Одеський національний університет імені І. І. Мечникова
Факультет математики, фізики та інформаційних технологій
Кафедра оптимального керування та економічної кібернетики

Кваліфікаційна робота

на здобуття ступеня вищої освіти «бакалавр»

«Пошук історій у даних: методи та практичні приклади»
«Finding insights in data: methods and practical examples»

Виконала: здобувачка денної форми навчання
спеціальності 113 Прикладна математика
Освітня програма «Прикладна математика»

Дідковська Вікторія Костянтинівна

Керівник ст. викл. Платонов В.В.
(науковий ступінь, вчене звання, прізвище та ініціали, підпис)
Рецензент
кандидат фіз.-мат. наук, доцент, Страхов Є.М.
(науковий ступінь, вчене звання, прізвище та ініціали)

Рекомендовано до захисту:
Протокол засідання кафедри
№ ___ від _____2025 р.

Завідувач кафедри

_____ _____
(підпис) (прізвище, ініціали)

Захищено на засіданні ЕК №_____
протокол № ____ від _____ 2025 р.
Оцінка_____/___ ___/_____
(за національною шкалою, шкалою ECTS, бали)

Голова ЕК

_____ _____
(підпис) (прізвище, ініціали)

Одеса – 2025

ЗМІСТ

ВСТУП

Останнім часом традиційні статистичні підходи до виявлення розподілів даних та локалізації викидів дедалі більше показують свої обмеження в контексті формулювання практичних рекомендацій для бізнес-замовника. Хоча класичні методи – такі як тестування гіпотез, використання гістограм і яскравих статистичних метрик – залишаються важливими та невід'ємними інструментами, вони часто фіксують лише факт відхилення без глибокого пояснення його походження чи побудови чіткого плану подальших дій. Дедалі частіше у бізнесі будь-якої ніші дійсно мало почути від дата аналітика, що є певний викид у даних або, що користувач "відпадає" на певній сторінці сайту чи при певному етапі реєстрації при аналізі воронок і поведінки користувача. Від спеціаліста хочуть чути реальні припущення і пропозиції до змін. Відтак сучасний аналіз даних переходить до розробки цілісних аналітичних «історій/інсайтів» (insights), що поєднують виявлення аномалій із їх контекстуалізацією й практичною інтерпретацією. Їхній пошук та алгоритм до цього, власне, і буде описано у цій роботі.

На початку, хочу окреслити короткий "helicopter view" для занурення у контекст. Якісний аналітичний інсайт починається з детектування аномалії, яке реалізується за допомогою порівняння локальних метрик (наприклад, рівня конверсії в окремих сегментах клієнтів, показників ретеншену користувачів чи об'єму транзакцій) із відповідними глобальними значеннями. Такий підхід дозволяє виділити статистично значущі відхилення з урахуванням варіаційних особливостей кожного сегменту даних і звести до мінімуму хибні спрацьовування. Наступним кроком є розміщення в контексті виявлених відхилень шляхом аналізу супутніх ознак: демографічних (вік, стать, регіон), тимчасових (день тижня, сезонні коливання) або поведінкових (кількість сесій, глибина

перегляду). Цей етап забезпечує виявлення потенційних факторів до пояснення та спрямовує гіпотези щодо причин появи аномалії. Практична інтерпретація формулює конкретні рекомендації для бізнесу, оцінюючи їх ймовірний вплив на ключові показники ефективності (KPI): зростання прибутковості, скорочення відтоку клієнтів або оптимізацію операційних витрат. За необхідності розробляється кількісна модель оцінки очікуваного ефекту впровадження запропонованих заходів.

Метою даної роботи є створення схеми аналізу, яка буде забезпечувати:

1. Автоматизоване виявлення відхилень за обраною бінарною цільовою змінною.
2. Ідентифікацію ключових пояснювальних ознак для глибокого розуміння механізмів виникнення виявлених патернів.
3. Формулювання текстового звіту – «аналітичного інсайту», для представлення перед бізнес-замовником з конкретними рекомендаціями.

Щоб продемонструвати універсальність та практичну цінність запропонованого підходу буде здійснено аналіз декількох публічно доступних датасетів. Завдяки цьому забезпечимо прозорість і відтворюваність усіх етапів – від попередньої обробки до формування кінцевих бізнес-рекомендацій.

РОЗДІЛ 1

ЗАГАЛЬНА МЕТОДИКА ФОРМУЛЮВАННЯ “ІНСАЙТІВ”

1.1 Визначення цільової метрики

Нехай є цільова бінарна змінна $Y \in \{0,1\}$ (наприклад, факт покупки білету або виживання в датасеті Titanic, про який поговоримо пізніше). Для будь-якої підмножини даних S обчислюють частку позитивних спостережень:

$$p_S = \frac{\sum_{i \in S} Y_i}{|S|},$$

яку порівнюють із загальною часткою у всьому наборі:

$$p_{\text{total}} = \frac{\sum_{i=1}^{N} Y_i}{N},$$

Якщо $|p_S - p_{\text{total}}| \geq \delta$, (де δ — обраний поріг, наприклад 0.2), підмножина S вважається аномальною.

1.2 Пошук аномальних значень категоріальних ознак

1. Для кожного значення v категоріальної ознаки X формують підмножину $S_v = \{\, i \mid X_i = v\}$.
2. Розраховують частку позитивних спостережень у цій підмножині:
   $$p_{S_v} = \frac{\sum_{i \in S_v} Y_i}{|S_v|}$$
3. Вираховують різницю з глобальною часткою:
   $\Delta_v = p_{S_v} - p_{\text{total}}$.

4. Застосовують алгоритм виявлення аномалій (наприклад, Isolation Forest) до вектору $\{\Delta_v\}_v$.
5. Відбирають ті категорії v, для яких модель позначила Δ_v як аномальне (мітка -1 в Isolation Forest).

1.3 Ідентифікація пояснюючих ознак

Після визначення аномальних категорій або інтервалів, треба знайти інші ознаки, що статистично відрізняються в цих підмножинах від загального масиву. Алгоритм побудований за такою схемою:

1. Нехай A — множина індексів, де спостерігається аномалія за ознакою X.
2. Для кожної іншої ознаки Z та для кожного її можливого значення z:

   - обчислити частку

   $$q_A(z) = \frac{\#\{\, i \in A \mid Z_i = z\}}{|A|},$$

   - обчислити частку у всьому датасеті

   $$q_{\text{total}}(z) = \frac{\#\{\, i \mid Z_i = z\}}{N}.$$

3. Якщо $|q_A(z) - q_{\text{total}}(z)| \geq \theta$ (поріг, наприклад 0.2), значення z вважається значущою особливістю, яка пояснює аномалію $X = v$.

РОЗДІЛ 2

ҐРУНТ ДО РОБОТИ

Основним поштовхом до роботи та ідеєю її створення стала стаття про пошук інсайтів, автором якої є Платонов Віталій Валерійович, старший викладач кафедри оптимального керування та економічної кібернетики ОНУ імені І. І. Мечникова **(ДОДАТОК 1)**

2.1 Детальний опис алгоритмічних компонентів коду

Нижче представлено опис двох ключових функцій, реалізованих у статті. Ці функції складають основу схеми автоматизованого пошуку інсайтів у даних.

2.1.1 Функція find_abnormal_categorical(df, col, target)

1.1. Завдання та загальна ідея

Метою **find_abnormal_categorical** є виявлення таких значень категоріальної ознаки **col**, для яких розподіл бінарної цільової змінної **target** суттєво відрізняється від загального розподілу. Ідея полягає в тому, щоб спочатку зібрати статистичну інформацію про кожну категорію (**Value**) у контексті частки позитивних випадків (наприклад, виживання на «Титаніку»), а потім за допомогою алгоритму виявлення аномалій визначити ті категорії, які «виглядають» аномально порівняно з іншими.

1.2. Розбір кроків реалізації

1. **Початковий фільтр**
    - Якщо аналізована ознака **col** ідентична цільовій змінній **target**, функція одразу повертає порожній список —

немає сенсу шукати аномалії в самій цільовій ознаці.

- Далі перевіряється, чи містить стовпець принаймні два унікальні значення; за відсутності—пакет статистичних тестів та алгоритмів не мають на чому працювати.

2. **Побудова зведеної статистики**
   - Виконується групування **df.groupby([col, target]).size()**, щоб для кожного значення **col** отримати два підрахунки: число спостережень, де **target=True**, та де **target=False**.
   - Методом **pivot** перетворюється результат у таблицю, в якій рядком відповідає одне значення **col**, а стовпцями — два лічильники. Порожні комірки заповнюються нулями.
   - Обчислюють загальну поточну кількість спостережень для кожного значення **col** (**size = sum(True, False)**) та відношення позитивних до негативних випадків (**ratio = True / False**). Цей коефіцієнт є індикатором «відхилення» категорії.

3. **Відсіювання «шумових» категорій**
   - Видаляються ті значення **col**, які трапляються надто рідко — менше ніж у 1 % всіх спостережень. Це необхідно, оскільки вкрай невелика вибірка може давати статистично нестабільні або некорисні для бізнесу патерни.

4. **Запуск алгоритму Isolation Forest**
   - Ініціалізується **IsolationForest(max_samples=100)**, оскільки наша задача полягає в одновимірному виявленні аномалій на основі метрики **ratio**.

