Хмельницький національний університет
Факультет програмування та комп'ютерних і телекомунікаційних систем
Кафедра комп'ютерних наук та інформаційних технологій

ДИПЛОМНА РОБОТА
МАГІСТРА

на тему *Виявлення аномалії в бухгалтерському звіті на базі штучного інтелекту*

Галузь знань *12 – Інформаційні технології*
Шифр і назва галузі знань

Спеціальність *122 – Комп'ютерні науки*
Шифр і назва спеціальності

конав: студент *2* курсу, група *КНм-19-1* Підпис *Б.Г. Гордійчук* Ініціали, прізще

Керівник: *к.т.н., доцент кафедри КНІТ* Підпис *Е.А. Манзюк* Ініціали, прізвище

Нормоконтроль: *к.т.н., доцент кафедри КНІТ* Підпис *Р.О. Багрій* Ініціали, прізвище

До захисту допускаю:
Зав. кафедри КНІТ, д.т.н., професор Підпис *О.В. Бармак* Ініціали, прізвище
7 12 2020 р.

Хмельницький 2020

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет *програмування та комп'ютерних і телекомунікаційних систем*

Кафедра *комп'ютерних наук та інформаційних технологій*

Освітній ступінь *магістр*

Галузь знань *12 – Інформаційні технології*

Спеціальність *122 – Комп'ютерні науки*

ЗАТВЕРДЖУЮ

Завідувач кафедри *комп'ютерних наук та інформаційних технологій*

_____

(підпис)

д.т.н., професор О.В. Бармак

«*7*» *9* *2020* року

ЗАВДАННЯ

НА ДИПЛОМНУ РОБОТУ МАГІСТРА

1. Тема дипломної роботи магістра: *«Виявлення аномалії в бухгалтерському звіті на базі штучного інтелекту»*

2. Завдання дано студенту *Гордійчук Богдан Геннадійоч*

(прізвище, ім'я, по батькові)

3. Керівник роботи *к.т.н., доцент Манзюк Едуард Андрійович*

(прізвище, ім'я, по батькові)

4. Затверджені наказом університету від «*9*» *9* *2020* р. № *92*

5. Зміст пояснювальної записки (перелік задач) та хідні дані:

Мета роботи – є розробка методу аналізу та визначення нетипових даних з використанням штучного інтелекту на базі фінансових даних. Предметом дослідження є набір даних фінансового характеру а також ознаки та класифікатори даних машинного навчання. Об'єктом дослідження є порівняння методів визначення нетипових даних на аномалії та викиди.

Реферат

Дипломна робота магістра присвячена виявленню аномалії в бухгалтерському звіті на базі штучного інтелекту.

Актуальність теми. В магістерській роботі розроблений і реалізований метод виявлення аномалій та викидів в даних фінансової сфери. Розроблювальна система пропонує системний підхід з визначення та аналізу даних та встановлення аномалій в даних. Такі системи можуть допомогти у встановленні нетипових даних та визначення даних, які можуть носити важливий характер. Цей напрямок в дослідженнях важливий з точки зору пошуку нетипових даних особливо з точки зору машинного навчання.

Метою дослідження є розробка методу аналізу та визначення нетипових даних з використанням штучного інтелекту на базі фінансових даних.

Для досягнення зазначеної мети поставлені наступні **задачі**:

- показати, що використання засобів машинного навчання дає змогу визначити нетипові дані;
- провести дослідження визначення ознак виявлення аномалій в даних;
- провести порівняння застосовності методів машинного навчання базуючись на результатах досліджень.

При цьому передбачається розв'язок таких **підзадач**, як

- попередня обробка даних та їх очистка;
- вибір моделей, ділення ознак і застосування методів машинного навчання для отримування інформації про відповідні параметри;
- тестування методів на основі правил і з використанням машинного навчання та вибір оптимального методу для кожного з параметрів;
- програмна реалізація системи визначення аномалій.

Об'єктом дослідження є порівняння методів визначення нетипових даних на аномалії та викиди.

Предметом дослідження є набір даних фінансового характеру а також ознаки та класифікатори даних машинного навчання.

Існуючі системи розпізнавання та визначення нетипових даних розвинуті в недостатній мірі. Тому проведені дослідження з визначенням найбільш ефективних методів та підходів по визначенню та інтерпретації даних. Дані можуть бути як аномаліями, викидами так і новими даними. Визначення причин та факторів впливу є важливим фактором.

