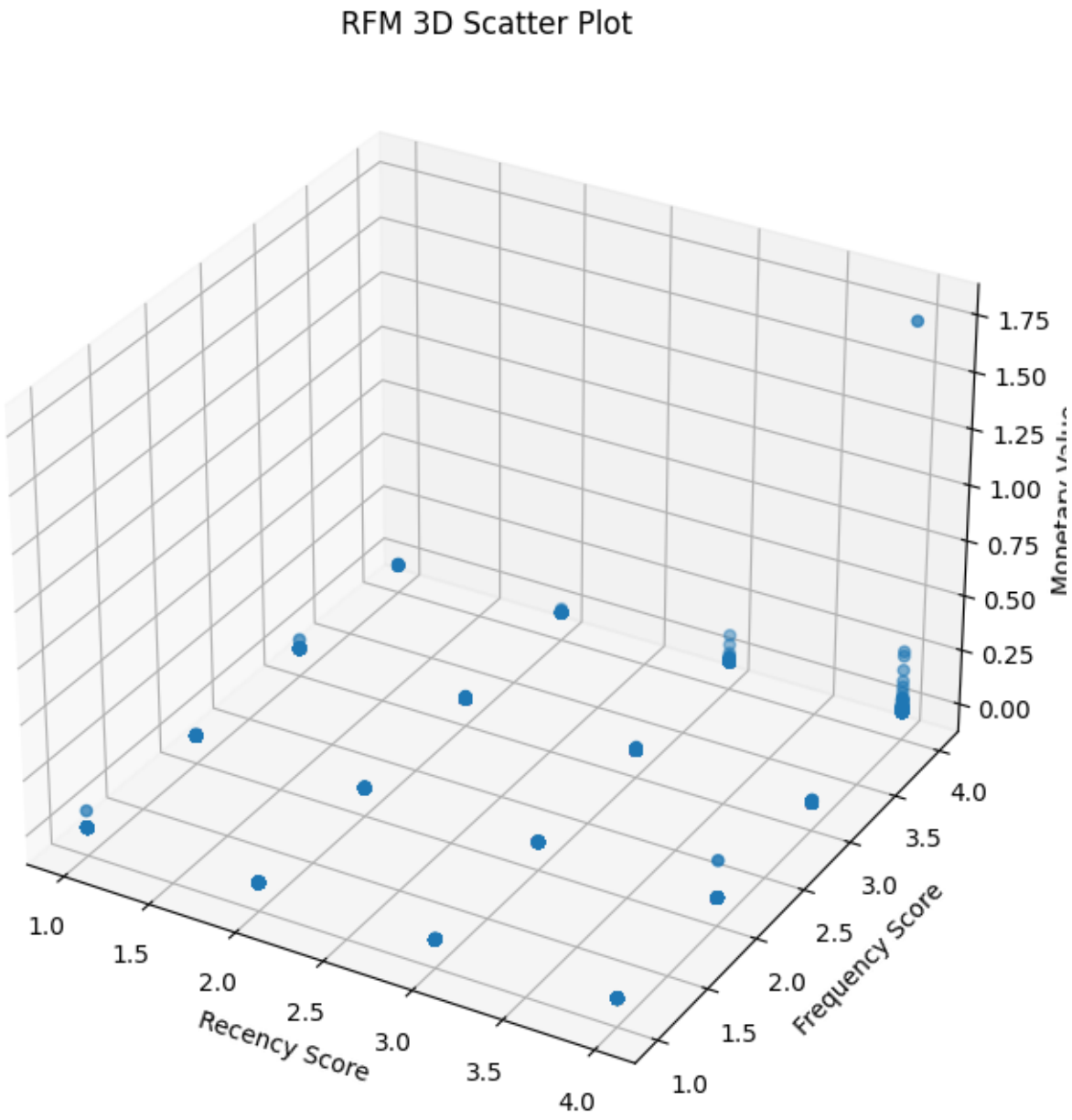


Рисунок 4.24 – Просторова візуалізація за трьома базисними ознаками RFM

Джерело: розроблено автором

Наступним кроком аналізу було визначення кількості та базисних статистичних показників кожного сегменту. Результати зображені на рисунку 4.25.

```
Segment Distribution:
 Other        3260
At Risk       438
Champions     417
Name: Segment, dtype: int64

Segment Profiles:
           Recency Frequency Monetary
             mean      mean     mean        sum
Segment
At Risk     255.9       1.0    160.9    70467.0
Champions     8.7      17.9  12601.4  5254783.4
Other        82.7       2.9   1161.7  3787228.8
```

Рисунок 4.25 – Загальні відомості про кожен сегмент клієнтів після аналізу RFM

Джерело: розроблено автором

Отримані результати по кожному сегменту:

- Чемпіони: Цей сегмент має найнижчу свіжість (в середньому 7,7 днів), високу частоту (в середньому 19,4 покупки) і найвищу грошову цінність. Це найкращі клієнти.
- У зоні ризику: клієнти в цьому сегменті не купували останнім часом (в середньому 267,6 днів), мають низьку частоту (в середньому 1 покупка) і витрачають відносно мало (в середньому 165,2).
- Інші: Різноманітний сегмент, який не підпадає під попередні категорії, з помірною періодичністю та частотою покупок і значним внеском у загальний обсяг продажів.

Для кращого розуміння кожного із сегментів було побудовано візуалізації що розкривають кожний показник RFM та чітко формують ознаки сегментування. Описані вище візуалізації можна побачити на рисунках 4.26-4.29.

