УДК 004.65

АЛГОРИТМІЧНИЙ ПІДХІД ДО АВТОМАТИЗОВАНОГО ПОШУКУ ІНСАЙТІВ У СТРУКТУРОВАНИХ ДАНИХ

Дідковська В.К.[1]; Платонов В.В.[2]; Платонова Є.В.[3]
[1,2,3]Одеський національний університет імені І.І. Мечникова
[1,2,3] Україна, Одеса
[1]viktoriia.didkovska@stud.onu.edu.ua;[2]v.platonov@onu.edu.ua;[3]jane.platonova@onu.edu.ua

Анотація. *У роботі запропоновано та обґрунтовано комплексний підхід для автоматизованого пошуку аналітичних «історій» (інсайтів) у структурованих даних, що виходить за межі класичного виявлення викидів. Метод поєднує статистичний аналіз та алгоритми машинного навчання, зокрема Isolation Forest, для ідентифікації аномальних значень у категоріальних ознаках та визначення факторів, що їх пояснюють. Підхід апробовано на відкритому датасеті Glassdoor, де він продемонстрував здатність автоматично генерувати нетривіальні, інтерпретовані бізнес-висновки.*

Ключові слова: *пошук інсайтів; аналіз даних; категоріальні дані; виявлення аномалій; алгоритм Isolation Forest; автоматизація звітності.*

Вступна частина. Сучасний бізнес-аналіз вимагає від спеціалістів не лише фіксації статистичних фактів, таких як викиди чи відхилення метрик, а й формування цілісних аналітичних «історій» з практичними рекомендаціями. Традиційні статистичні інструменти часто не дають глибокого пояснення причин аномалій. Відтак, актуальною задачею є розробка алгоритмічних підходів, що автоматизують процес виявлення значущих патернів та їх контекстуалізацію для бізнес-користувачів.

Метою є створення та апробація схеми аналізу, яка забезпечує:

1. Автоматизоване виявлення статистично значущих відхилень у підгрупах даних відносно бінарної цільової змінної.
2. Ідентифікацію ключових ознак, що пояснюють природу цих відхилень.
3. Автоматичне формування структурованого текстового звіту з виявленими інсайтами для подальшого використання.

Основна частина. Запропонований метод складається з двох ключових етапів.

Етап 1: Пошук аномальних значень категоріальних ознак. Для кожної категорії v ознаки X обчислюється відхилення Δv частки позитивних спостережень цільової змінної від загального середнього. Далі, до вектору відхилень $\{\Delta v\}v$ застосовується алгоритм Isolation Forest для ідентифікації тих категорій, чиє відхилення є статистично аномальним. Цей підхід дозволяє виявляти не просто викиди, а саме ті групи даних, поведінка яких суттєво відрізняється від загальної тенденції.

Етап 2: Ідентифікація пояснюючих ознак. Після виявлення аномальної підмножини A (наприклад, компанії в певній галузі), для кожної іншої ознаки Z аналізується, чи її розподіл в межах A значуще відрізняється від її загального розподілу в усьому датасеті. Якщо різниця часток $|q_A(z) - q_{total}(z)|$ перевищує заданий поріг θ, ознака $Z=z$ вважається такою, що пояснює аномалію.

Практична реалізація підходу втілена у двох програмних функціях: find_abnormal_categorical та describe_sel. Наприклад, при аналізі датасету «Титанік», де

цільовою змінною є виживання пасажирів, функція describe_sel для аномальної категорії «Стать = жінки» автоматично генерує наступний текстовий звіт:

Підмножина «Стать = жінки» складає 35.2% від загальної вибірки. У цій групі показник виживання становить 74.2%, що на 91% вище, ніж у середньому по датасету (38.8%). Значущі особливості: - Клас квитка = 1: частка у вибірці 42% проти 24% у загальній. - Порт посадки = Cherbourg: частка у вибірці 34% проти 23% у загальній.

Апробація методу. Для демонстрації ефективності було використано датасет «Glassdoor Company Insights», що містить дані про ~10 тис. компаній США з такими категоріальними ознаками, як галузь, розмір компанії та локація. Цільовою змінною було визначено бінарну ознаку «має високий рейтинг» (рейтинг > 4.0).

Алгоритм автоматично виявив галузеві аномалії. Наприклад, сектори «Коледжі та університети» та «Розробка програмного забезпечення» були ідентифіковані як аномально успішні, з часткою компаній з високим рейтингом ~53% та ~47% відповідно. На противагу, галузі «Ресторани та кафе» та «Універмаги» показали аномально низькі результати (лише 8-9% та 7% компаній з високим рейтингом).

Додатково було введено метрику «Коефіцієнт найму» (Hiring Ratio) — відношення кількості вакансій до кількості співробітників, що дозволило виявити компанії, які знаходяться у фазі стрімкого зростання.

Висновки. Запропонований підхід довів свою практичну цінність. Його новизна полягає у комбінації методу Isolation Forest з аналізом співвідношень у категоріях, що дозволяє виявляти неочевидні закономірності. Метод є предметно-незалежним і може бути застосований для аналізу клієнтів, продуктів, транзакцій у будь-якій галузі. Автоматична генерація текстових інсайтів скорочує час на рутинний аналіз та відповідає сучасним трендам доповненої аналітики, де звітність стає все більш автоматизованою.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Joyshil, M. (2023). *USA Company Insights: Glassdoor Scraped Data 2023*. Kaggle. Retrieved from https://www.kaggle.com/datasets/joyshil0599/glassdoor-company-insightsscraped-data-collection

[2] Microsoft Learn. (2025, January 31). *Anomaly detection in Power BI*. Retrieved from https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-anomaly-detection

[3] Zhe, D. (2018, January). *Automating Anomaly Detection*. Retrieved from https://dongyuzheng.com/files/Automating_Anomaly_Detection.pdf

Didkovska V., Platonov V., Platonova Y.

An algorithmic approach for automated insights finding in structured data

Abstract. *This paper proposes and substantiates a comprehensive approach for the automated discovery of analytical "stories" (insights) in structured data, moving beyond classical outlier detection. The method combines statistical analysis and machine learning algorithms, particularly Isolation Forest, to identify abnormal values in categorical features and determine their explanatory factors. The approach was tested on the public Glassdoor dataset, where it demonstrated the ability to automatically generate non-trivial, interpretable business conclusions.*

Keywords: *insight finding; data analysis; categorical data; anomaly detection; Isolation Forest algorithm; reporting automation.*