- Підготовлений вектор коефіцієнтів **ratio** передається в **clf.fit()**, після чого **clf.predict()** маркує кожну категорію як нормальну (**1**) або аномальну (**–1**).

- Повертається список тих значень **col**, які отримали мітку **–1**.

1.3. Примітки та можливі розширення

- Коли **False == 0**, **ratio** може вийти нескінченним. У коді значення **np.inf** замінюється на найбільше ціле, щоб уникнути збоїв при подальшому навчанні.
- Поріг у 1 % для мінімальної частоти можна зробити параметром, щоб у спеціалізованих задачах виявляти рідкісні, але важливі для бізнесу категорії.
- Хоча зараз функція працює лише з бінарною **target**, її можна узагальнити, використавши one-vs-rest підхід або аналіз ентропії в багатокласовому випадку.

2.1.2 Функція describe_sel(df, sel, target, sel_var, truetargetname, threshold_variables=0.2, number_of_bins=20)

2.1. Завдання та загальна ідея

Коли виявлено аномальну підмножину **sel** (де **sel_var = конкретне значення**), потрібно з’ясувати, які інші ознаки в цій вибірці «пояснюють» її відмінність від загальних даних. Функція **describe_sel** автоматично складає текстовий опис із базовими статистиками та переліком значущих ознак, що мають суттєве відхилення у частці прояву.

2.2. Детальний розбір алгоритму

1. Підготовка та базова статистика

   - Визначаємо розмір **sel** та його питому вагу відносно всього **df**, оформлюємо результат у відсотках: $\frac{|sel|}{|df|}$

   - Розраховуємо загальну частку позитивних спостережень $p_{\text{total}} = \frac{\sum df[\text{target}]}{|df|}$

     та частку в **sel**

     $p_{\text{sel}} = \frac{\sum sel[\text{target}]}{|sel|}$

   - Додаємо до тексту відсоток відхилення **(higher/lower)**.

2. Перевірка значущості вибірки

   - Якщо абсолютна різниця $|p_{\text{sel}} - p_{\text{total}}|$ виявляється меншою за 20 % від **p_total**, детальний аналіз інших ознак припиняється, оскільки вибірка недостатньо аномальна.

3. Аналіз кожної ознаки

   - Для стовпців з кількістю унікальних значень ≤ 50 або типом **object** ознака вважається категоріальною. Інакше створюється від 1 до **number_of_bins** бінів за допомогою **pd.qcut**.

- Для кожного можливого значення z (категорії або інтервалу) обчислюють:
  - частку $\text{Share}_1 = \frac{\#\{\, i \in \text{sel} \mid Z_i = z\}}{|\text{sel}|}$
  - частку $\text{Share}_2 = \frac{\#\{\, i \in df \mid Z_i = z\}}{|df|}$

Якщо $|\text{Share}_1 - \text{Share}_2| \geq \text{threshold\_variables}$, значення z вважається значущим і описується в тексті:

Feature = z

Share = X% in selection vs Y% overall

- При цьому для числових ознак тимчасово додається стовпець із біном **binind**, який видаляється після обчислень, щоб не змінювати оригінальний **sel**.

4. Формування підсумкового тексту

- У накопиченій змінній **res** послідовно додаються блоки: загальні статистики, різниця у **target**, а потім пункти з значущими характеристиками.

- Результатом є докладний опис вибірки, готовий для безпосереднього включення в текст аналітичного звіту або презентацію.

2.3. Особливості та рекомендації

- Значення **threshold_variables = 0.2** відповідає мінімуму 20 % абсолютної різниці в частках. Його можна коригувати залежно від задачі: для більш чутливого аналізу зменшити, для консервативного — збільшити.

- При роботі з великими датасетами доцільно попередньо знижувати розрядність даних або використовувати методи семплінгу, щоб уникнути надмірного споживання ресурсів під час групувань і бінування.

- Можна доповнити **describe_sel** частиною коду, що автоматично генерує прості візуалізації (гістограми, bar-діаграми) у SVG чи matplotlib, і вставляє посилання на зображення прямо в текстовий звіт.

Тож, докладний опис кожного з етапів функцій **find_abnormal_categorical** та **describe_sel** демонструє їхню внутрішню логіку, а також важливі параметри налаштування та можливі напрямки розвитку для більш гнучкого й масштабованого аналізу даних у контексті автоматичного формування інсайтів.

РОЗДІЛ 3

ПЕРЕВІРКА НА ТЕСТОВОМУ НАБОРІ “ТИТАНІК”

3.1 Постановка задачі класифікації

Задача полягає у прогнозуванні виживання пасажирів корабля «Титанік» на основі датасету із 891 записом про пасажирів. Залежна змінна Survived набуває значень 1 (вижив) або 0 (не вижив). Це класична задача бінарної класифікації, метою якої є побудова моделі, здатної за характеристиками пасажира (стать, вік, клас квитка, тариф, порт посадки тощо) передбачити його ймовірність вижити. Код програми можна знайти у **ДОДАТКУ 2**.

3.2 Підготовка даних

1. Видаляємо стовпці, які не є інформативними: PassengerId, Name, Ticket (не несуть корисної практичної інформації для моделі).
2. Обробляємо пропуски:
    - Cabin було видалено через >75% пропусків.
    - Age та Embarked: записи з пропущеними значеннями було видалено (отримано 712 повних записів).
3. Кодуємо категоріальні змінні:
    - Sex закодовано як бінарна ознака (female=1, male=0).
    - Embarked перетворено на dummy-представлення (S, C, Q).
4. Перевірка шкалювання: значення ознак Age і Fare знаходяться приблизно в однаковому діапазоні, масштабування не застосовувалося.

3.3 Пошук інсайтів

Щоб глибше зрозуміти структури даних та ключових чинників, що впливають на виживання, виконано дослідження даних (EDA) і пошук інсайтів:

1. Кореляційний аналіз: побудували кореляційну матрицю між числовими та бінарними ознаками і цільовою змінною Survived (чи вижила особа).
    - Survived найбільше корелює із Sex_female (≈0.54), що говорить про більший шанс виживання у жінок
    - Значення змінної Pclass має найбільшу від'ємну кореляцію з фактом виживання (≈ –0.34), що свідчить про зниження шансів виживання у пасажирів нижчого класу.
    - Fare показав помірну позитивну кореляцію (≈0.26), що вказує на те, що вища вартість квитка пов'язана з кращими шансами вижити.
2. Груповий аналіз: обчислено частку вижившх для кластерів за категоріями:
    - Стать: жінки виживали в ~74%, чоловіки – у ~19%.
    - Клас квитка: 1-й клас – 63%, 2-й – 47%, 3-й – лише 24% виживаність.
    - Порт посадки: пасажири, що сіли в Cherbourg (C), виживали в ~61% випадків, тоді як з Southhampton (S) – у 34%, з Queenstown (Q) – у 37%.
3. Розподіл за віком та сімейним статусом:
    - Пасажири віком до 16 років виживали в ~60% випадків (діти мали пріоритет).
    - Особи з великим числом супутників ($SibSp + Parch \geq 4$) показали знижену виживаність (~20%), тоді як подорожуючі наодинці ($SibSp = 0$ та $Parch = 0$) – ~38%.

4. Виявлені нетипові випадки (аномалії):
    - Декілька чоловіків 3-го класу, але з високою вартістю квитка (>100), вижили всупереч загальній тенденції.
    - Невелика група жінок 3-го класу без супутників мала нижчі шанси вижити (~40%) порівняно з жінками 1-го класу (понад 90%).

Ключові інсайти:

- Стать і клас квитка є найбільш сильними показниками виживання.
- Вартість квитка та порт посадки слугують додатковими індикаторами соціально-економічного статусу.
- Діти та невеликі родини мали вищі шанси вижити, в той час як занадто великі сім’ї — нижчі.

РОЗДІЛ 4

ЗАСТОСУВАННЯ МЕТОДУ ПОШУКУ ІНСАЙТІВ ДО МИНУЛОРІЧНОЇ (КУРСОВОЇ) РОБОТИ

У минулому році я займалася курсовою роботою за темою “Профілактика банкрутства”, метою якої було виявлення ключових факторів, що впливають на фінансові результати компаній і ризик їхнього банкрутства. Для досягнення цієї мети було проведено огляд існуючих методів аналізу фінансової стійкості (Модель Е. Альтмана та У. Бівера), описано процес збору та підготовки даних, а також розроблено та протестовано моделі машинного навчання на реальних фінансових даних компаній. У результаті було випрацювано ефективний механізм оцінки ймовірності банкрутства компанії. У цьому році було прийнято рішення спробувати знайти корисні інсайти та приховані закономірності у датасеті із курсової роботи для виявлення загальних тенденцій у поведінці фінансових показників успішних і неуспішних компаній. З кодом можна ознайомитися у **ДОДАТКУ 3**. Датасет містить дані американських компаній за період 1999–2018 років. Кожен запис відповідає фінансовим показникам певної компанії за конкретний рік, разом із міткою status_label, що вказує на статус компанії – «alive» (діюча) або «failed» (збанкрутіла). Загальний обсяг вибірки – 78 682 записи, які охоплюють тисячі окремих компаній різного масштабу. Частка компаній, що зазнали невдачі (банкрутства), відносно невелика – лише близько 6,6% записів мають мітку *failed* (решта ~93,4% – *alive*). Така диспропорція свідчить про імбаланс класів у даних, що необхідно враховувати при інтерпретації результатів.