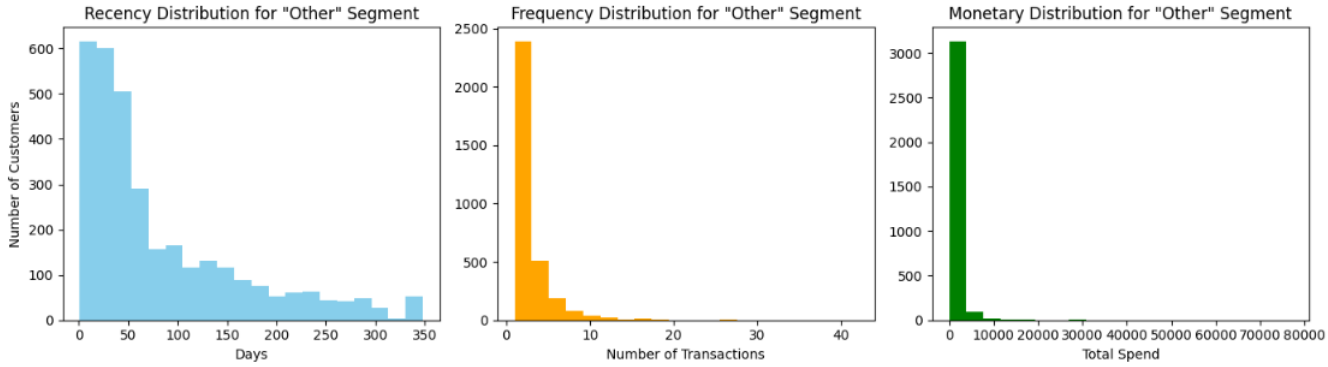


Рисунок 4.26 – Візуалізація RFM ознак для сегменту інші

Джерело: розроблено автором

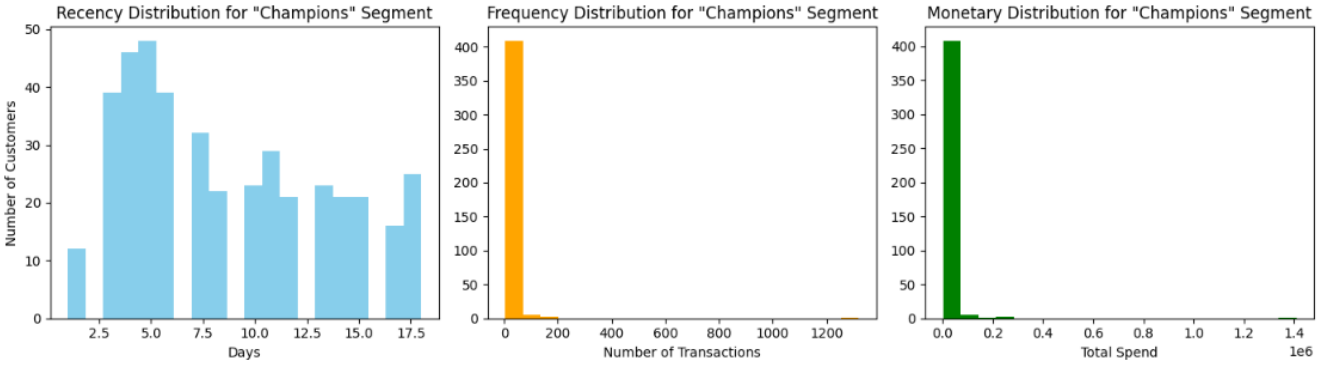


Рисунок 4.27 – Візуалізація RFM ознак для сегменту чемпіони

Джерело: розроблено автором

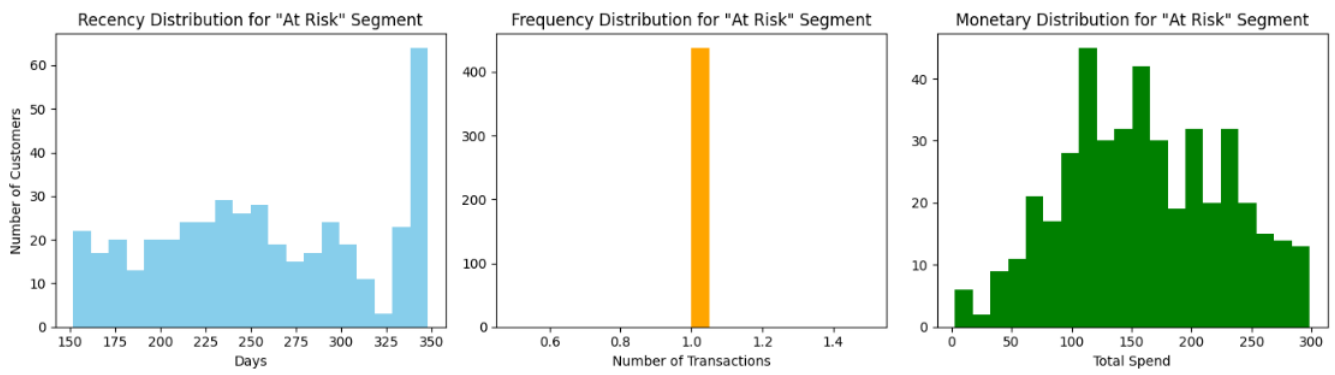


Рисунок 4.28 – Візуалізація RFM ознак для сегменту інші

Джерело: розроблено автором

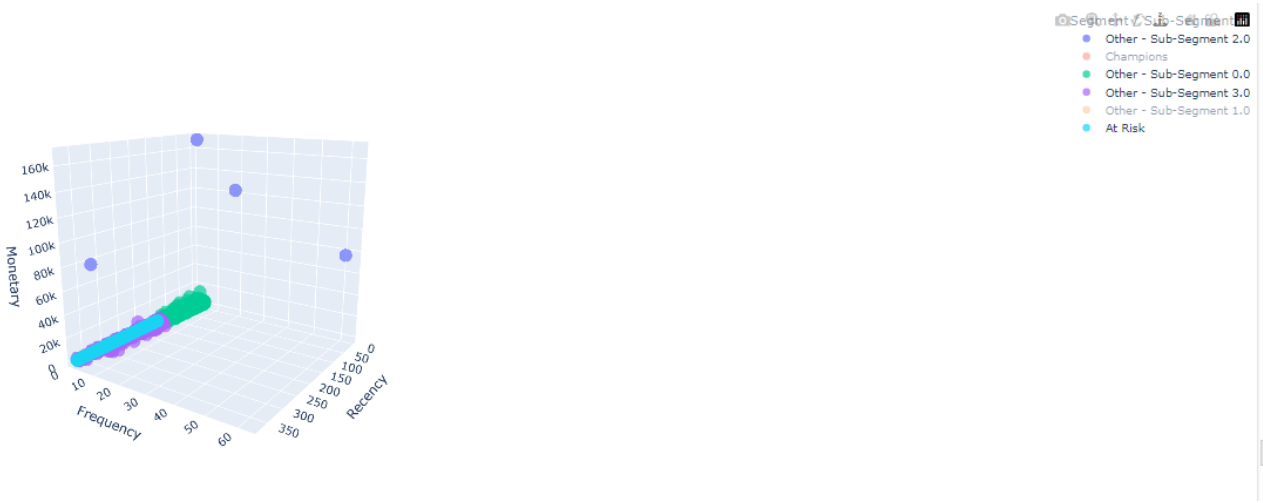


Рисунок 4.29 – Просторова візуалізація RFM ознак для усіх сегментів

Джерело: розроблено автором

Як результат після формування статистичних класів було сформовано наступні класифікаційні моделі:

- Випадковий ліс
- Логістична регресія
- Дерева рішень
- К найближчих сусідів

Для кожної моделі машинного навчання було сформаовано тестові та тренувальні набори. Після тренування моделей для класифікації сегментів клієнтів було отримано наступні резульати. Результати побудови тренування та валідації моделей показано на рисунках 4.30 – 4.31.

