УДК 004.67 DOI: 10.18372/2073-4751.79.19385

Сулейманова С.Р.,
e-mail: 3647768@stud.nau.edu.ua

РОЗВІДУВАЛЬНИЙ АНАЛІЗ ТА ВІЗУАЛІЗАЦІЯ НАБОРУ ДАНИХ НА ПРИКЛАДІ ПІДПРИЄМСТВА ЕЛЕКТРОННОЇ ТОРГІВЛІ

Національний авіаційний університет

Вступ

За даними *Avada Commerce* у 2024 році глобальний ринок електронної комерції продовжує демонструвати стійке зростання. Останній звіт *eMarketers* передбачає значне зростання в секторі електронної комерції, оцінюючи зростання глобальних продажів на 9,4% у 2024 році [3].

В Україні, за даними інтернет-агенції *Promodo* в 2023 році доля електронної комерції 10% від загального обсягу роздрібної торгівлі. Загальний обсяг продажів товарів, проданих онлайн, склав 182 мільярди гривень [8]. В Україні прийнятий та функціонує Закон «Про електронну комерцію» №675-*VIII* від 03 вересня 2015 року. Даний законодавчий акт встановлює загальні правила та особливості здійснення електронної торгівлі в Україні [2].

Дана сфера потребує постійного проведення інтелектуального аналізу даних для підтримки прийняття стратегічних рішень. Аналіз даних включає в себе процес збору, очищення, організації та інтерпретації великого обсягу даних з метою виявлення цінних знань, закономірностей та патернів. Використання машинного навчання та аналізу даних у прогнозуванні тенденцій стало потужним інструментом [1]. Прикладами використання є персоналізація покупок, сегментація клієнтів, оптимізація логістики та ціноутворення, аналіз поведінки користувачів та боротьба з шахрайством [4]. Для використання алгоритмів Машинного навчання необхідно провести розвідувальний аналіз даних. Підходи до проведення цього аналізу буде розглянуто в даній статті.

Огляд інструментів для розвідувального аналізу даних

Для проведення інтелектуального аналізу даних існує багато програмних засобів: пакети *Python* для аналізу даних (*numpy* – обробка багатовимірних масивів та математичні операції над ними [12], *pandas* – маніпулювання та аналіз даних, індексування, групування, злиття даних, функціональність для часових рядів [13]; *scikit-learn* – бібліотека машинного навчання для створення та тренування алгоритмів кластеризації, класифікації, регресії [15]; *matplotlib* – бібліотека, яка підтримує лінійні графіки, діаграми розсіювання, стовпчасті, секторні діаграми, тощо [11]; *seaborn* – бібліотека, що базується на *matplotlib* дозволяє створювати комплексні інтерактивні графіки [16]). Для проведення аналізу даних також використовуються платформи для візуалізації даних *Tableau* [18], *Power BI* [14], *Looker* [10], *Apache Superset* [5]. Ці програмні засоби дозволяють обробляти даних з різних джерел даних, націлені на створення інтерактивної візуалізації та орієнтовані на бізнес-аналітику.

У статті буде запропоновано комбінований підхід до розвідувального аналізу даних за допомогою пакетів мови програмування *Python* та платформи *Tableau*. Основними перевагами перерахованих аналітичних пакетів *Python* є ліцензія на вільне використання багатоплатформність, детальна документація, широкий набір функцій. Платформа *Tableau* використовується на численних підприємствах, має зрозумілий інтерфейс та широкий набір візуалізацій, вбудованих функцій для аналітики (кластеризації, розпізнавання та побудова трендів), можливість публікувати дашборди на сервері.

Опис набору даних

Дані для дослідження взяті з *Online Retail II* датасету [6]. Цей набір даних містить усі транзакції, здійснені у

роздрібному онлайн-магазині, що базується в Великобританії між 01/12/2009 та 09/12/2011. Компанія в основному продає подарункові вироби широкого призначення. Велика частина клієнтів компанії – оптові продавці.

В цьому наборі даних міститься інформація про номер замовлення, унікальний номер товару, опис товару, кількість придбаних одиниць, дата замовлення, ціна за одну одиницю, унікальний номер клієнта та країна, що описані в табл. 1.