Було проаналізовано широкий набір фінансових показників (усього 18 змінних для кожної компанії за рік), серед яких: загальні активи, сукупна виручка (Total Revenue), чистий дохід (Net Income), EBITDA (прибуток до сплати відсотків, податків і амортизації), валовий прибуток та інші. Також було розраховано додаткові показники, такі як валова маржинальність (Gross Profit / Net Sales) та Z-рахунок Альтмана – інтегральний індекс фінансової стійкості.

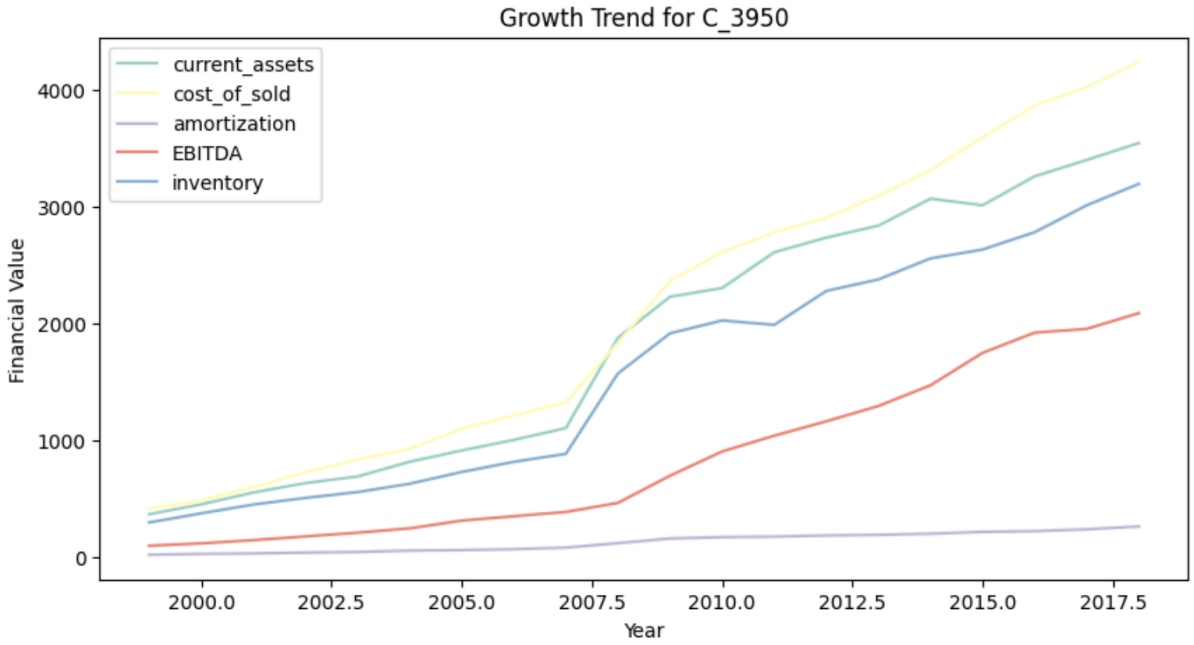
4.1 Методологія аналізу

Аналіз даних здійснювався за допомогою методів Exploratory Data Analysis (EDA). Первинно було виконано описову статистику всіх фінансових змінних: розраховано середні значення, медіани, квартилі, мінімальні та максимальні значення. Це дозволило оцінити масштаб і варіабельність даних, а також виявити наявність значних перекосів розподілів чи екстремальних значень. Для ідентифікації викидів застосовано як класичні статистичні критерії (значення за межами діапазону [Q1–1.5*IQR, Q3+1.5*IQR]), так і алгоритм Isolation Forest. Останній слугував для автоматичного виявлення нетипових спостережень у розподілах фінансових показників, що потенційно можуть спотворювати оцінки або ж самі по собі є предметом інтересу (наприклад, надзвичайно великі збитки або прибутки компаній). Поряд із одновимірним аналізом, проведено кореляційний аналіз між усіма парними фінансовими показниками. Було побудовано матрицю кореляцій Пірсона, щоб визначити, які показники мають сильні взаємозв’язки. Високі коефіцієнти кореляції ($p > 0.7$) можуть свідчити про колінеарність показників або про те, що вони відображають споріднені аспекти діяльності.

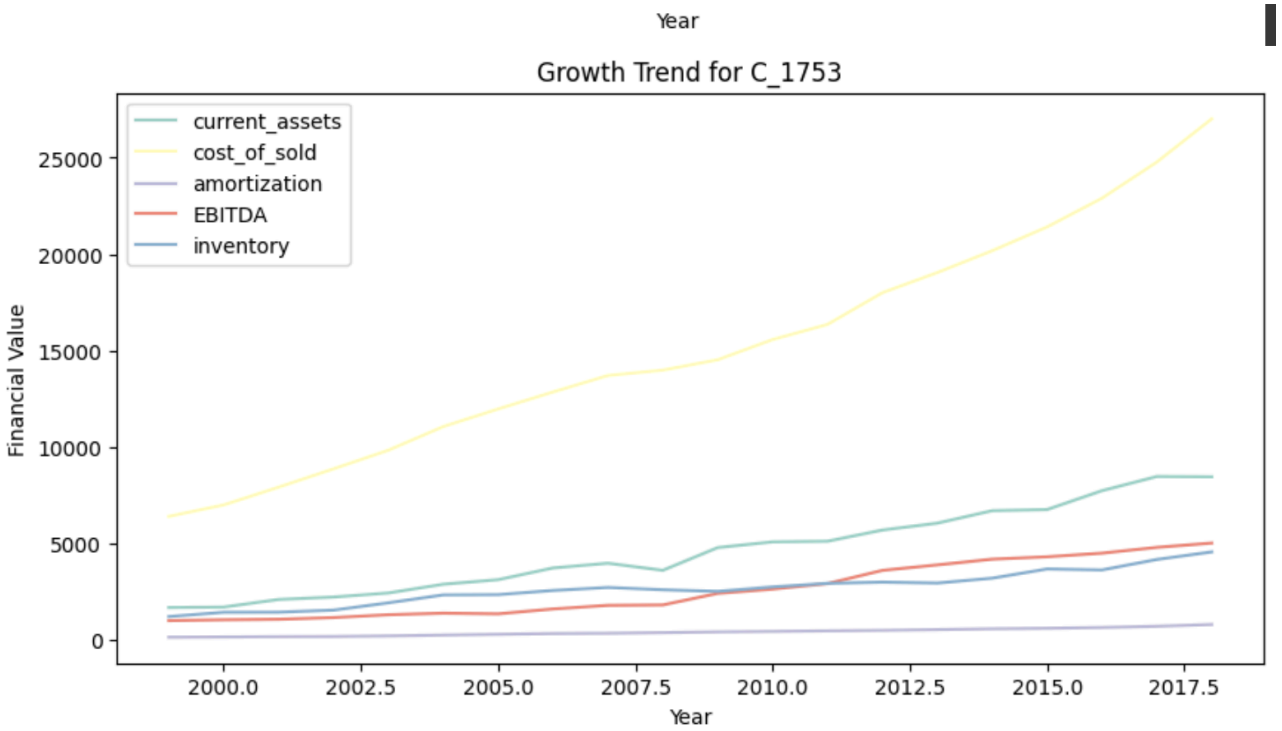
Виявлення таких зв’язків важливе для інтерпретації: наприклад, сильна кореляція між двома показниками може означати, що для пошуку інсайтів достатньо аналізувати один із них. Низькі ж кореляції підкреслюють, що кожен показник дає унікальну інформацію.

Далі застосовано методи сегментаційного аналізу для дослідження впливу окремих факторів на виживання компаній. Зокрема, дані групувалися за принципом «високі проти низьких» значень певного показника з наступним порівнянням частки вижившихкомпаній у кожній групі. Таким чином перевірено гіпотези: чи мають компанії з високою виручкою, великою ринковою вартістю або високою рентабельністю істотно вищі шанси залишитися на ринку. Для прикладу, було вибрано підмножину компаній з загальною виручкою вище медіани та окремо – компанії, ринкова капіталізація яких перевищує 75-й перцентиль (топ-25% за Market Value). Для обох груп обчислено частку живих (не збанкрутілих) компаній і порівняно з середнім рівнем по вибірці. Аналогічно аналізувався вплив валової маржі: вибірка була поділена на компанії з маржинальністю вище та нижче медіани.