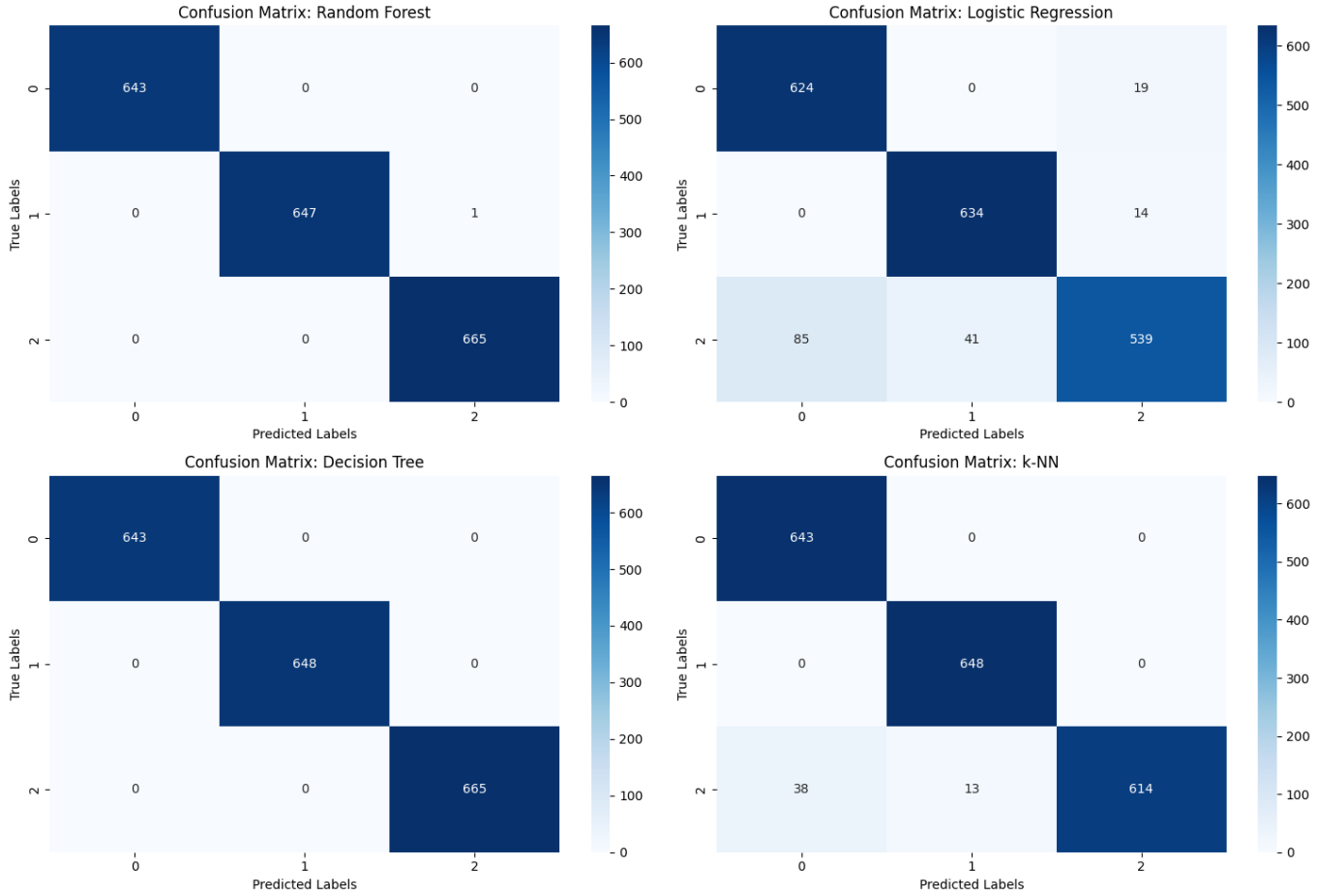


Рисунок 4.30 – Крос валідація моделей та матриця похибки для кожної з моделей

Джерело: розроблено автором

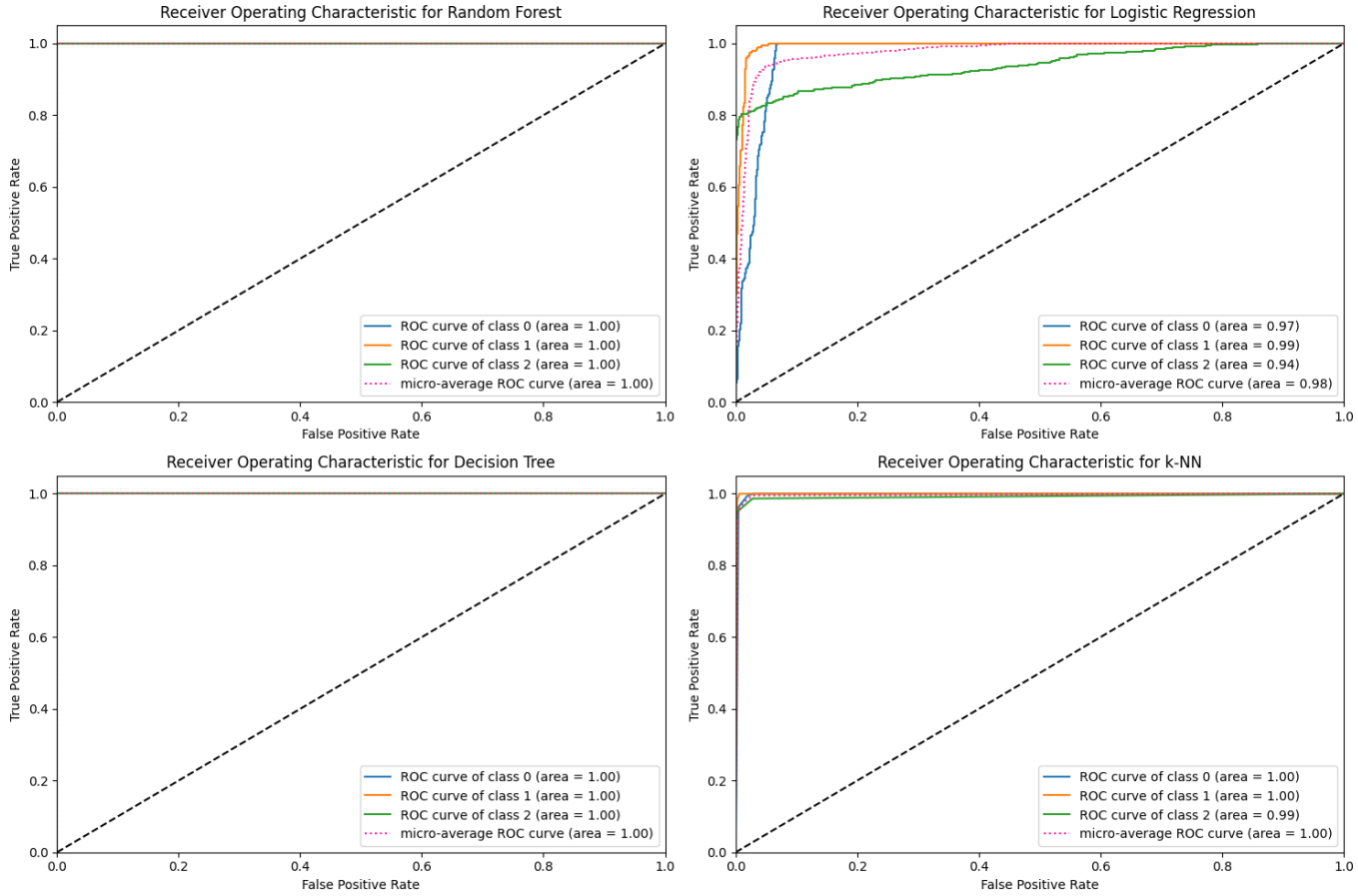


Рисунок 4.31 – Криві робочої характеристики приймача (ROC)