Таблиця 1. Опис набору даних

Назва колонки (англійською)	Назва колонки (українською)	Тип даних	Опис
Invoice	Номер замовлення	*str*	Унікальний номер замовлення
Stock Code	Номер одиниці	*str*	Унікальний номер товару
Description	Опис товару	*str*	Опис товару
Invoice Date	Дата замовлення	*datetime*64	Дата замовлення
Customer ID	Номер клієнта	*int*64	Унікальний номер клієнта
Quantity	Кількість	*int*64	Кількість кожного продукту (позиції) на транзакцію
Amount	Продажі	*float*64	Продажі товару: (Кількість*Продажі)
Price	Ціна	*float*64	Ціна продукту за одиницю в фунтах стерлінгів
Country	Країна	*str*	Країна адреси доставки

Попередня обробка даних та інтерпретація результатів

Розвідувальний аналіз даних був проведений у середовищі *Tableau* та за допомогою мови програмування *Python*. Для ретельного вивчення даних та розробки точних прогнозних моделей в майбутніх дослідженнях, необхідно провести попередню обробку даних, яка може складатися з таких кроків: конвертація даних у необхідні формати, очистка, фільтрація, агрегування, логарифмування, візуалізація даних, детектування та видалення аномалій.

Для вибору країни для аналізу було побудовано мапу кількості проданих одиниць (рис. 1). Було визначено, що найбільша країна по кількості проданих одиниць – це Велика Британія. Подальший аналіз був проведений на основі даних цієї країни.

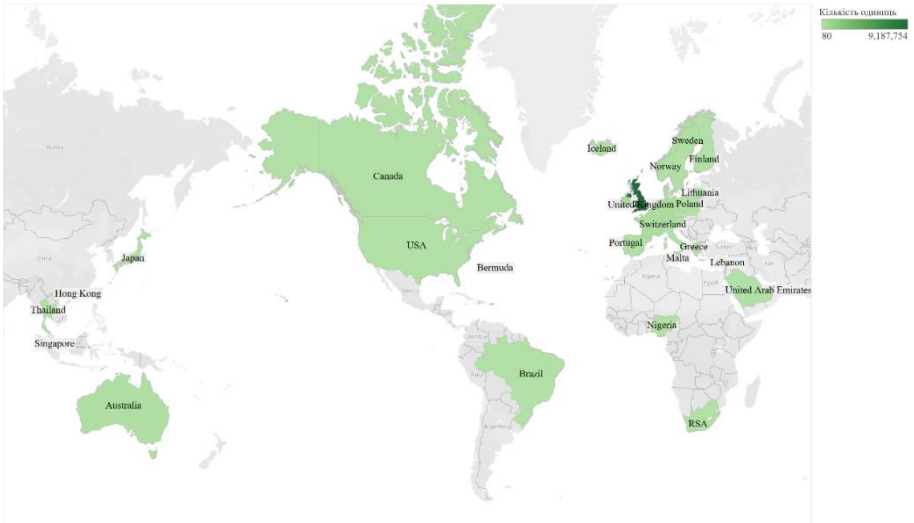

Рис. 1. Мапа кількості проданих одиниць

Наступним кроком була очистка та фільтрація даних. З набору даних було видалено: повернення товару, записи із боргами, комісія банку, пошти та інтернет-сервісів, ручні введення продажів та корегування, записи зі знижками, товари з ціною менше 0.5 фунтів стерлінгів.

Для візуального аналізу та детектування аномалій було проведено агрегування кількості проданих одиниць по дням замовлення (рис. 2).