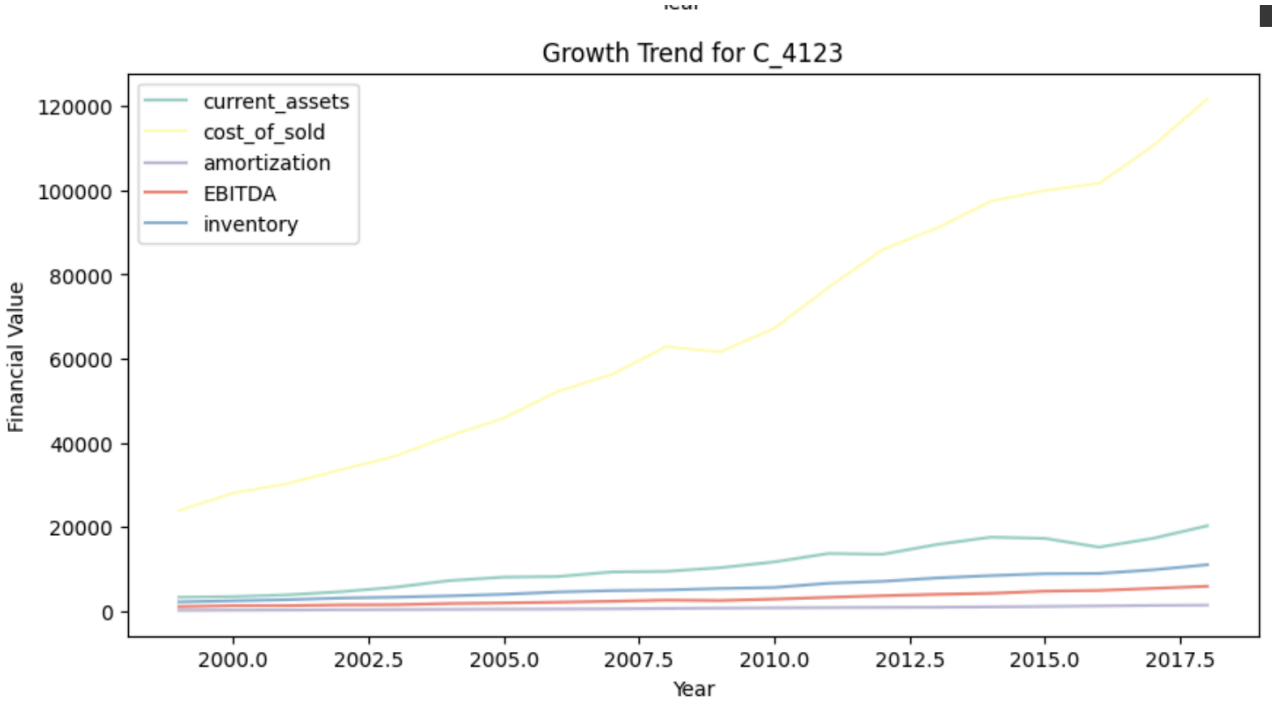
Крім того, досліджено динамічні аспекти фінансових показників. Запроваджено поняття “стійко зростаючих” компаній – тих, що демонструють постійне зростання ключових фінансових метрик протягом кількох років поспіль. Для цього кожну компанію було проаналізовано в розрізі часу: для низки фінансових показників (активи, виручка, прибутки тощо) перевірялося, чи зростали їхні значення з року в рік. Компанія відносилася до стійко зростаючих, якщо принаймні протягом трьох послідовних років спостерігалося невпинне зростання більшості основних показників. Частка таких компаній та їх характеристики (наприклад, середній темп зростання, максимальна тривалість періоду зростання) були зафіксовані. Для ілюстрації тенденцій побудовано графіки динаміки фінансових показників окремих компаній із найбільш стабільним зростанням (рис. 1, рис. 2, рис. 3)



(рис. 1)


Year
Growth Trend for C_1753
current_assets
cost_of_sold
amortization
EBITDA
inventory
Financial Value
25000
20000
15000
10000
5000
0
2000.0
2002.5
2005.0
2007.5
2010.0
2012.5
2015.0
2017.5
Year

(рис. 2)


Growth Trend for C_4123
current_assets
cost_of_sold
amortization
EBITDA
inventory
Financial Value
120000
100000
80000
60000
40000
20000
0
2000.0
2002.5
2005.0
2007.5
2010.0
2012.5
2015.0
2017.5
Year

(рис. 3)

Нарешті, для оцінки ризику банкрутства застосовано Z-рахунок Альтмана – відому модель прогнозування неплатоспроможності. На основі фінансових даних кожного запису розраховано значення Z за формулою Едварда Альтмана, яка комбінує п’ять фінансових коефіцієнтів (ліквідність, прибутковість, левередж та оборотність активів). За отриманим значенням Z всі випадки класифіковано на три зони ризику: *високий ризик* банкрутства ($Z < 1.81$), *середній ризик* (1.81–2.99) та *низький ризик* ($Z \geq 2.99$). Цю класифікацію зіставлено з фактичними мітками alive/failed, щоб оцінити, наскільки моделі фінансових ризиків узгоджуються з реальними результатами.

Таким чином, застосовано комплексний підхід, що поєднує описову статистику, виявлення аномалій, групове порівняння та аналіз існуючих фінансових індикаторів ризику. Нижче наведено ключові результати цього аналізу та обговорення отриманих інсайтів.

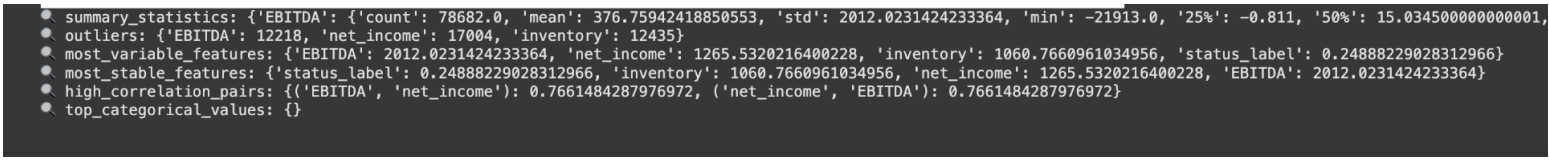
4.2 Результати аналізу та інсайти

Аналіз розподілу фінансових змінних показав, що дані охоплюють надзвичайно широкий діапазон значень, характерний для компаній різного масштабу – від відносно малих підприємств до корпоративних гігантів. Для прикладу, показник *EBITDA* варіюється від значних від’ємних величин до дуже великих додатних. Середнє значення EBITDA становить ~376.8 (умовних одиниць), стандартне відхилення – понад 2000, при цьому мінімальне значення – близько -21 913, а максимальне сягає ~81 730. Подібна картина і з *чистим прибутком (Net Income)*: середня ~129.4, $\sigma \approx 1265$, мінімум глибоко негативний, максимум – дуже високий.

Аналіз розподілу фінансових змінних показав, що дані охоплюють надзвичайно широкий діапазон значень, характерний для компаній різного масштабу – від відносно малих підприємств до корпоративних гігантів. Для прикладу, показник *EBITDA* варіюється від значних від'ємних величин до дуже великих додатних. Середнє значення EBITDA становить ~376.8 (умовних одиниць), стандартне відхилення – понад 2000, при цьому мінімальне значення – близько -21 913, а максимальне сягає ~81 730. Подібна картина і з *чистим прибутком (Net Income)*: середня ~129.4, $\sigma \approx$ 1265, мінімум глибоко негативний, максимум – дуже високий. Той факт, що середні значення значно менші за стандартні відхилення, а розмах величезний, свідчить про сильну асиметрію розподілу та наявність численних екстремальних значень. Іншими словами, більшість компаній мають відносно помірні показники прибутковості, тоді як декілька випадків – надзвичайно великі (або негативні) значення, які суттєво підвищують середнє і дисперсію.

Викиди справді становлять помітну частку вибірки. За оцінками, кількість нетипово великих або малих значень для окремих показників обчислюється десятками тисяч. Наприклад, за допомогою Isolation Forest було виявлено, що близько 17 004 записів мають аномальні значення чистого прибутку, ~12 218 – аномальні EBITDA, та ~12 435 – аномально великі запаси. Це означає, що приблизно 15–22% даних по цих показниках можуть розглядатися як статистичні викиди. Їх наявність зумовлена тим, що в датасеті одночасно присутні фінансово малі компанії і надзвичайно великі корпорації, а також збиткові фірми поруч з дуже прибутковими. Для подальшого аналізу та візуалізацій ці екстремальні спостереження були враховані обережно: вони можуть свідчити про цікаві випадки (наприклад, значні збитки перед банкрутством або стрибкоподібні прибутки), але також можуть спотворювати тренди, якщо не розглядати їх окремо.

Більшість пар фінансових показників продемонстрували слабкий або помірний кореляційний зв’язок (рис. 4). Лише декілька пар мали коефіцієнти кореляції вище 0.7. Найбільш сильною виявилася очікувана кореляція між EBITDA та Net Income (p ~0.77). Ці дві величини обидві характеризують прибутковість, тож висока взаємозалежність логічна: компанії з високим EBITDA, як правило, мають і високий чистий прибуток, і навпаки. Окрім цієї пари, жодна інша пара фінансових змінних не відзначилася настільки тісним зв’язком.



(рис. 4)

Це важливий інсайт: фінансові метрики охоплюють різні аспекти діяльності компаній, і більшість з них не дублюють одна одну напряму. Наприклад, виручка слабо корелює із прибутком (компанія може мати великі обсяги продажів, але низьку рентабельність), активи не повністю визначають ринкову вартість (на останню впливають очікування інвесторів, структура капіталу тощо), а запаси чи дебіторська заборгованість майже не корелюють з EBITDA. Отже, кожна група показників – ліквідність, масштаб, прибутковість, заборгованість – робить свій внесок у загальну картину фінансового стану. Відсутність надмірної мультиколінеарності означає, що для пошуку інсайтів варто розглядати ці показники в комплексі, оскільки жоден окремий показник (окрім згаданих пар) не може бути однозначним маркером іншого.