Джерело: розроблено автором

Крива робочої характеристики приймача (ROC) - це графічна діаграма, яка показує діагностичну здатність системи бінарного класифікатора при зміні її порогу дискримінації. Вона створюється шляхом побудови графіка залежності частоти правильних спрацьовувань (TPR) від частоти хибних спрацьовувань (FPR) при різних значеннях порогового значення. Площа під кривою (AUC) є мірою здатності моделі розрізняти позитивні та негативні класи. ROC-криві у побудованих візуалізації порівнюють ефективність чотирьох різних моделей: Випадковий ліс, логістична регресія, дерево рішень і k-NN (k-найближчих сусідів). Ось що показують результати:

Випадковий ліс і дерево рішень:

- Обидві моделі показують ROC-криві, які майже ідеально вирівнюються з верхнім лівим кутом графіка, що свідчить про відмінну ефективність класифікації з AUC 1,00 для всіх класів.
- Мікросередня ROC-крива, яка об’єднує внески всіх класів, також показує AUC 1,00, що підтверджує відмінну продуктивність моделі.

Логістична регресія:

- ROC-криві дуже близькі до верхнього лівого кута, але не такі досконалі, як у випадкового лісу або дерева рішень. Показники AUC коливаються від 0,94 до 0,99 для різних класів, що свідчить про високу, але не ідеальну здатність розрізняти класи.
- Мікросередня ROC-крива також показує високий показник AUC - 0,98, що свідчить про хорошу загальну продуктивність.

k-NN (k-найближчих сусідів):

- Криві ROC для k-NN ідентичні кривим для випадкового лісу та дерева рішень, що дивно, оскільки k-NN зазвичай працює інакше, ніж методи на основі дерев. Він показує AUC 1.00 для всіх класів.

Вихідний код завантажений на GitHub репозиторій та доступний а за посиланням – https://github.com/Demoniolik/master-diploma

ВИСНОВКИ

Метою даного дослідження була розробка інтелектуальної технології використання інтелектуального аналізу даних про клієнтів з платформ електронної комерції з метою отримання дієвих висновків та підвищення ефективності стратегій платформи. В роботі було сформовано мету роботи, обрано інструменти та засоби реалізації.

В рамках аналізу предметної області було обрано та зкореговано шлях та методи дослідження. Важливою частиною було саме розуміння актуальності в такому аналізі. Таким чином було виявлено високий попит на використання аналізу даних та дата майнингу у e-commerce платформах.

Важливою частиною роботи було визначення методів та підходів для аналізу. Сукупність методів було влаштовано у коректному порядку для отримання більш точних та оптимізованих результатів. Таким чином виконано порівняння та використання методів, які мають оптимальне застосування щодо набору даних та поставлених задач. Як результат, визначено такий порядок процесів аналізу: збір та уніфікація даних, обробка та очищення набору, статистичний аналіз для розширення даних та формування груп та кластерів, візуалізація отриманих результатів та ідентифікація ознак, побудова моделей машинного навчання для прогнозування та формування інсайтів.

В рамках кваліфікаційної роботи магістра проведено комплексний аналіз даних користувачів електронної комерції для пошуку закономірностей та формування представлення купівельної поведінки користувачів. На першому етапі було оброблено та очищено дані для попереднього аналізу. В результаті, виявлено високу кількість аномальних чинників таких як: негативні кількість товарів та ціна. Такі аномалії надали змогу підійти до пошуку та розуміння логіки, яка лежить за ними, що призвело до отримання вичерпної інформації та сформування припущення, що негативна кількість це повернення. На основі припущення було

побудовано логістичну регресійну модель для прогнозування клієнтів, які можуть повертати товари.

Протягом другої фази аналізу було проведено RFM аналіз, що дозволило вирахувати важливі ознаки. Ці ознаки були використанні для кращого розуміння купівельних звичок клієнтів та побудови класифікаційних моделей. Такі моделі мають багато призначень та в рамках цієї роботи вони були використанні для формування класів, які в майбутньому можуть бути використанні для формування персоналізованих пропозицій та різних стратегій щодо залучення клієнтів. Загально, всі ці моделі були натреновані на однакових наборах даних та, таким чином, було перевірено точність класифікації кожної з них між собою та отримано доволі хороші результати.

Інформаційна технологія обробки та аналізу даних покупців e-commerce платформ показала високу ефективність: в ідентифікації аномалій та знаходженні додаткових показників що їх розкривають, в класифікації клієнтів та отриманні інформації про шаблони їх поведінки. Ці результати приносять користь для побудови стратегій заохочення користувачів, оптимізації логістики та загальних бізнес процесів платформ.

Результати досліджень було представлено на VI Всеукраїнській конференції молодих вчених, курсантів та студентів «Інформаційна безпека та інформаційні технології» (ІБІТ – 2023).

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. A Data Mining Based Approach to Customer Behaviour in an Electronic Settings [Електронний ресурс] – Доступ до ресурсу: https://www.scirp.org/journal/paperinformation.aspx?paperid=92828 (Дата звернення: 07.10.23)

2. Information mining of customers preferences for product specifications determination using big sales data [Електронний ресурс] – Доступ до ресурсу: https://www.sciencedirect.com/science/article/pii/S2212827122006692 (Дата звернення: 09.10.23)

3. Efficient mining of intra-periodic frequent sequences [Електронний ресурс] – Доступ до ресурсу: https://www.sciencedirect.com/science/article/pii/S2590005622000960?via%3Dihub (Дата звернення: 11.10.23)

4. Anomaly Detection in Data Mining: A Comprehensive Guide 101 [Електронний ресурс] – Доступ до ресурсу: https://hevodata.com/learn/anomaly-detection-in-data-mining/ (Дата звернення: 11.10.23)

5. Customer analysis: Definition, benefits & how to perform it the right way [Електронний ресурс] – Доступ до ресурсу: https://www.paddle.com/resources/customer-analysis (Дата звернення: 18.10.23)

6. How to unlock business insights through customer data analysis way [Електронний ресурс] – Доступ до ресурсу: https://monday.com/blog/crm-and-sales/customer-data-analysis/ (Дата звернення: 21.10.23)

7. Customer Data Analysis Best Practices You Need to Know way [Електронний ресурс] – Доступ до ресурсу: https://learn.g2.com/customer-data-analysis-best-practices (Дата звернення: 27.10.23)

8. Customer lifetime value: The customer compass [Електронний ресурс] – Доступ до ресурсу: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/customer-lifetime-value-the-customer-compass (Дата звернення: 30.10.23)

9. The Impact of big data analytics on E-commerce Personalization and Marketing Strategies [Електронний ресурс] – Доступ до ресурсу: https://www.linkedin.com/pulse/impact-big-data-analytics-e-commerce-personalization-marketing (Дата звернення: 12.11.23)