Запропонована методика визначення аномалій. Отримані аналітичні співвідношення для імовірнісних характеристик відповідних випадкових полів.

Синтезовані алгоритми виявлення аномалій при наявності випадкових перешкод в умовах невідомих параметрів.

Проведений аналіз ефективності виявлення запропонованих і відомих алгоритмів, що дозволяє виробити рекомендації з вибору необхідних значень параметрів для забезпечення заданих характеристик виявлення.

Достовірність результатів забезпечується проведенням всебічного оцінювання та порівняння ефективності різних методів.

У процесі виробництва на всіх стадіях і циклах ділової активності працівники підприємства (навмисно або не навмисно) можуть прибігати до викривлень і фальсифікації. У результаті таких дій підприємства можуть мати більші матеріальні та моральні втрати, тому однієї з основних задач аудита є явища помилок і фактів шахрайства, а також здійснення необхідних заходів щодо попередження можливих втрат підприємства.

Практична значимість дослідження полягає в тому, що описані методи та отримані результати застосовуються для виявлення аномалій в даних. Сфера практичного застосування може бути розповсюджена на дослідження даних будь-якого типу.

У результаті застосування методики явища аномалій вдалось сполучити ефективний метод пошуку викидів у даних *LOF* з алгоритмом глибоких автоенкодерних нейронних мереж. Це забезпечило останньому високу стійкість до викривлень у даних одночасно зі значним збільшенням продуктивності системи при побудові моделі. Стійкість до викривлень зазначена як зниження точності класифікації при різних рівнях шуму в даних, яке при використанні

запропонованої методики виявилося суттєво менше зниження точності при застосуванні інших алгоритмів. Збільшення продуктивності становить $p/2$ раз для кожного атрибута об'єкта даних, де p — число значень атрибута, перевіряемого на аномальність.

Для задоволення вимог сучасного бізнесу потрібне автоматичне виявлення аномалій, яке може надавати точну інформацію в режимі реального часу незалежно від того, скільки метрик потрібно відстежувати. Дійсно автоматизовані системи виявлення аномалій повинні включати виявлення, ранжування та групування даних, усуваючи потребу у великих групах аналітиків.

Апробація дипломної роботи.

Основні положення і результати роботи опубліковані в збірнику наукових праць – Гордійчук Б. Г. Виявлення аномалій в даних / Б. Г. Гордійчук, Е. А. Манзюк, Т. К. Скрипник // Збірник наукових праць за матеріалами Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук - 2020» Хмельницький, 2020, – С.72-74.

Структура та обсяг роботи. Дипломна робота магістра складається з завдання, реферату, змісту, вступу, 4 розділів, висновків, переліку посилань із 38 найменувань та додатків. Загальний обсяг дипломної роботи магістра становить 75 сторінок, в роботі наведено 30 рисунків та 1 таблицю.

Ключові слова: викиди, класифікація, машинне навчання, нетипові дані.

Зміст

Вступ

Революція в інформаційних технологіях, що привела до повсюдного використання комп'ютерів, а також дігіталізації культурного та інтелектуального надбання людства, привела до усвідомлення важливості проблем, пов'язаних з аналізом накопиченої інформації та подоланням парадокса наявності інформаційного голоду (нестачі знань) на тлі достатку первинної інформації (даних). Ці проблеми актуалізувалися лише в останнє десятиліття XX століття, коли інформаційні технології стали перетворюватися в необхідний атрибут життя сучасної людину та здобувати самодостатній характер. Активізація розробки методів, що дозволяють ефективно досліджувати оцифрованну інформацію, орієнтуватися в безмежному океані накопичених даних, витягаючи з них потрібні в цей момент знання, згодом привела до того, що такі методи виділилися в специфічний підрозділ штучного інтелекту, що одержав назва «інтелектуальний аналіз даних».