Однією з гіпотез було припущення, що компанії з великими доходами мають більшу ймовірність залишитися діяльними. Для перевірки

цього всі компанії було поділено на дві рівні групи за рівнем Total Revenue (загальної виручки): вище медіанного значення та нижче. Результати виявилися дещо несподіваними – частка «живих» компаній майже однакова в обох групах. Зокрема, серед компаній з виручкою вище медіани частка тих, що не збанкрутували, становить ~93,44%, що практично дорівнює 93,37% для всього датасету. Різниця менш ніж 0,1% знаходиться в межах статистичної похибки. Таким чином, сам по собі високий обсяг виручки не гарантує довгострокового виживання. Компанія може генерувати значні продажі, але водночас зазнавати високих витрат чи збитків, що ставить під загрозу її існування. Даний інсайт узгоджується з тим, що необхідно враховувати не лише масштаби бізнесу, але і його ефективність. Висока виручка без належної рентабельності або управління витратами не убезпечує від банкрутства.

Розмір компанії та фінансова стійкість. Іншим важливим фактором, що розглядався, є розмір компанії з точки зору її ринкової капіталізації (Market Value). На відміну від виручки, великі компанії демонструють помітно кращу стабільність. Група компаній, чия ринкова вартість входить до топ-25% (четвертий квартиль), має частку живих компаній близько 96,6%. Це приблизно на 3 відсоткові пункти вище за середній рівень 93,4%. Така різниця є статистично значущою, враховуючи великий обсяг вибірки. Отже, компанії з великою ринковою вартістю значно рідше переживають банкрутство. Це свідчить про те, що розмір і ринкова успішність (цінність компанії в очах інвесторів) створюють своєрідний “запас міцності”. Великі корпорації зазвичай мають кращий доступ до капіталу, диверсифікований бізнес, впізнаваний бренд – всі ці чинники допомагають переживати економічні спади. Крім того, висока ринкова оцінка часто пов’язана з довірою інвесторів та кредиторів, що дозволяє таким компаніям легше рефінансувати борги чи залучати додаткові

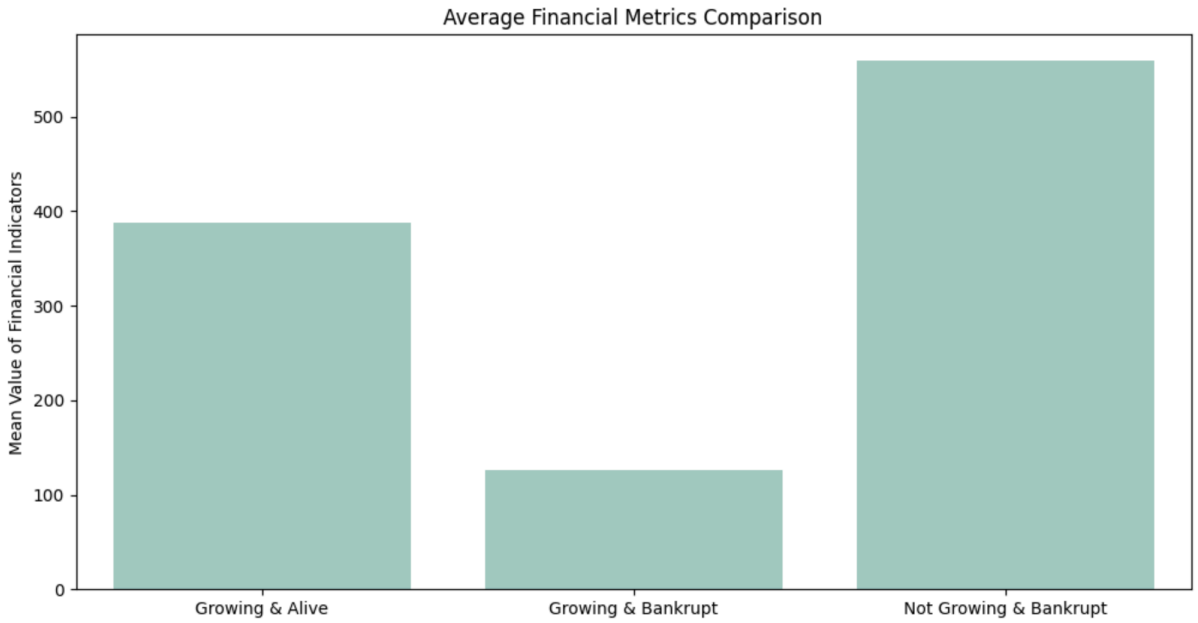
ресурси у разі тимчасових труднощів. Таким чином, величина компанії прямо корелює з її фінансовою стійкістю.

Оцінюючи прибутковість як можливий фактор успішності, було розраховано валову маржинальність – відношення валового прибутку до чистого обсягу продажів (Gross Profit / Net Sales). Цей показник відбиває ефективність основної діяльності: який відсоток від продажів залишається після собівартості продукції. Компанії з високою валовою маржею (вище медіани, тобто топ-50%) дещо перевершили менш ефективні за виживаністю: 94,3% проти 93,4% залишилися діяльними. Різниця ~1% хоч і невелика, але на масштабах тисяч компаній є показовою. Висока маржинальність дає підприємству певну перевагу в довгостроковому існуванні. Це логічно, адже більша маржа означає, що компанія отримує більше прибутку з кожної одиниці продажу, що створює фінансовий буфер для покриття постійних витрат, інвестицій чи непередбачених втрат. Компанії з високою рентабельністю можуть легше пережити періоди спаду продажів, оскільки їх запас прибутковості більший. Натомість низькомаржинальні бізнеси працюють «на межі» і при найменшому шоку (зростанні цін сировини, падінні попиту) можуть скотитися в збитки.

Цікавим результатом є ідентифікація групи так званих стійко зростаючих компаній. Близько 21,5% усіх компаній у вибірці продемонстрували здатність нарощувати свої фінансові показники протягом тривалого часу. До цієї категорії віднесено компанії, які щонайменше три роки поспіль показували приріст за ключовими показниками (такими як активи, виручка, прибуток тощо) без падінь. Вражає те, що п’ята частина компаній спромоглася на такий стабільний розвиток, адже підтримувати зростання постійно складно навіть у сприятливих ринкових умовах. Серед цих компаній виявлено декілька з особливо високим індексом зростання: наприклад, дві компанії мали

сумарний показник приростів, рівний 313 (це сумарна кількість річних покращень по різних метриках), що свідчить про десятиліття успішного розширення. Графічний аналіз підтверджує, що такі підприємства щороку збільшували обсяги продажів, активів, прибутків тощо практично без спадів.

Як правило, стабільно зростаючі компанії залишаються діяльними – серед них дуже низька частка банкрутств. Це можна пояснити тим, що сам критерій відбору (постійне зростання) відсіює більшість фінансово слабких або стагнуючих фірм, які більш схильні до невдач. Компанія, що щорічно нарощує показники, найімовірніше володіє конкурентними перевагами: наприклад, ефективним менеджментом, успішною стратегією, інноваційним продуктом чи міцною позицією на ринку. Такі переваги не тільки забезпечують зростання, але й захищають від банкрутства. Таким чином, виявлення понад 20% компаній з тривалим зростанням демонструє, що послідовний розвиток є досяжним і веде до підвищення шансів на довгострокове виживання. З іншого боку, близько 78,5% компаній не показали стабільного росту – серед них і знаходиться більшість випадків банкрутств. Це ще раз підкреслює, що відсутність зростання або коливальний розвиток можуть бути маркерами майбутніх проблем (рис. 5)



(рис. 5)

Практичний висновок полягає в тому, що аналіз коефіцієнтів фінзвітності дозволяє досить рано виявляти групу ризику, хоча й не забезпечує 100% точності для кожної фірми. Для підвищення точності прогнозу доцільно додатково врахувати інші фактори (наприклад, динаміку показників, макроекономічне оточення, галузеву специфіку), але базові фінансові індикатори залишаються фундаментом оцінки ризиків.

Однак, бачимо, що у цьому датасеті не вдалося знайти “гучних” інсайтів. У випадку з Titanic у нас був невеликий табличний набір з чітко вираженими категоріальними ознаками (наприклад: стать, клас, вік), які мали сильні кореляції із виживанням. Це дозволяло відразу побачити значні інсайти — наприклад, що жінки чи пасажири першого класу виживали значно частіше.

У цьому ж корпоративному датасеті ситуація інша:

1. Висока складність та багатовимірність

   Тут одразу 18+ фінансових показників, які відображають різні аспекти діяльності (ліквідність, прибутковість, заборгованість, масштаб бізнесу тощо). Жоден з них не «домінує» над іншими так яскраво, як у Titanic «стать» чи «клас».

2. Великий розкид і сильна асиметрія

   Чистий прибуток, EBITDA, виручка, активи мають розподіли з величезними хвостами й численними викидами. Через це статистичні зведення (середнє, медіана) мало показують без додаткових трансформацій (логарифмування, нормалізації, усічення хвостів).

3. Нечіткий зв’язок із цільовою змінною

   Як показав аналіз, сама по собі висока виручка майже не впливала на шанси компанії «вижити», на відміну від чистоти кореляцій у Titanic. Слабкі кореляції роблять прості групування малоефективними.

4. Збалансованість класів та контекст

   У Titanic ми мали майже рівну скорельованість classes (38% вижили), а тут лише ≈6.6% компаній банкрутували. Такий сильний класовий дисбаланс ускладнює виявлення закономірностей саме по банкрутству без додаткових методів (зважування, оверсемплінгу, моделювання ризику).