10. Ecommerce Machine Learning Brings the Future to the Present of Online Shopping [Електронний ресурс] – Доступ до ресурсу: https://www.bigcommerce.com/articles/ecommerce/machine-learning/ (Дата звернення: 12.11.23)

11. How do you use data mining to influence customer behavior? [Електронний ресурс] – Доступ до ресурсу: https://www.linkedin.com/advice/3/how-do-you-use-data-mining-influence-customer-behavior (Дата звернення: 12.11.23)

12. Unleashing Sales Potential: Data Science for Customer Lifetime Value (CLV) Prediction [Електронний ресурс] – Доступ до ресурсу: https://www.linkedin.com/pulse/unleashing-sales-potential-data-science-customer-lifetime-sawla (Дата звернення: 13.11.23)

13. Data Mining [Електронний ресурс] – Доступ до ресурсу: https://www.qlik.com/us/data-analytics/data-mining#:~:text=,data%20to%20make%20accurate%20predictions (Дата звернення: 13.11.23)

14. Functional and Nonfunctional Requirements: Specification and Types [Електронний ресурс] – Доступ до ресурсу: https://www.altexsoft.com/blog/functional-and-non-functional-requirements-specification-and-types/ (Дата звернення: 05.12.23)

15. Functional vs Non Functional Requirements [Електронний ресурс] – Доступ до ресурсу: https://www.geeksforgeeks.org/functional-vs-non-functional-requirements/ (Дата звернення: 05.12.23)

16. Non-functional Requirements: Examples, Types, How to Approach [Електронний ресурс] – Доступ до ресурсу: https://www.altexsoft.com/blog/non-functional-requirements/ (Дата звернення: 05.12.23)

17. An Introduction to Python for Data Analytics [Електронний ресурс] – Доступ до ресурсу: https://www.linkedin.com/pulse/introduction-python-data-analytics-sid-mehandru-p99kf?trk=article-ssr-frontend-pulse_more-articles_related-content-card (Дата звернення: 09.11.23)

18. Pandas for data management and data analysis [Електронний ресурс] – Доступ до ресурсу: https://svitla.com/blog/pandas-for-data-management-and-data-analysis (Дата звернення: 09.11.23)

19. NumPy: the absolute basics for beginners [Електронний ресурс] – Доступ до ресурсу: https://numpy.org/doc/stable/user/absolute_beginners.html (Дата звернення: 09.11.23)

20. Scikit-learn – Доступ до ресурсу: https://www.nvidia.com/en-us/glossary/data-science/scikit-learn/ (Дата звернення: 09.11.23)

21. What is Matplotlib in Python? [Електронний ресурс] – Доступ до ресурсу: https://www.scaler.com/topics/matplotlib/matplotlib-in-python/ (Дата звернення: 09.11.23)

22. An introduction to seaborn [Електронний ресурс] – Доступ до ресурсу: https://seaborn.pydata.org/tutorial/introduction (Дата звернення: 10.11.23)

23. Getting Started with Plotly in Python – Доступ до ресурсу: https://plotly.com/python/getting-started/ (Дата звернення: 10.11.23)

24. What is python Bokeh? [Електронний ресурс] – Доступ до ресурсу: https://www.educative.io/answers/what-is-python-bokeh (Дата звернення: 10.11.23)

25. What is Jupyter Notebook? – Доступ до ресурсу: https://domino.ai/data-science-dictionary/jupyter-notebook (Дата звернення: 10.11.23)

26. Colaboratory [Електронний ресурс] – Доступ до ресурсу: https://research.google.com/colaboratory/faq.html (Дата звернення: 10.11.23)

27. What Is Data Preparation and Why Is It Important? – Доступ до ресурсу: https://blogs.oracle.com/analytics/post/what-is-data-preparation-and-why-is-it-

important#:~:text=Put%20simply%2C%20data%20preparation%20is,something%20digestible%20by%20analytics%20tools (Дата звернення: 10.11.23)

28. What are the benefits of data preparation? [Електронний ресурс] – Доступ до ресурсу: https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation#:~:text=Done%20properly%2C%20data%20preparation%20also,data%20management%20and%20analytics%20costs (Дата звернення: 10.11.23)

29. Data Preprocessing Techniques: 6 Steps to Clean Data in Machine Learning [Електронний ресурс] – Доступ до ресурсу: https://www.scalablepath.com/data-science/data-preprocessing-phase (Дата звернення: 10.11.23)

30. What is Data Mining? [Електронний ресурс] – Доступ до ресурсу: https://www.qlik.com/us/data-analytics/data-mining#:~:text=,of%20new%2C%20unseen%20data%20instances

31. Time Series Analysis: Definition, Types, Techniques, and When It's Used [Електронний ресурс] – Доступ до ресурсу: https://www.tableau.com/learn/articles/time-series-analysis (Дата звернення: 10.11.23)

32. Statistics and probability [Електронний ресурс] – Доступ до ресурсу: https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/mean-median-basics/a/mean-median-and-mode-review

33. Feature Engineering: Scaling, Normalization, and Standardization (Updated 2023) [Електронний ресурс] – Доступ до ресурсу: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

34. Z-scores review [Електронний ресурс] – Доступ до ресурсу: https://www.khanacademy.org/math/statistics-probability/modeling-distributions-of-data/z-scores/a/z-scores-review

35. What is RFM Analysis? [Електронний ресурс] – Доступ до ресурсу: https://www.actioniq.com/blog/what-is-rfm-analysis/ (Дата звернення: 10.11.23)

36. SMOTE for Imbalanced Classification with Python [Електронний ресурс] – Доступ до ресурсу: https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-

imbalance-using-smote-techniques/#h-smote-synthetic-minority-oversampling-technique (Дата звернення: 10.11.23)

37. Logistic regression [Електронний ресурс] – Доступ до ресурсу: https://datatab.net/tutorial/logistic-regression (Дата звернення: 10.11.23)

38. What is a Decision Tree? [Електронний ресурс] – Доступ до ресурсу: https://www.ibm.com/topics/decision-trees (Дата звернення: 10.11.23)

39. Decision Trees [Електронний ресурс] – Доступ до ресурсу: https://scikit-learn.org/stable/modules/tree.html (Дата звернення: 10.11.23)

40. What is random forest? [Електронний ресурс] – Доступ до ресурсу: https://www.ibm.com/topics/random-forest (Дата звернення: 10.11.23)

41. Відомості про діаграму IDEF0 [Електронний ресурс] – Доступ до ресурсу: https://www.edrawmax.com/article/the-complete-guide-to-understand-idef-diagram.html (Дата звернення: 14.11.23)

42. Відомості про діаграму декомпозиції [Електронний ресурс] – Доступ до ресурсу: https://www.visual-paradigm.com/guide/uml-unified-modeling-language/what-is-use-case-diagram/ (Дата звернення: 14.11.23)

43. Відомості про діаграму діяльності [Електронний ресурс] – Доступ до ресурсу: https://ppt-online.org/751856 (Дата звернення: 15.11.23)

44. Відомості про діаграму потоку даних [Електронний ресурс] – Доступ до ресурсу: https://www.lucidchart.com/pages/data-flow-diagram (Дата звернення: 22.11.23)

ДОДАТОК А ПЛАНУВАННЯ РОБІТ

А.1 Ідентифікація мети проекту

Призначення інтелектуальної технології

Даний аналіз призначений для покращення бізнес процесів e-commerce платформ. Основними віхами є отримання ефективного аналізу уже існуючої системи процесів та їх покращення на основі статистичного аналізу та технік дата майнингу. Таким чином бізнес отримує вичерпну інформацію щодо функціонування поточних процесів та розуміння того як можна покращити: продажі, логістику, таргетинг та отримати загальне розуміння хто є кінцевий клієнт та як з ним можна взаємодіяти.