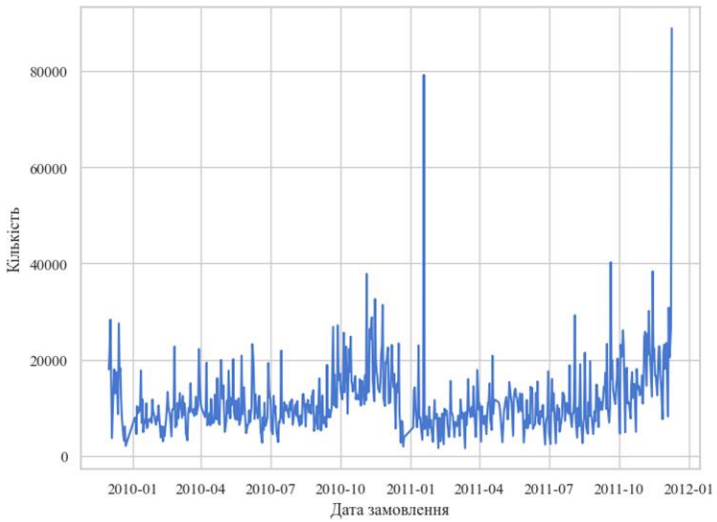

Рис. 2. Динаміка проданих одиниць до детектування аномалій

З візуального аналізу графіку на рис. 2 можна зробити висновок про наявність декількох аномальних піків у даних, які можуть викривити результати дослідження. Для корегування цих рядків у наборі даних було застосовано формулу на основі інтерквартильного розмаху (IQR) із використанням 5-го та 95-го процентилю:

$$IQR = Q3 - Q1,$$

де, $Q1$ − 5-ий процентиль, $Q3$ −95-ий процентиль.

Тоді нижня ($L1$) та верхня ($L2$) границі даних, які не будуть вважатися аномальними, обчислюються як:

$$L1 = Q1 - 1.5 * IQR,$$
$$L2 = Q3 + 1.5 * IQR.$$

Після розрахунку границь аномальні значення було замінено граничними та побудовано графік кількості проданих одиниць по дням замовлень (рис. 3).

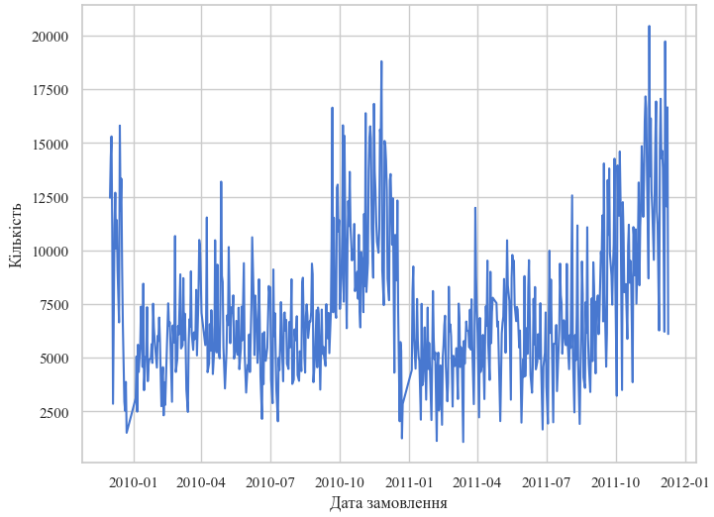

Рис. 3. Динаміка кількості проданих одиниць після детектування аномалій

На основі рис. 3 можна зробити наступні припущення та висновки: у часовому ряді наявні трендові складові, чітко простежуються сезонні та циклічні складові, відслідковуються різкі піки та падіння.

На початку графіка можна спостерігати високу волатильність у даних з частими стрибками та падіннями до квітня 2010 року. Від середини 2010 року орієнтовано до липня 2011 року кількість замовлень більш стабільна, хоча також є помітні піки і спади. З серпня 2011 року до кінця графіка кількість замовлень починає поступово зростати з новими високими піками до кінця 2011 року, що може свідчити про певний сезонний або ринковий фактор.

Найвищий пік продажів припадає на листопад 2011 року, що може бути пов'язане з підготовкою до святкових періодів, таких як Різдво чи Новий рік. Можна зробити припущення щодо присутності подібних піків наприкінці року в листопаді чи грудні, оскільки цей період зазвичай приносить високий попит на іграшки та подарунки.

Різкі падіння та зростання можуть бути пов'язані з економічними або ринковими змінами, рекламою або зовнішніми факторами. Виходячи з візуального аналізу даних, для прогнозування рекомендовано використовувати методи, які будуть враховувати сезонність та наявність трендів. Прикладами таких методів є *SARIMA* (*Seasonal ARIMA*), *Holt-Winters*, *XGBoost*, *LSTM* (*Long Short-Term Memory*, нейронні мережі) [7], *Facebook Prophet* [19].