Таким чином, відсутність яскравих інсайтів пояснюється природою фінансових даних: вони складніші, багатовимірні, з викидами та слабкими простими зв’язками. Щоб отримати такі інсайти, варто застосувати глибший аналіз: сегментацію, трансформацію ознак, нелінійні моделі й більш вибагливі візуалізації у контексті подібних датасетів. Це допоможе знайти справді цікаві закономірності, які чітко пояснять, що саме впливає на стійкість компанії. Тим не менш, у моїй роботі було прийнято рішення рухатися у бік іншого датасету із більшою кількістю категоріальних ознак, а відповідно і більш ефективного та якісного пошуку інсайтів.

РОЗДІЛ 5

ПОШУК НОВОГО ДАТАСЕТУ ТА ЗАСТОСУВАННЯ МЕТОДУ ПОШУКУ ІНСАЙТІВ ДО НЬОГО

5.1 Опис та порівняння нового датасету

При пошуку нового датасету я зупинилася на наборі даних «Glassdoor Company Insights: Scraped Data Collection» на платформі Kaggle — це повна компіляція інформації, зібраної з Glassdoor, провідної платформи для оглядів співробітників і рейтингів компаній, а також вакансій у ніші IT. Він охоплює приблизно 10 тисяч компаній (записів) з США і має структуру з 9 основних полів, що включають як числові показники, так і категоріальні змінні. Файл CSV "glassdoor_usa_company.csv" складається з таких стовпців:

1. Company Name: назва компанії.

2. Company Rating: рейтинг або загальна оцінка задоволеності, присвоєна працівниками або користувачами.

3. Company Reviews: кількість відгуків або відгуків, отриманих для компанії на Glassdoor.

4. Company Salaries: інформація, пов’язана із заробітною платою, яку пропонує компанія, як-от середні зарплати, діапазони зарплат або дані, пов’язані із зарплатою.

5. Company Jobs: подробиці про вакансії або посади, доступні в компанії.

6. Location: Географічне розташування штаб-квартири або головного офісу компанії.

7. Number of Employees: приблизна кількість працівників, які працюють у компанії.

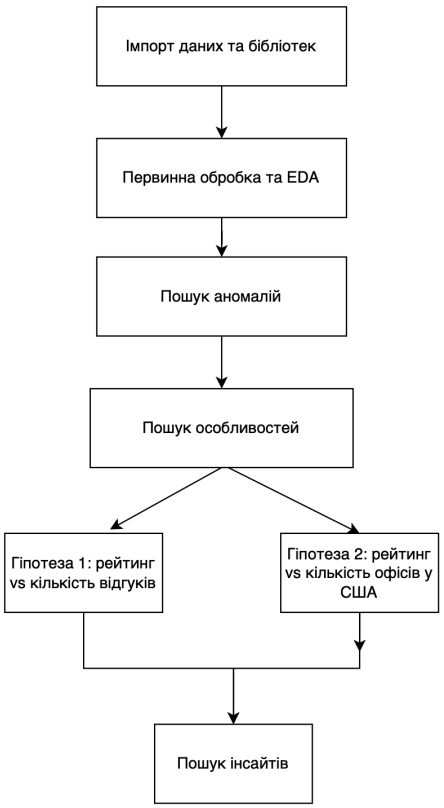
8. Industry Type: галузь або сектор, до якого належить компанія (наприклад, технології, охорона здоров'я, фінанси).

9. Company Description: короткий опис або огляд компанії, що висвітлює її основні види діяльності, продукти або послуги.

Цей набір даних надає інформацію про різні компанії США, включаючи їхні рейтинги, відгуки, зарплати, можливості працевлаштування, місцезнаходження, кількість працівників, тип галузі та описи. На відміну від попередньої спроби, де використовувався більш обмежений і менш структурований набір даних, обраний датасет пропонує значно ширший спектр змінних і спостережень, необхідних для глибшого пошуку інсайтів. Структура містить декілька категоріальних змінних (галузь, розмір компанії, локація), що забезпечує можливість групувати компанії за змістовними ознаками. Також присутні кількісні показники (рейтинг, кількість відгуків, вакансій, тощо), які дозволяють виконувати числовий аналіз і пошук статистичних залежностей. Інтерпретація результатів з цього датасету є зручнішою. Змінні такі як галузь чи розмір компанії мають реальний зміст для бізнес-аналізу, і знайдені на їх основі інсайти легко пояснити не спеціалістам. Наприклад, зрозуміло, що різні галузі можуть мати різний рівень задоволеності співробітників, або що компанії різного масштабу стикаються з різними викликами в управлінні персоналом – ці гіпотези прямо випливають з категоріальних даних.

5.2 Пошук інсайтів

Для виявлення інсайтів у бізнес-даних компаній було застосовано поєднання статистичних методів та алгоритмів машинного навчання. Зокрема, здійснено кореляційний аналіз числових показників, статистичне виявлення викидів (аномальних значень) за допомогою міжквартильного розмаху, а також використано алгоритм Isolation Forest для пошуку аномалій у категоріальних даних. Додатково розраховано спеціальні співвідношення показників (напр. коефіцієнт найму) для висвітлення нетривіальних характеристик компаній. Такий комплексний підхід дозволив автоматично ідентифікувати ключові тенденції та аномальні випадки в датасеті. Із кодом можна ознайомитися у **ДОДАТКУ 4.** Також пропоную подивитися на Flow-діаграму для простої візуалізації роботи програми (рис. 6)



(рис. 6)

Кореляційний аналіз. На першому етапі розраховано матрицю кореляцій між числовими змінними набору даних (включно з кількістю відгуків співробітників, кількістю зазначених зарплат, числом відкритих вакансій, розміром компанії тощо). Виявлено декілька сильних позитивних кореляцій. Найбільш високий коефіцієнт кореляції (близько 0.84) спостерігається між кількістю відгуків про компанію та кількістю зазначених зарплат її співробітників. Це свідчить про те, що ці дві величини масштабуються разом і фактично відображають розмір та популярність компанії на ринку праці. Іншими словами, великі компанії, що мають багато працівників, як правило, отримують і багато відгуків, і мають значну кількість інформації про зарплати на спеціалізованих платформах. Також зафіксовано очікувано високий зв'язок між числовим рейтингом компанії та індикатором «високого рейтингу» (двійковою ознакою, що вказує чи перевищує рейтинг поріг 4.0) – кореляція ~0.81. Це тривіальна залежність, оскільки індикатор прямо похідний від значення рейтингу, але вона підтверджує узгодженість визначення високо оцінених компаній (біля 27.8% усіх компаній у вибірці мали рейтинг вище 4.0). Натомість інших пар змінних із суттєвою кореляцією (понад 0.3 за модулем) не виявлено, що вказує на відсутність сильних лінійних залежностей між, наприклад, розміром компанії та її рейтингом чи іншими показниками у досліджених даних.

Пошук числових аномалій. Для виявлення незвичних або екстремальних значень у числових полях було застосовано метод на основі меж коробкового графіка (1.5 IQR). Квартильний аналіз визначив порогові значення, за межами яких спостереження вважаються статистичними викидами. Згідно з цим критерієм, до топ-аномалій потрапили насамперед дуже великі компанії за масштабами представленості на платформі. Наприклад, компанія *Amazon* виділяється найбільшою кількістю відгуків

співробітників (понад 168 тисяч) та зазначених зарплат (понад 201 тисячу) – ці показники значно перевищують типові значення і були ідентифіковані як аномально високі. Схожим чином, *Deloitte* має понад 167 тисяч зафіксованих зарплат, а *Accenture* – понад 145 тисяч відгуків, що значно виходить за межі меж квартильного розмаху і позначено як викид. Такі величини свідчать про винятковий розмір і глобальну присутність цих організацій: вони залучають величезну кількість співробітників, які залишають відгуки та діляться зарплатною інформацією, що і вирізняє їх на фоні інших. Окремо алгоритм виявив явний артефакт даних: для однієї організації (благодійного фонду *Susan G. Komen*) помилково зазначено число офісів 801889, що було класифіковано як екстремально аномальне значення. Ця знахідка підкреслює важливість автоматизованого контролю якості даних – вона вказує на можливу помилку парсингу або введення, яку слід скоригувати перед подальшим аналізом.

Аналіз аномалій у категоріальних ознаках. Важливим завданням було виявити, чи існують галузі або інші категорії компаній, у яких співвідношення високих та низьких рейтингів відрізняється відносно середнього по вибірці. Для цього використано алгоритм Isolation Forest, що н спеціалізується на пошуку аномальних спостережень. Алгоритм було застосовано до відношення кількості високо оцінених компаній до кількості низько оцінених у межах кожної категорії. Іншими словами, для кожної категоріальної ознаки (такої як галузь компанії чи діапазон чисельності персоналу) розраховано кількісний показник – *коефіцієнт успішності* (відношення компаній з рейтингом >4.0 до компаній з рейтингом ≤4.0). Далі Isolation Forest автоматично позначив ті категорії, для яких цей коефіцієнт є аномально високим або низьким у порівнянні з іншими.