Мета створення інтелектуальної технології

Метою даної роботи є формування ефективного та вичерпного аналізу користувачів e-commerce платформи для формування стратегій покращення бізнес процесів.

Цільова аудиторія

Основними зацікавленими особами є маркетологи бізнес аналітики та загалом інвестори e-commerce платформи. Фактично зацікавленими особами можуть служити і більша кількість людей залежить від структури компанії.

А.2 Планування змісту структури робіт проекту

Важливою частиною розробки проекту є ідентифікація та фіксація проміжних етапів проекту. На табличці А.2 можна побачити основні етапи проекту.

Таблиця А.2 – Етапи проекту

№	Склад і зміст робіт	Строк розробки (у робочих днях)
1	Постановка задачі проекту	10 днів
2	Складання технічного завдання	5 днів
3	Підготовка та пошук набору даних	10 днів
4	Первинний аналіз та ідентифікація релевантності набору даних	10 днів
5	Очищення даних та підготовка до аналізу	7 дні
6	Статистичний аналіз та підготовка даних до побудови моделей	15 днів
7	Візуалізація та формування проміжних звітів	10 днів
8	Побудова моделей	10 днів
9	Навчання та корегування моделей	7 днів
10	Завершення роботи	3 дні
	Загальна тривалість робіт	87 днів

Джерело: розроблено автором

А.3 Побудова календарного графіку виконання проекту

Побудова календарного плану надає можливість відслідковувати та ефективно розподіляти ресурси та час для оптимального виконання проектної роботи. Для візуалізації процесів та відповідальних осіб було побувдовано діаграми WBS та OBS, які можна побачити на рисунках А.1 та А.2. Діаграма Ґанта ілюструє весь перелік робіт з урахуванням тривалості да строків виконання кожної задачі її можна побачити на рисунку А.3-А.4.