У цей час інтелектуальний аналіз даних є однією зі сфер, що найбільше активно розбудовуються, міждисциплінарних досліджень, що швидко рекрутує послідовників не тільки серед компьютерщиків, розроблювачів систем штучного інтелекту, але та серед економістів, соціологів, психологів, медиків і ін. XXI-е століття, часто називають століттям комп'ютерних технологій, неможливо представити без інтелектуального аналізу даних. Сучасна людина навіть не усвідомлює, що просто здійснюючи пошук інформації в Інтернет, він задіє цілий комплекс методів інтелектуального аналізу даних. Дослідники, аналізуючи емпіричні дані, застосовують сучасні пакети статистичної обробки, які містять у собі інтелектуальні методи. Інтелектуальний аналіз даних поступово починає визнаватися соціологічним співтовариством: 1) у якості корисного інструмента обробки емпіричної інформації; 2) як необхідного атрибута нового витка в розвитку. Показовим є зростаючий інтерес розроблювачів інтелектуальних технологій до соціологічної специфіки, створення ними спеціалізованих систем

інтелектуального аналізу даних. При цьому слід зазначити, що поява нових технологій, що допомагають у роботі з інформацією, у науковому співтоваристві соціологів сприймається беззастережно позитивно. Однак назва «інтелектуальний аналіз даних» викликає недовіру, як щось загадкове та неприродне. Інтеллектуальный аналіз даних – одна з кон'юнктурних назв прикладної статистики, що замість нього цілком природно використовувати термін статистична технологія. Це актуалізує дослідження термінологічних інновацій, що мають метою вербалізацію суті процесу добування знань із наявних даних у формі адекватного найменування.

У ряді робіт пропонуються локальні алгоритми виявлення на багатомірних даних. Розглядається застосування адаптивних алгоритмів компенсації корреліpованих перешкод і алгоритмів виявлення аномалій. Незважаючи на велику кількість публікацій із проблем синтезу алгоритмів виявлення аномалій, у цей час мало досліджена їхня ефективність. Існуючі розв'язки не дають задовільного результату в задачах аналізу ефективності та порівняння відповідних алгоритмів виявлення, особливо алгоритмів виявлення аномалій на даних.

Таким чином, представляється досить актуальної задача розробки та дослідження алгоритмів виявлення аномалій в умовах апріорної невизначеності відносних даних.

Актуальність теми. В магістерській роботі розроблений і реалізований метод виявлення аномалій та викидів в даних фінансової сфери. Розроблювальна система пропонує системний підхід з визначення та аналізу даних та встановлення аномалій в даних. Такі системи можуть допомогти у встановленні нетипових даних та визначення даних, які можуть носити важливий характер. Цей напрямок в дослідженнях важливий з точки зору пошуку нетипових даних особливо з точки зору машинного навчання.

Метою дослідження є розробка методу аналізу та визначення нетипових даних з використанням штучного інтелекту на базі фінансових даних.

Для досягнення зазначеної мети поставлені наступні **задачі**:

– показати, що використання засобів машинного навчання дає змогу визначити натипові дані;

– провести дослідження визначення ознак виявлення аномалій в даних;

– провести порівняння застосовності методів машинного навчання базуючись на результатах досліджень.

При цьому передбачається розв'язок таких **підзадач**, як

– попередня обробка даних та їх очистка;

– вибір моделей, ділення ознак і застосування методів машинного навчання для отримування інформації про відповідні параметри;

– тестування методів на основі правил і з використанням машинного навчання та вибір оптимального методу для кожного з параметрів;

– програмна реалізація системи визначення аномалій.

Об'єктом дослідження є порівняння методів визначення нетипових даних на аномалії та викиди.

Предметом дослідження є набір даних фінансового характеру а також ознаки та класифікатори даних машинного навчання.

Існуючі системи розпізнавання та визначення нетипових даних розвинуті в недостатній мірі. Тому проведені дослідження з визначенням найбільш ефективніх методів та підходів по визначенню та інтерпритації даних. Дані можуть бути як аномаліями, викидами так і новими даними. Визначення причин та факторів впливу є важливим фактором.

Запропонована методика визначення аномалій. Отримані аналітичні співвідношення для імовірнісних характеристик відповідних випадкових полів.

Синтезовані алгоритми виявлення аномалій при наявності випадкових перешкод в умовах невідомих параметрів.

Проведений аналіз ефективності виявлення запропонованих і відомих алгоритмів, що дозволяє виробити рекомендації з вибору необхідних значень параметрів для забезпечення заданих характеристик виявлення.