Результати показали наявність галузевих аномалій у контексті задоволеності працівників. Зокрема, виділено кілька індустрій з незвично високою часткою високих рейтингів: до них належать *«Коледжі та університети»* та *«Розробка програмного забезпечення»*. У категорії «Коледжі та університети» понад половина компаній (≈53%) мають рейтинг вище 4.0, що вдвічі перевищує середній показник по всіх галузях. Це свідчить про особливо високий рівень задоволеності співробітників у освітньому секторі, можливо через стабільні умови праці або соціальну значущість роботи. Аналогічно, галузь IT показала близько 47% компаній з високим рейтингом, що також нетипово багато. Для контрасту, алгоритм виявив галузі з аномально низькою часткою високих оцінок. Найбільш проблемними виявилися *«Ресторани та кафе»* та *«Універмаги, одяг і взуття»*. У сегменті громадського харчування лише приблизно 8–9% компаній досягли рейтингу понад 4.0, тобто частка задоволених працівників є вкрай малою. Це суттєво нижче від середнього ~28% і вказує на систематично нижчий рівень задоволення працею в цій сфері – ймовірно, через специфіку умов роботи, рівень оплати чи високий стрес у сфері обслуговування. Схожа ситуація і в роздрібній торгівлі одягом: частка «щасливих» співробітників ледве досягає 7%, що робить цю категорію найнижчою за показником рейтингової успішності. Ці галузеві аномалії узгоджуються і з аналізом середніх значень рейтингу по індустріях: так, середня оцінка компаній у сфері харчування та роздрібної торгівлі одягом знаходиться серед найнижчих (близько 3.6–3.7 балів з 5), тоді як освітні установи мають одні з найвищих середніх рейтингів (~4.05). Виявлення таких відхилень є цінним інсайтом прикладного характеру: воно сигналізує, що умови праці та корпоративна культура суттєво відрізняються між секторами, і що, наприклад, робота в академічному чи IT-секторі зазвичай отримує кращі відгуки від персоналу, ніж робота в харчовому чи роздрібному секторах.

Співвідношення вакансій до чисельності персоналу. Для додаткового аналізу поведінкових характеристик компаній введено показник *інтенсивності найму* – відношення кількості відкритих вакансій в компанії до загальної кількості її співробітників. Цей коефіцієнт покликаний автоматично виявити компанії, що надзвичайно активно наймають персонал відносно свого розміру (можливо, швидко зростаючі або з високою плинністю кадрів). Було розраховано значення коефіцієнта найму для всіх компаній і відібрано топ-10 максимальних. Аналіз показав, що найбільші значення належать переважно відомим консалтинговим та технологічним фірмам. Зокрема, компанія *Deloitte* продемонструвала вкрай високий коефіцієнт: згідно з даними, вона мала понад 167 тисяч відкритих вакансій при оцінній чисельності персоналу ~3000 осіб (за категорією «1001–5000 працівників»). Такий розрив (більше 50 вакансій на одного співробітника) є винятковим і може вказувати на агресивну стратегію розширення або на значну присутність компанії на ринку праці (можливо, з урахуванням глобальних філій). Серед технологічних гігантів виділяється *Amazon*: для нього коефіцієнт найму оцінено ~13 вакансій на кожні 15 тисяч працівників, що теж є високим показником, хоча й не таким екстремальним. Загалом, більші корпорації (особливо з категорії 10 тис.+ працівників) демонструють суттєву інтенсивність найму. Подібні виявлені інсайти щодо найму дозволяють ідентифікувати компанії, які знаходяться в стадії стрімкого росту або переживають значні організаційні зміни. Це має прикладне значення для розуміння динаміки ринку праці: наприклад, настільки високий коефіцієнт найму може сигналізувати про високий попит на фахівців у відповідній галузі або про значну кадрову ротацію.

Висновки щодо поведінки компаній. Застосовані обчислення і методи аналітики надали низку змістовних висновків про характеристики компаній. По-перше, масштаб компанії сильно впливає на її

представленість у даних: найбільші гравці (Amazon, IBM, Deloitte тощо) вирізняються екстремально високими показниками активності (відгуків, вакансій, тощо), що робить їх статистичними викидами у вибірці. По-друге, простежено секторні відмінності у задоволеності співробітників: сфери освіти та високих технологій демонструють значно кращі середні рейтинги компаній і вищу частку позитивних відгуків, тоді як роздрібна торгівля та громадське харчування суттєво відстають. Це означає, що галузеві особливості – такі як умови праці, рівень оплати, стабільність зайнятості – впливають на оцінки роботодавців їхніми працівниками. По-третє, автоматичний аналіз коефіцієнтів найму висвітлив компанії з винятковою інтенсивністю найму, що може вказувати на швидке розширення бізнесу або високий рівень вакансій (потенційно пов’язаний з ростом або текучістю кадрів). Усі ці інсайти були отримані без необхідності ручного перегляду тисяч записів – завдяки поєднанню методів прикладної математики та машинного навчання вдалося виявити неочевидні закономірності в даних про компанії. Таким чином, продемонстровано ефективність автоматизованого підходу до аналізу бізнес-даних: він не лише підтвердив очікувані тренди (приміром, кореляцію метрик масштабу компанії), але й надав нові знання, корисні для подальшого дослідження управління персоналом та оцінки корпоративного середовища.

5.3 Математична частина роботи

Також важливо звернути увагу на математичну частину роботи. Було додано нову ознаку - бінарна мітка “Має високий рейтинг” (Is High Rated).

```
def is_popular(x):
    if x <= 4:
        return False
    else:
        return True

df['Is HIgh Rated'] = df['Company rating'].map(is_popular)
```

Тут створюється ознака, яка вказує, чи перевищує рейтинг компанії порогове значення 4.0. Формально:

$$\text{IsHighRated}(r) = \begin{cases} \text{True}, & r > 4.0 \\ \text{False}, & \text{else} \end{cases}$$

Рейтинг було категоризовано:

```
def mapper(x):
    if x < 3.6:
        return 'low'
    elif x <= 4.5:
        return 'medium'
    elif x > 4.5:
        return 'high'
```

Формально:

$$\text{RatingDistribution}(r) = \begin{cases} \text{'low'}, & r < 3.6 \\ \text{'medium'}, & 3.6 \leq r \leq 4.5 \\ \text{'high'}, & r > 4.5 \end{cases}$$

Це зручно для зведених таблиць (pivot tables) і графіків.

Пошук аномалій здійснюється за допомогою коефіцієнту високих/низьких рейтингів та IsolationForest. У коді використовується функція find_abnormal_categorical, яка застосовує метод IsolationForest для виявлення аномальних категорій. Спочатку всередині кожної категорії (наприклад, галузі) обчислюється співвідношення:

$$\text{ratio} = \frac{\#(\text{IsHighRated} = \text{True})}{\#(\text{IsHighRated} = \text{False})}$$

Потім це значення ratio подається на вхід моделі IsolationForest:

```
clf = IsolationForest(max_samples=100)
clf.fit(categories['ratio'].values.reshape(-1, 1))
```

Модель IsolationForest оцінює ступінь «ізольованості» кожної точки (у цьому випадку — категорії) від загальної маси й класифікує деякі категорії як аномальні, тобто такі, що сильно відхиляються від типового шаблону поведінки.

Також була додана нова метрика “Коефіцієнт найму” (Hiring Ratio)

```
df['Hiring Ratio'] = df['Company Jobs'] / df['Number of Employees']
```

Вона показує, наскільки інтенсивно компанія відкриває нові вакансії відносно кількості поточних працівників.

Топ-10 компаній із датасету з найвищим коефіцієнтом найму (Hiring Ratio) можуть свідчити про стартапи або фірми, що перебувають на піку розширення.

5.4 Цінність методу, новизна і висновки щодо проєкту

Одним словом, основна новизна та ключова цінність застосованого підходу полягає у наступному:

- Комбінація IsolationForest із ratio high/low - це не класична задача Outlier Detection. Модель працює не з сирими даними, а з співвідношенням оцінок у категоріях, що рідко використовується в прикладній аналітиці.

- Автоматичне створення інсайтів у структурованому вигляді - функція detect_insights збирає текстові висновки на кожному кроці.
- Hiring Ratio як метрика зростання - Цей індикатор побудовано вручну, і він корисний для виявлення стартапів або перегрітих компаній. Може знайти корисне застосування у бізнесі.

До того ж автоматично збираються кореляції між усіма числовими ознаками (corr > 0.3);

- через IQR шукаються аномалії;
- розраховується Hiring Ratio — індикатор зростання/перегріву компаній.

Ці інсайти одразу потрапляють до insights[] як список текстових висновків придатних до вставки в звіт. Це вже напівавтоматизована система аналітики, що скорочує час на ручну перевірку. Хоча дані у датасеті - про компанії, але застосовані методи є незалежними від предметної області, а також засновані на класичних статистичних підходах, тожї х можна застосувати до аналізу: клієнтів, продуктів, транзакцій, регіональних відмінностей тощо (будь-яких структурованих даних), що говорить про масштабованість підходу на різні ніші та бізнеси. Усі показники мають просту інтерпретацію — наприклад: IsHighRated=True говорить про те, що компанія надійна, Hiring Ratio > 1 → говорить про швидке зростання компанії.

РОЗДІЛ 6

ПОДАЛЬШИЙ РОЗВИТОК ІДЕЇ

6.1 Автоматизація звітності та інтеграція з BI

Інструменти візуалізації бізнес-аналітики (BI) дедалі частіше інтегрують елементи штучного інтелекту для автоматизації підготовки звітів та виявлення аналітичних висновків. Наприклад, сервіси на зразок Rollstack представляють собою автоматизовані системи звітності, що об’єднують можливості Tableau, Power BI та інших BI-платформ із програмами для створення презентацій (як-от PowerPoint чи Google Slides). Це дає змогу формувати повноцінні аналітичні звіти без необхідності ручної обробки. Подібні рішення значно пришвидшують оновлення аналітичних матеріалів і зменшують навантаження на аналітиків, звільняючи їм час для стратегічних завдань.