Кореляційний аналіз та аналіз попиту

Наступним кроком буде проведення кореляційного аналізу змінних за допомогою коефіцієнта кореляції Спірмана, який використовується у припущенні нелінійного зв’язку між змінними. На рис. 4 зображено теплову карту коефіцієнтів кореляції між змінними: «Ціна», «Кількість» «Дохід». Також було перевірено статистичну значущість отриманих коефіцієнтів – усі коефіцієнти є статично значущими на рівні 5%.

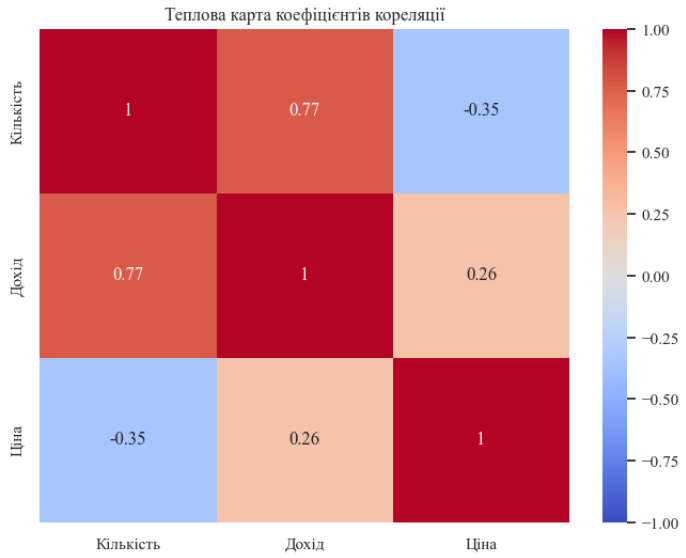

Рис. 4. Теплова карта коефіцієнтів кореляції

Можна зробити наступні висновки: основним рушієм доходу є кількість проданих одиниць – сильна позитивна кореляція між кількістю та доходом це підтверджує. Наявний негативний зв'язок між ціною та кількістю: зі збільшенням ціни кількість проданих одиниць зменшується, вищі ціни можуть стримувати попит на продукцію. Встановлено позитивний зв'язок між ціною та доходом: хоча високі ціни зменшують кількість, вони збільшують загальний дохід через більший внесок кожної одиниці товару.

Кластеризація даних

Для більш детального вивчення даних з 2010 по 2011 роки було проведено кластеризацію даних за ознаками: логарифмована двічі «Ціна», логарифмовані «Кількість» та «Дохід» за допомогою методу *k-means* [17]. Для вибору кількості кластерів було використано метод Ліктя [9]. Оптимальною кількістю кластерів стали 3 кластери.

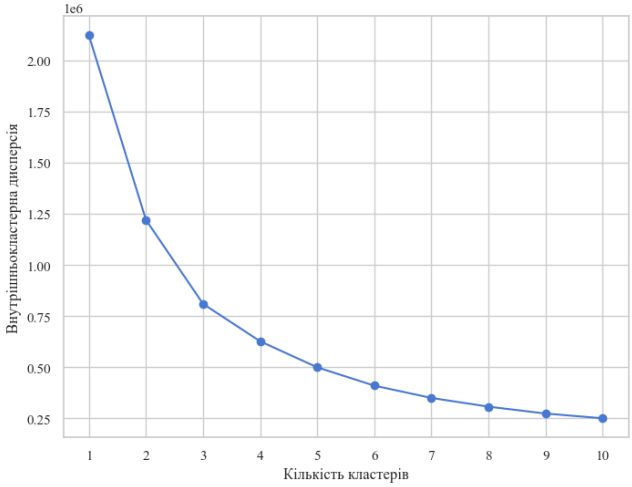

Рис. 5. Внутрішньокластерна дисперсія для різної кількості кластерів

На рис. 6 зображено результати кластеризації даних. Трактувати кластери можна наступним чином: кластер «0» характеризується середніми значеннями ціни та кількості, що в сукупності забезпечує середній дохід. Це товари з оптимальним балансом ціни та кількості, що забезпечують стабільний продаж. Кластер «1» характеризується найменшими значеннями у кількості. Це товари, які продаються у невеликих кількостях за низькою ціною, що призводить до низького доходу. Кластер «2» має найвищі значення як по кількості, так і по ціні. Це товари, які продаються у великих обсягах із високою ціною, що призводить до найбільшого доходу.