Достовірність результатів забезпечується проведенням всебічного оцінювання та порівняння ефективності різних методів. У процесі виробництва на всіх стадіях і циклах ділової активності працівники підприємства (навмисно або не навмисно) можуть прибігати до викривлень і фальсифікації. У результаті таких дій підприємства можуть мати більші матеріальні та моральні втрати, тому однієї з основних задач аудита є явища помилок і фактів шахрайства, а також здійснення необхідних заходів щодо попередження можливих втрат підприємства.

Практична значимість дослідження полягає в тому, що описані методи та отримані результати застосовуються для виявлення аномалій в даних. Сфера практичного застосування може бути розповсюджена на дослідження даних будь-якого типу.

У результаті застосування методики явища аномалій вдалось сполучити ефективний метод пошуку викидів у даних *LOF* з алгоритмом глибоких автоенкодерних нейронних мереж. Це забезпечило останньому високу стійкість до викривлень у даних одночасно зі значним збільшенням продуктивності системи при побудові моделі. Стійкість до викривлень зазначена як зниження точності класифікації при різних рівнях шуму в даних, яке при використанні запропонованої методики виявилося суттєво менше зниження точності при застосуванні інших алгоритмів. Збільшення продуктивності становить p /2 раз для кожного атрибута об'єкта даних, де p — число значень атрибута, перевіряемого на аномальність.

Для задоволення вимог сучасного бізнесу потрібне автоматичне виявлення аномалій, яке може надавати точну інформацію в режимі реального часу незалежно від того, скільки метрик потрібно відстежувати. Дійсно автоматизовані системи виявлення аномалій повинні включати виявлення, ранжування та групування даних, усуваючи потребу у великих групах аналітиків.

Розділ 1

Аналіз аномалій та викидів та практичного використання систем ідентифікації

1.1 Опис предметної області

Суть явища обману та помилок. У процесі виробництва на всіх стадіях і циклах ділової активності працівники підприємства (навмисно або не навмисно) можуть прибігати до викривлень і фальсифікації. У результаті таких дій підприємства можуть мати більші матеріальні та моральні втрати, тому однієї з основних задач аудита є явища помилок і фактів шахрайства, а також здійснення необхідних заходів щодо попередження можливих втрат підприємства. Для ефективної роботи з явища та усуненню помилок і зловживань в 1982 р. був розроблений і затверджений міжнародний норматив аудита "Обман і помилка" Комітету з аудиторської практики і інструкція "Відповідальність аудиторів у зв'язку зі зловживання, іншими аномаліями та помилками". З 1 січня 1999 р. в Україні придбав силу національний норматив аудита № 7 " Про помилки та шахрайстві". Метою цього нормативу є "зобов'язання тлумачення та використання термінів "шахрайство" і "помилка" з позицій підготовки аудиторського висновку, значення ризику аудита та впливу шахрайства та помилок на вірогідність фінансової звітності клієнта".

Шахрайство передбачає навмисний неправильний показ фінансової інформації одним або декількома посадовим особа зі структури керівництва підприємства, що служать (може здійснюватися шляхом маніпуляцій, фальсифікацій і змін записів на рахунках бухгалтерського обліку, в облікових реєстрах або документах; навмисного неправильного віднесення до активів різних статей; знищення або пропуски записів операцій або документів; відбиття операцій без розкриття їх змісту; підготовки та використання в обліку фальсифікованих первинних документів).

Помилка стосується ненавмисних порушень у відбитті фінансової інформації, які виникають у результаті арифметичних, граматичних або інших помилок у записах облікових даних; ненавмисного пропуску або неправильного уявлення про окремі факти; неправильного відбиття рахункових обладнань; різних відхилень від правил контролю над діяльністю службових осіб.

Обман є наслідком навмисного порушення у відображенні фінансової інформації одним або декількома посадовцями серед керівництва підприємства. Він може здійснюватися шляхом фальсифікації, підробки або зміни записів на рахунках бухгалтерського фінансового обліку або неправильного віднесення до активів або пасивів окремих статей; знищення або пропуски окремих господарських операцій або документів і т.п.

Аудит повинен проводитися так, щоб забезпечити гарантію розкриття істотних неточностей у бухгалтерському обліку та звітності.

Розрізняють два основні види неточностей:

- помилки,
- відхилення від норми.

Помилка — це ненавмисне викривлення даних бухгалтерського обліку та звітності, а відхилення від норми, навпаки, — ненавмисне викривлення показників обліку та звітності.