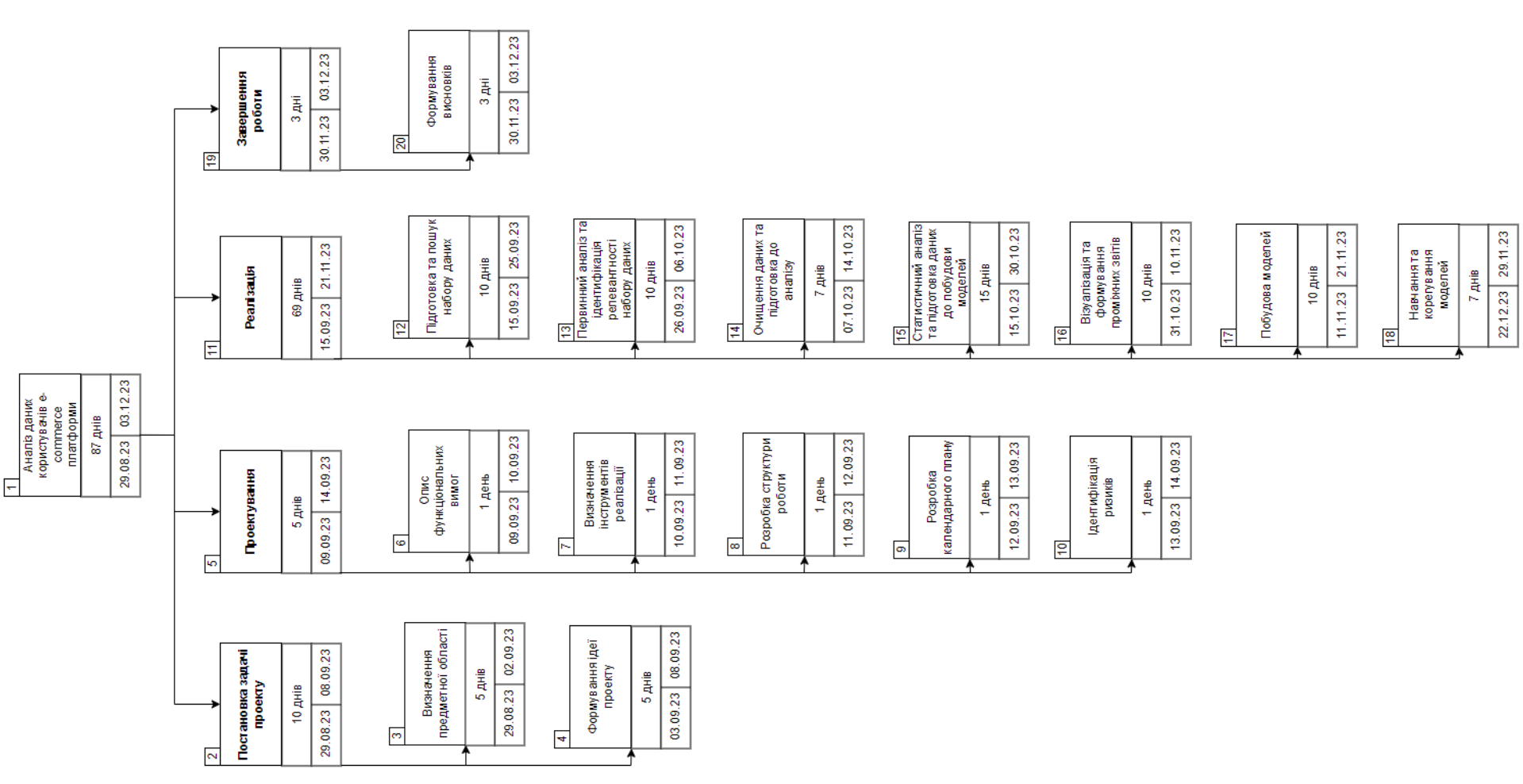


Рисунок А.1 – Діаграма декомпозиції проекту засобами WBS

Джерело: розроблено автором

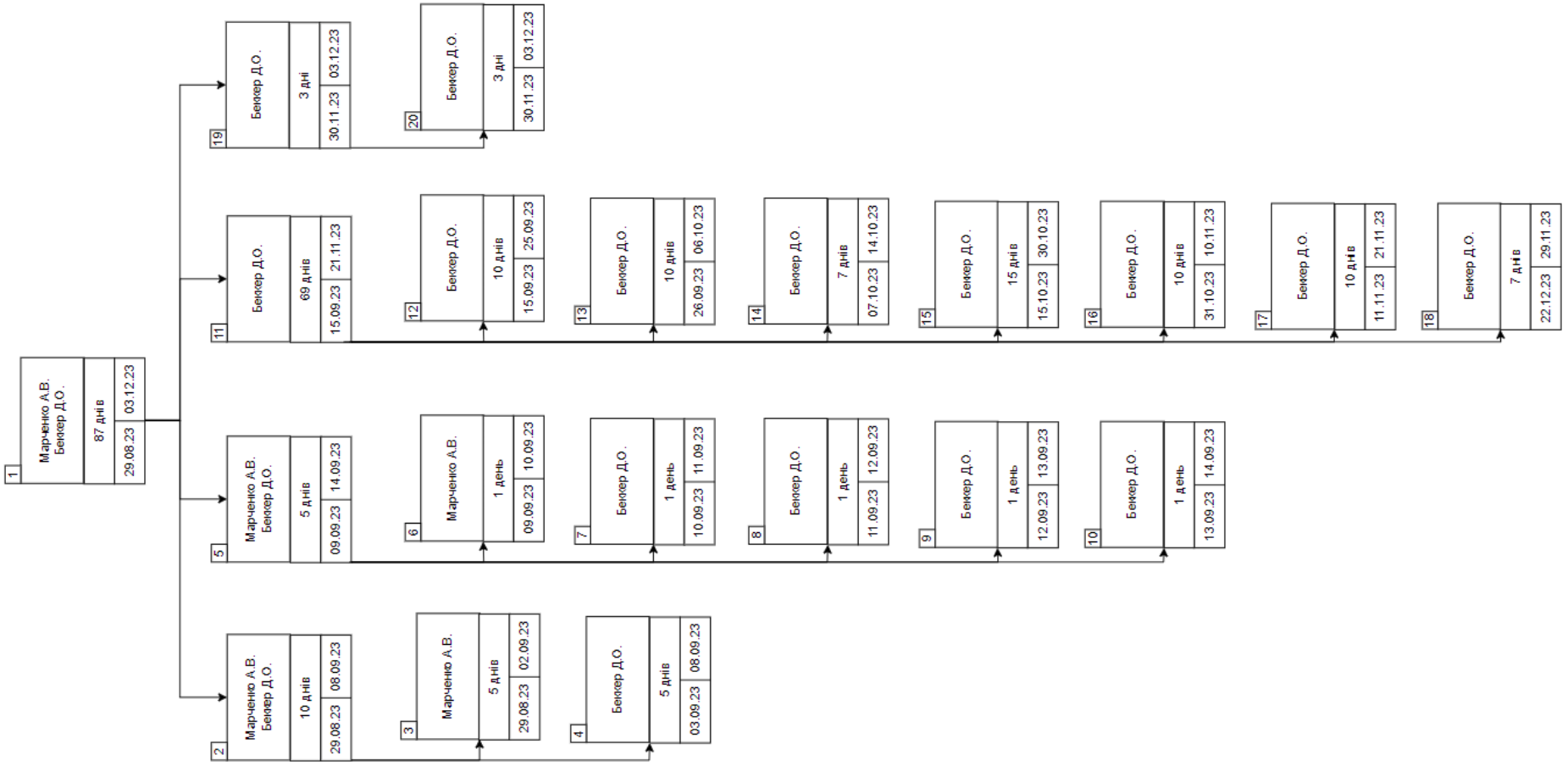
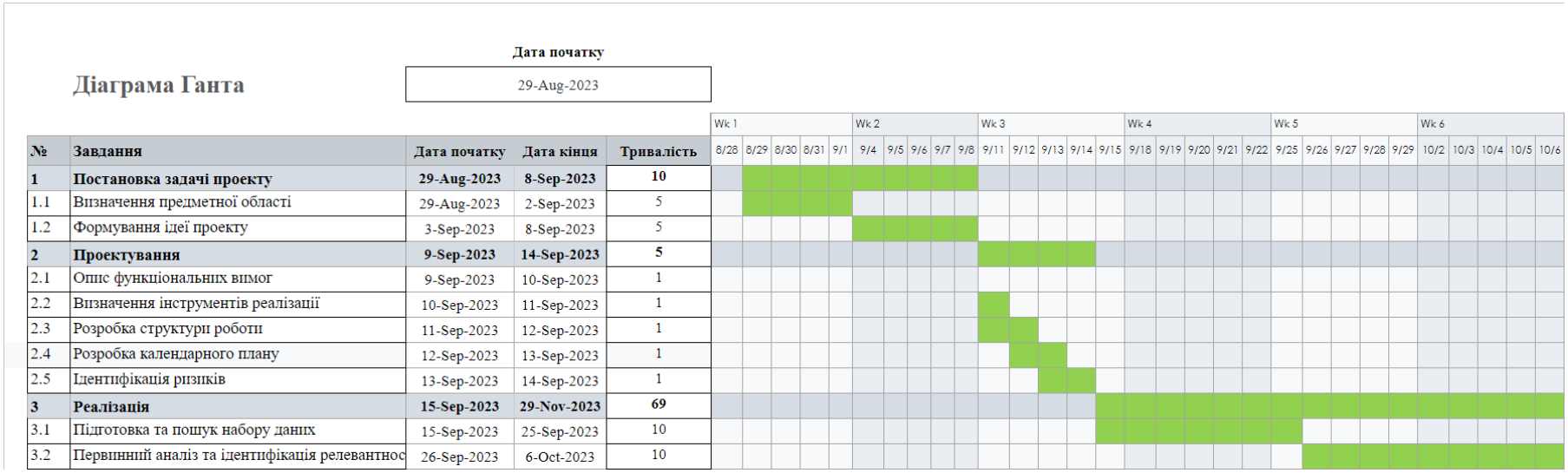