Графік на рис. 6 демонструє чітке групування товарів за кластерами, що може свідчити про існування окремих категорій продуктів (наприклад, дешеві, середньої вартості та дорогі продукти).

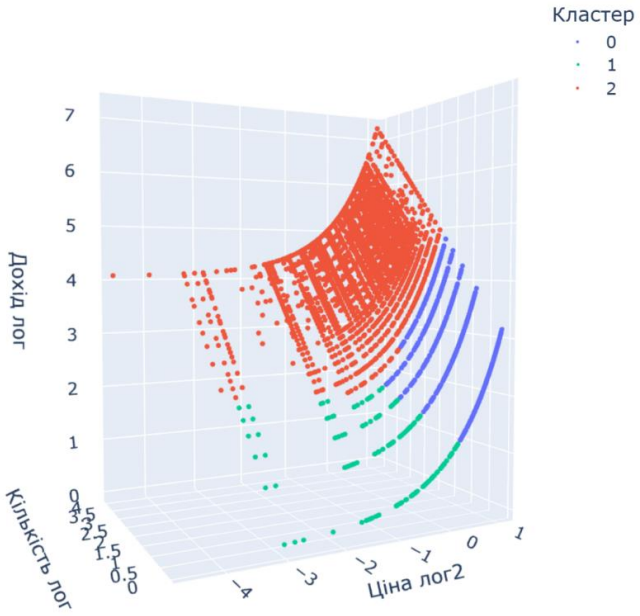

Рис. 6. Кластеризовані дані

Постановка задачі

Метою дослідження є, за допомогою обраних інструментів провести розвідувальний аналіз набору даних для підготовки до вирішення наступних задач:

- Оцінка моделі часового ряду.
- Виявлення трендових, сезонних та циклічних складових часового ряду.
- Кластеризація клієнтів.
- Класифікація нових клієнтів.
- Прогнозування кількості проданих одиниць.

Висновки

Для проведення розвідувального аналізу набору даних було обрано мову програмування *Python* та систему *Tableau*. Попередня обробка даних показала проблеми в даних, які було усунуто за допомогою обраних інструментів. Після побудови мапи продажів для подальшого аналізу було обрано Велику Британію, як країну з найбільшою кількістю проданих одиниць. В даних спостерігається чітка сезонність, наявність трендів та циклів, що буде впливати на подальший вибір моделей для аналізу. Було проведено аналіз попиту та кореляційний аналіз. Також, за допомогою кластерного аналізу було висунуто припущення щодо наявності різних типів товарів. Усі висновки, отримані на основі розвідувального аналізу даних будуть впливати на процеси побудови моделей для вирішення задач, описаних у постановці задачі. Подальші дослідження будуть направлені на побудову та порівняння моделей за допомогою методів машинного навчання.

Література

1. Пінцак І. Використання машинного навчання та аналізу даних для прогнозування тенденцій у електронній комерції. *Information Technology: Computer Science, Software Engineering and Cyber Security*. 2024. № 1. С. 80–88. DOI: 10.32782/it/2024-1-10.

2. Про електронну комерцію: Закон України від 01.01.2024 № 675-VIII. URL: https://zakon.rada.gov.ua/laws/show/675-19#Text (дата звернення: 01.09.2024).

3. Що чекає на український e-commerce у 2024 році: розбираємо ключові тренди? URL: https://rau.ua/novyni/ukr-e-commerce-2024-trendi/ (дата звернення: 10.09.2024).

4. 12 Best Machine Learning Strategies for E-commerce Businesses. URL: https://www.prefixbox.com/blog/machine-learning-for-ecommerce/ (дата звернення: 24.09.2024).

5. Apache Superset. The Apache Software Foundation. URL: https://superset.apache.org/ (дата звернення: 01.10.2024).

6. Chen D., Sain S. L., Guo K. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*. 2012. Vol. 19, no. 3. P. 197–208. URL: https://doi.org/10.1057/dbm.2012.17.