Якщо аудиторові вдалося знайти зловживання, то він повинен насамперед з'ясувати, як вони позначилися на фінансовій звітності, а, отже, взяти під сумнів правильність цієї інформації. Однак якщо аудитор дійшов висновку, що наявність незаконних дій стать під сумнів інформацію фінансової звітності, оскільки вона неправильно відображає реальний стан справ, він зобов'язаний відповідно змінити свій висновок. Аудитор може також переглянути та власне відношення до адміністрації. Якщо вона знала про явлених аудитором фактах незаконних дій, але не повідомила про них раніше, виникає серйозний сумнів щодо довіри до такої адміністрації.

Якщо клієнт відмовляється признати змінений аудиторський висновок або не ухвалює ніякі заходи, щоб виправити недоліки в обліку та звітності, то аудитор може відмовитися від подальшого виконання замовлення. Однак цей розв'язок має серйозні наслідки та прийняття його вимагає консультацій.

Особливу увага слід звернути на відповідальність адміністрації (не аудитор, а адміністрація фірми-клієнта відповідає за правильність облікової політики, за створення та використання відповідної системи внутрішньогосподарського контролю та за об'єктивне складання фінансової звітності). Звітність публічних фірм може включати звіт про обов'язки їх адміністрації та про відносини з аудиторськими фірма.

1.2 Аудиторські докази

Для складання обґрунтованих звітів аудитор повинен мати повну, достовірну та надійну інформацію — аудиторські докази. Стандарти щодо кількості, якість і процедур одержання аудиторських доказів установлені в національному нормативі аудита "Аудиторські докази" № 14. Джерела інформації можуть бути облікова система, первинні документи, матеріальні акти, робітники підприємства, контрагенти.

Усі первинні документи та облікові записи повинні вестися відповідно до вимог постанови Кабінету Міністрів України " Положення про організацію бухгалтерського обліку та звітності в Україні" від 3 квітня 1993 р. № 250" Положення про документальне забезпечення записів у бухгалтерському обліку", затвердженого Наказом Міністерства фінансів України від 14 травня 1995 р. № 88, Закону України " Про бухгалтерський облік і фінансової звітності в Україні" від 16 липня 1999 р. № 996-XIV (зі зміна та доповнення). Цей Закон набув чинності з 1 січня 2000 року. Відповідно до Програми реформування системи бухгалтерського обліку із застосуванням міжнародних стандартів, затвердженої постановою Кабінету Міністрів України від 28 жовтня 1998 р. № 1706, набув

чинності цілий ряд положень (стандартів) бухгалтерського обліку, інструкцій і методичних вказівок.

Кількість і якість інформації, якої може володіти аудитор, багато в чому визначає якість аудиторського висновку, його адекватність фінансовому та майновому стану підприємства.

Аудиторські докази — це інформація, отримана аудитором, що служить базою для підготовки аудиторського висновку (звіту). Аудитор повинен одержати таку кількість аудиторських доказів, яка дозволила б йому підготувати аудиторський висновок.

Аудиторські докази включають первинні документи та облікові записи, які лежать в основі складання звітності, а також інформацію, яка підтверджується з інших джерел.

Прямі докази — це дані, які підтверджуються первинними документа та обліковими записами.

Прямі докази підрозділяють на:

матеріальні — це документи та натуральні об'єкти,

нематеріальні — це інформація контрагентів підприємства, на якому проводиться аудит.

Аудиторські докази одержують шляхом об'єднання тестів систем контролю та процедур перевірки на істотність.

На достатність аудиторських доказів впливає ряд факторів:

- оцінка аудитором характеру та величини ризику, властивого як на рівні фінансової звітності, так і на рівні залишків на рахунках або по класу операцій;
- характер систем обліку та внутрішнього контролю та оцінка ризиків контролю;
- досвід, придбаний у процесі попередніх аудиторських перевірок;
- результати аудиторських процедур з можливими виявленя фактами помилок або шахрайства;

– джерело та надійність наявної інформації.

Надійність аудиторських доказів залежить від джерела їх одержання — внутрішнього або зовнішнього, а також від їхнього характеру.

По характеру розрізняють:

– візуальні;

– документальні;

– усні докази.

Аудиторські докази, отримані від зовнішніх джерел (наприклад, підтвердження, отримане від третьої особи), більш надійні, чим отримані із внутрішніх джерел.