Рисунок А.2 – Діаграма декомпозиції проекту засобами OBS

Джерело: розроблено автором



Діаграма Ганта

Дата початку: 29-Aug-2023

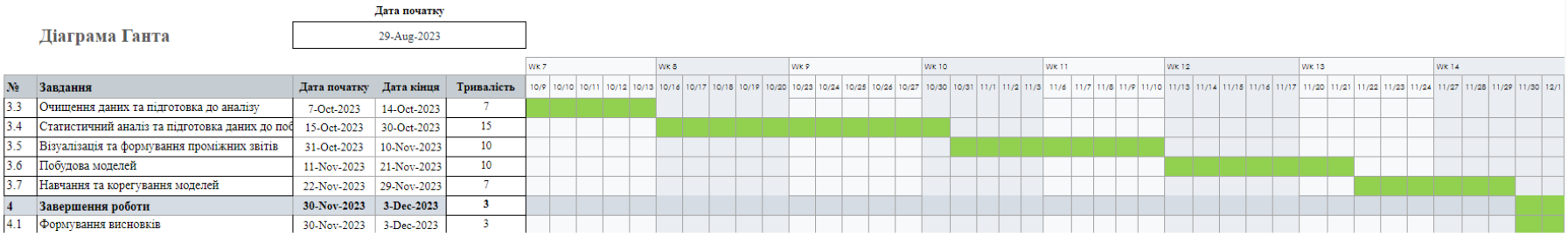
№	Завдання	Дата початку	Дата кінця	Тривалість
1	**Постановка задачі проекту**	**29-Aug-2023**	**8-Sep-2023**	**10**
1.1	Визначення предметної області	29-Aug-2023	2-Sep-2023	5
1.2	Формування ідеї проекту	3-Sep-2023	8-Sep-2023	5
2	**Проектування**	**9-Sep-2023**	**14-Sep-2023**	**5**
2.1	Опис функціональних вимог	9-Sep-2023	10-Sep-2023	1
2.2	Визначення інструментів реалізації	10-Sep-2023	11-Sep-2023	1
2.3	Розробка структури роботи	11-Sep-2023	12-Sep-2023	1
2.4	Розробка календарного плану	12-Sep-2023	13-Sep-2023	1
2.5	Ідентифікація ризиків	13-Sep-2023	14-Sep-2023	1
3	**Реалізація**	**15-Sep-2023**	**29-Nov-2023**	**69**
3.1	Підготовка та пошук набору даних	15-Sep-2023	25-Sep-2023	10
3.2	Первинний аналіз та ідентифікація релевантнос	26-Sep-2023	6-Oct-2023	10

Wk 1 | Wk 2 | Wk 3 | Wk 4 | Wk 5 | Wk 6

8/28 8/29 8/30 8/31 9/1 9/4 9/5 9/6 9/7 9/8 9/11 9/12 9/13 9/14 9/15 9/18 9/19 9/20 9/21 9/22 9/25 9/26 9/27 9/28 9/29 10/2 10/3 10/4 10/5 10/6

Рисунок А.3 – Перша частина діаграми Ґанта

Джерело: розроблено автором



Діаграма Ганта

Дата початку: 29-Aug-2023

№	Завдання	Дата початку	Дата кінця	Тривалість
3.3	Очищення даних та підготовка до аналізу	7-Oct-2023	14-Oct-2023	7
3.4	Статистичний аналіз та підготовка даних до пос	15-Oct-2023	30-Oct-2023	15
3.5	Візуалізація та формування проміжних звітів	31-Oct-2023	10-Nov-2023	10
3.6	Побудова моделей	11-Nov-2023	21-Nov-2023	10
3.7	Навчання та корегування моделей	22-Nov-2023	29-Nov-2023	7
4	**Завершення роботи**	**30-Nov-2023**	**3-Dec-2023**	**3**
4.1	Формування висновків	30-Nov-2023	3-Dec-2023	3

Wk 7 | Wk 8 | Wk 9 | Wk 10 | Wk 11 | Wk 12 | Wk 13 | Wk 14

10/9 10/10 10/11 10/12 10/13 10/16 10/17 10/18 10/19 10/20 10/23 10/24 10/25 10/26 10/27 10/30 10/31 11/1 11/2 11/3 11/6 11/7 11/8 11/9 11/10 11/13 11/14 11/15 11/16 11/17 11/20 11/21 11/22 11/23 11/24 11/27 11/28 11/29 11/30 12/1

Рисунок А.4 – Друга частина діаграми Ґанта

Джерело: розроблено автором

А.4 Планування ризиків проекту

Для успішного планування та виконання поставлених задач потрібно сформувати ризики та їх вплив на загальне виконання завдань. Таким чином можна зрозуміти та знайти рішення основних проблем з якими можна зіткнутися. Таке розуміння надає можливість більш гнучко підійти до планування проекту та швидко зрозуміти та зреагувати на основні потенційні проблеми. Перелік ризиків та їх вплив можна побачити на таблиці А.4.1.

Таблиця А.4.1 – Ризики проекту

Ризик	Результат впливу	Оцінка впливу	Оцінка ймовірності	Ранг
Проблеми з якістю даних	Неточні або неповні дані клієнта впливають на прийняття рішень	Середній: вплив на аналітику та отримані результати	Помірний	Середній
Відсутність досвіду аналізу даних	Нездатність ефективно аналізувати дані клієнтів за допомогою інтелектуальних технологій	Високий: погіршення якості дослідження, можливе неправильне тлумачення результатів	Помірний	Середній
Часові обмеження	Недостатньо часу для ретельного дослідження, аналізу та документування	Високий: вплив на глибину та якість дипломної роботи	Високий	Високий
Технічна складність	Проблеми інтеграції з існуючими системами електронної комерції, що призводить до затримок	Середній: затримка завершення проекту	Високий	Високий
Недостатній доступ до реальних даних	Обмежена доступність різноманітних і репрезентативних даних про клієнтів для аналізу	Середній: обмеження надійності та узагальненості результатів	Помірний	Середній

Продовження таблиці А.4.1 – Ризики проекту

Складність літературних джерел	Складність і специфічність досліджень в цій темі	Середній: вплив на якість роботи	Середній	Середній
Обмежена доступність ресурсів	Недостатній доступ до обчислювальних ресурсів, програмного забезпечення або спеціалізованих інструментів	Високий: вплив на дослідницькі можливості та результати	Помірний	Високий

Джерело: розроблено автором

На основі попередньо визначених ризиків можна збудувати заходи по їх запобіганню або зменшенню. На основі таких заходів будуються стратегії та оптимізація роботи загалом.

Таблиця А.4.2 – Стратегії по оптимізації та вирішенню ризиків

Ризик	Результат впливу	Заходи запобігання виникненню ризиків	Стратегія управління ризиками
Проблеми з якістю даних	Неточні або неповні дані клієнта впливають на прийняття рішень	Впровадження процедури перевірки та очищення даних та використання інструментів забезпечення якості даних.	Мінімізація
Відсутність досвіду аналізу даних	Нездатність ефективно аналізувати дані клієнтів за допомогою інтелектуальних технологій	Пошук навчальних програм, аналіз схожих досліджень в даній сфері	Мінімізація
Часові обмеження	Недостатньо часу для ретельного дослідження, аналізу та документування	Розробка реалістичного графіку проекту з основними етапами, визначення пріоритетності завдань та критичного шляху	Мінімізація

Продовження таблиці А.4.2 – Стратегії по оптимізації та вирішенню ризиків

Технічна складність	Проблеми інтеграції з існуючими системами електронної комерції, що призводить до затримок проекту	Проведення ретельного аналізу системи перед впровадженням	Мінімізація
Недостатній доступ до реальних даних	Обмежена доступність різноманітних і репрезентативних даних про клієнтів для аналізу	Вивчення відкритих джерел даних, що стосуються галузі, пошук способів покращення якості дослідження на основі існуючих даних	Відхилення
Складність літературних джерел	Складність і специфічність досліджень в цій темі	Вивчення додаткових джерел	Мінімізація
Обмежена доступність ресурсів	Недостатній доступ до обчислювальних ресурсів, програмного забезпечення або спеціалізованих інструментів	Оптимізація використання ресурсів за допомогою ефективного планування	Делегування

Джерело: розроблено автором