7. García-Aroca C. et al. An algorithm for automatic selection and combination of forecast models. *Expert Systems with Applications*. 2024. 121636. DOI: 10.1016/j.eswa.2023.121636.

8. How Ukrainian eCommerce Survived 2023. Annual Indicators & Forecast 2024. URL: https://www.promodo.com/research/ukrainian-ecommerce-2023#obsyag-ukrayinskogo-rinku-2023 (дата звернення: 11.09.2024).

9. Syakur M.A. et al. Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*. 2018. Т. 336. 012017. DOI: 10.1088/1757-899x/336/1/012017.

10. Looker Studio Overview. URL: https://lookerstudio.google.com/ (дата звернення: 01.10.2024).

11. Matplotlib – Visualization with Python. URL: https://matplotlib.org/ (дата звернення: 02.10.2024).

12. NumPy. The fundamental package for scientific computing with Python. URL: https://numpy.org/ (дата звернення: 02.10.2024).

13. Pandas. Python Data Analysis Library. URL: https://pandas.pydata.org/ (дата звернення: 02.10.2024).

14. Power BI. Uncover powerful insights and turn them into impact. URL: https://www.microsoft.com/en-us/power-platform/products/power-bi (дата звернення: 01.10.2024).

15. Scikit-learn: machine learning in Python. URL: https://scikit-learn.org/stable/ (дата звернення: 03.10.2024).

16. Seaborn: statistical data visualization. URL: https://seaborn.pydata.org/ (дата звернення: 03.10.2024).

17. Sinaga K. P., Yang M.-S. Unsupervised K-Means Clustering Algorithm. *IEEE Access*. 2020. Т. 8. С. 80716–80727. DOI: 10.1109/access.2020.2988796.

18. Tableau: Business Intelligence and Analytics Software. URL: https://www.tableau.com/ (дата звер-нення: 01.10.2024).

19. Taylor S. J., Letham B. Forecasting at Scale. *The American Statistician*. 2018. Т. 72, № 1. С. 37–45. DOI: 10.1080/00031305.2017.1380080.

Сулейманова С.Р.

РОЗВІДУВАЛЬНИЙ АНАЛІЗ ТА ВІЗУАЛІЗАЦІЯ НАБОРУ ДАНИХ НА ПРИКЛАДІ ПІДПРИЄМСТВА ЕЛЕКТРОННОЇ ТОРГІВЛІ

У цій статті представлений підхід до розвідувального аналізу та візуалізації набору даних на прикладі підприємства електронної торгівлі. В дослідженні розглядаються ключові етапи розвідувального аналізу даних: попередню обробку даних, візуалізацію, видалення аномалій, кореляційний та кластерний аналізи для підготовки даних до вирішення задач машинного навчання у майбутніх дослідженнях. До цих задач входять: оцінка моделі часового ряду, виявлення тренду, сезонних та циклічних складових часового ряду, кластеризація клієнтів, класифікація нових клієнтів, прогнозування кількості проданих одиниць по кластерам клієнтів. Розроблений підхід може бути використаний для аналізу інших наборів даних електронної торгівлі.

***Ключові слова**: розвідувальний аналіз; візуалізація даних; часові ряди; кореляційний аналіз; кластерний аналіз; машинне навчання.*

Suleimanova S.R.

EXPLORATORY DATA ANALYSIS AND VISUALIZATION ON THE EXAMPLE OF AN E-COMMERCE ENTERPRISE

This article presents an approach to exploratory data analysis and visualization using the example of an e-commerce company. The study examines key stages of exploratory data analysis, including data preprocessing, visualization, anomaly detection, correlation analysis, and cluster analysis, aimed at preparing data for solving machine learning tasks in future research. These tasks include estimating a time series model, identifying trends, seasonal and cyclical components of time series, customer clustering, new customer classification, and predicting the quantity of items sold within customer clusters. The proposed approach can be applied to the analysis of other e-commerce datasets.

***Keywords**: exploratory data analysis; data visualization; time series; correlation analysis; cluster analysis; machine learning.*