Також, у Power BI вже реалізовано функції автоматичного виявлення аномалій у часових даних із поясненням результатів та інтерактивною візуалізацією. Завдяки цьому користувачі можуть буквально за кілька кліків виявляти відхилення у даних і розуміти можливі причини без необхідності ручного аналізу великої кількості показників.

Таким чином, впровадження розробленого підходу у BI-середовища (наприклад, у вигляді модуля для виявлення інсайтів у Tableau чи Power BI) дозволить автоматично доповнювати дашборди поясненнями виявлених аномалій і стислими аналітичними висновками.

6.2 Підтримка нових типів даних і завдань

Розроблений підхід можна легко розширити на більш складні сценарії. Його можна, наприклад, адаптувати для багатокласових задач

(наприклад, класифікації різних видів аномалій або подій) з використанням підходів one-vs-rest (стратегія розбиває багатокласову класифікацію на одну задачу бінарної класифікації на клас) чи softmax-виходів (багатозмінна логістична функція). Також можливе розширення на роботу з текстовими та часовими даними: наприклад, у разі аналітики клієнтських відгуків застосувати NLP-методи для векторизації тексту і виявлення незвичних тем чи сентименту, а для часових рядів – сезонні моделі та статистичне прогнозування відхилень. Прикладом подібного розширення в реальних BI-сценаріях є використання алгоритмів виявлення аномалій на часових даних – вони дозволяють автоматично виділяти «сплески» чи «провали» у показниках (наприклад, у продажах чи трафіку) і попереджати про них користувача. Таким чином, підхід може працювати не лише з табличними числовими даними, але й з даними інших форматів, гнучко поєднуючись з існуючими модулями аналізу.

6.3 Вдосконалення та перспективи функціоналу

Додаткові можливості покращення методу:

- Адаптивне визначення порогів. Замість фіксованого порогу для відокремлення аномалій можна застосувати динамічний поріг, який змінюється залежно від контексту та часу. Наприклад, у системах моніторингу Dynatrace «адаптивні пороги» самостійно підлаштовують поріг під дані за останні дні, зменшуючи кількість хибних спрацьовувань при зростанні навантаження. Аналогічно, наша система може автоматично коригувати чутливість у залежності від сезонності або нових трендів.

- Інтерпретація за допомогою SHAP (SHapley Additive exPlanations, теоретико-ігровий підхід до пояснення виходу будь-якого режиму

машинного навчання, допомагає зрозуміти, які ознаки найбільш важливі для прогнозу). Використання методів пояснень (наприклад, SHAP) дозволить робити висновки не лише «що» вважається аномалією, а й «чому». Як показано у дослідженнях XAI, для непрозорих моделей (наприклад, нейромереж чи ансамблів) можна отримати ваги ознак через SHAP, щоб пояснити рішення моделі. У нашому випадку це означало б: якщо в графіку продажів виділено аномальний сплеск, система може повідомити, що «пік відбувся через одноразову акцію чи помилку в даних».

- Інтеграція з A/B-тестуванням. Аномалії можна зв'язати із проведеними експериментами. Відомо, що деякі збої помітні лише в певних когортах користувачів. Об'єднання нашого підходу з системою A/B-тестування дозволить автоматично аналізувати, чи експеримент вплинув на якусь метрику (і чи це відхилення є аномалією). Наприклад, за прикладом інструменту SMAD (пропонується як інструмент для автоматичного виявлення аномалій у глобальних показниках (метрики, де користувачі не є частиною жодного A/B-тесту, тобто вони перебувають у статус-кво), а також виявляють статистично значні відмінності між статус-кво та експериментальними когортами в активних тестах A/B. Після виявлення, електронний лист має бути надіслано відповідним сторонам. На високому рівні SMAD приймає дані щовечора від Amazon Redshift, виконує глобальний і експериментальний метричний аналіз і надсилає електронні листи, якщо це можливо) можна налаштувати оповіщення не тільки про глобальні аномалії, але й про статистично значущі відмінності між контрольними й експериментальними групами.

6.4 Цінність для бізнесу

Реалізація таких можливостей забезпечить значну користь для компаній будь-якого масштабу. По-перше, автоматизація звітності та інсайтів зменшує навантаження на аналітиків, прискорює прийняття рішень і знижує людський фактор. Великі компанії отримають єдину платформу для моніторингу сотень чи тисяч показників у реальному часі, де автоматично визначатимуться відхилення і формуватимуться звіти. Малі та середні підприємства можуть ефективніше використовувати обмежені ресурси: навіть без команди аналітиків вони отримають повідомлення про критичні аномалії (наприклад, різкий спад продажів чи незвичну активність у веб-трафіку) з поясненням причин, що також оптимізує фінансування. Згідно з дослідженнями, інструменти, що «добавляють аномалії» в BI, допомагають виділяти важливі патерни без необхідності глибокого аналізу. Всі ці аспекти роблять систему цінною і для невеликих стартапів (яким потрібна швидка діагностика проблем з мінімальним бюджетом), і для великих корпорацій (які мають складні екосистеми даних і потребують масштабованих рішень).

ВИСНОВОК

У роботі детально продемонстровано та обгрунтовано, а також застосовано на прикладах підхід автоматичного пошуку інсайтів. Виявлено, що він працює з більш високою точністю на датасетах із вищим вмістом категоріальних даних. Визначено напрямок подальшого розвитку ідеї, а також підкреслено її переваги.

Поєднання виявлення аномалій із автоматизованою генерацією звітів відповідає сучасним трендам доповненої аналітики. Згідно з прогнозами Gartner, до 2025 року більшість аналітичних звітів будуть автоматично згенерованими «історіями» з даних, а пояснювані методи аналізу аномалій стають необхідними для прийняття рішень у критичних сферах.

На прикладах (Titanic, фінансовий набір, Glassdoor-дані) підхід показав свою практичність. Автоматизоване виявлення аномалій і генерація звітів мають прямий бізнес-вплив: вони дозволяють оперативно виявляти фінансові порушення, шахрайство чи системні збої. Це підвищує прозорість операцій і допомагає компаніям своєчасно реагувати на ризики, покращуючи прийняття рішень та довіру до даних.

У майбутньому можливе розширення набору ознак, використання глибших моделей та аналіз даних у реальному часі. Також актуальним є впровадження нових технік генерації текстових описів і інтеграція з інтерфейсами на основі штучного інтелекту. Експерти прогнозують, що автоматизоване створення аналітичних історій займе домінуючу роль у BI-системах найближчим часом.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Automating Anomaly Detection, Dongyu Zhe, 01.2018
URL: https://dongyuzheng.com/files/Automating_Anomaly_Detection.pdf

2. Business Intelligence in the Age of AI: Tools & Strategy for 2025, 05.12.2024
URL: https://www.rollstack.com/articles/business-intelligence-in-the-age-of-ai-tools-strategy#:~:text=Rollstack%E2%80%99s%20AI%20Report%20generation%20connects,to%20focus%20on%20strategy%20rather

3. Anomaly detection, Microsoft Learn, 01.31.2025
URL: https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-anomaly-detection

4. Auto-adaptive thresholds for anomaly detection, 05.04.2024
URL: https://docs.dynatrace.com/docs/discover-dynatrace/platform/davis-ai/anomaly-detection/concepts/auto-adaptive-threshold

5. Kaggle, USA Company Insights: Glassdoor Scraped Data 2023, 2023
URL: https://www.kaggle.com/datasets/joyshil0599/glassdoor-company-insightsscraped-data-collection

6. An introduction to explainable AI with Shapley values
URL: https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html

7. A Survey on Explainable Anomaly Detection, ZHONG LI, YUXUAN ZHU, and MATTHIJS VAN LEEUWEN, Leiden Institute of Advanced Computer Science (LIACS), Leiden University, The Netherlands, 11.07.2023

URL: https://arxiv.org/pdf/2210.06959#:~:text=by%20making%20the%20anomaly%20detection,175%5D%2C%20and

8. Power BI Documentation
URL: https://learn.microsoft.com/ru-ru/power-bi/

9. Kaggle, Titanic Dataset
URL: https://www.kaggle.com/competitions/titanic/data

ДОДАТОК 1 (Стаття “Як шукати історії (insight) у даних” Платонова Віталія Валерійовича)

URL:

https://docs.google.com/document/d/1CozWHlFkvTfTezKt-159RLK3_QkvAGc05OJFAfPIY8E/edit?tab=t.0

ДОДАТОК 2

Посилання на Collab Notebook:
https://colab.research.google.com/drive/1bQXVLZgb_gx5nLkLWYDYiVBP5m_5_W5C?usp=sharing

ДОДАТОК 3

Посилання на Collab Notebook:
https://colab.research.google.com/drive/1pWR0gQyz2sahq_tRGHFE1Ls3wG-jCzZb?usp=sharing

ДОДАТОК 4

Посилання на Collab Notebook:
https://colab.research.google.com/drive/1-3W2ORTu_RkoHP1kv07mcb0tl2HO7slb?usp=sharing