Аудиторські докази, отримані від внутрішніх джерел інформації підприємства, яке перевіряється, надійні у випадку наявності ефективних систем обліку та внутрішнього контролю.

Аудиторські докази, отримані за допомогою проведених тестів, надійніше тих, які отримані від працівників підприємства.

Аудиторські докази у формі документів або письмових підтверджень надійніше, чим усні підтвердження.

По доказовості документи, які перевіряються, діляться на:

– первинні документи — письмові свідчення, які фіксують і підтверджують господарські операції, включаючи розпорядження та дозволу адміністрації (власника) на їхнє проведення;

– облікові записи (облікові регістри) — носії спеціального формату (паперові, машинні) у вигляді відомостей, ордерів, машинограм і т.п., призначені для хронологічного, систематичного або комбінованого нагромадження, обґрунтування та узагальнення інформації первинних документів, призначених до обліку;

– головна книга — регістр синтетичного обліку, у якому відбиваються залишки на початок і кінець звітного періоду, обороти по дебету та кредиту відповідних рахунків;

– фінансова звітність — баланс, звіт про фінансові результати, звіт про рух грошових коштів, звіт про власний капітал;

– інвентаризаційні матеріали (описи, порівнювальні відомості, рахунки природнього збитку);

– розрахунки, декларації, кошториси, калькуляції, договори, контракти, установчі документи, статути, накази, розпорядження, бізнес-плани;

– оперативна, статистична, податкова звітність;

– матеріали перевірок і ревізій, проведених органа податкової служби, державної контрольно-ревізійної служби, статистики, банків і ін.;

– матеріали внутрішньогосподарського контролю (внутрішнього аудита);

– дані документального та фактичного контролю, експертних перевірок, лабораторних аналізів, контрольних вимірів, проведених за участю аудиторів;

– письмові та усні дані, пояснювальні та доповідні записки матеріально відповідальних і посадових осіб, суб'єктів підприємницької діяльності, замовників і документи, які робітники підприємства ведуть за власною ініціативою, неофіційні документи, допоміжні документи — документи, у яких викладена думка осіб, що працюють на підприємстві.

Неофіційні та допоміжні документи доказового значення не мають, однак можуть бути використані при виборі напрямку проведення дослідження та перевірки.

Як машинне навчання може дозволити явища аномалії. Як людина, наш мозок завжди налаштований на явища чогось з "нормального" або "звичайного матеріалу". Коротше кажучи, деяка аномалія, яка не вписується в звичайну картину. При зростанні даних, інструменти науки про дані також шукають аномалії, які не підписуються на нормальний потік даних. Наприклад, "надзвичайно висока" кількість спроб входу може вказувати на потенційну

кібератаку, або великий похід у транзакції за кредитними картками за короткий період потенційно може бути шахрайством з кредитними картками.

1.3 Машинне навчання у фінансах

Перш ніж розглянемо деякі програми у фінансах машинного навчання, спочатку опишемо, що таке машинне навчання.

При цьому явища аномалій перед обличчям безперервного потоку неструктурованих даних з різних джерел має свої виклики. Прикладом завдання припустимо, що більшість операцій з кредитними картками є законною та належною, шукати серйозні відхилення в кількох транзакціях, які ходять за межі "нормального" діапазону досить складно.

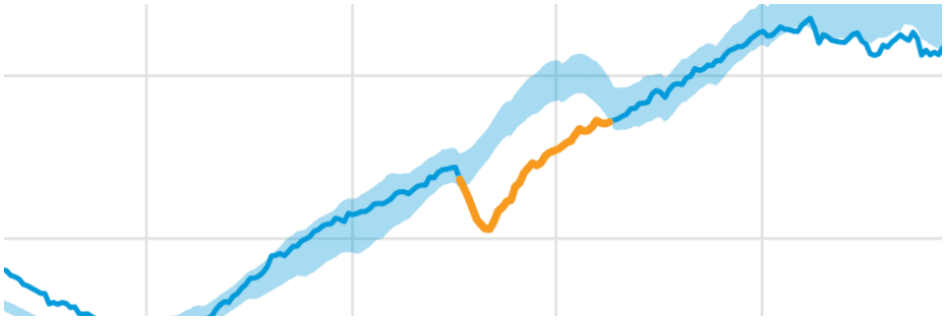


Рисунок 1.1 – Виявлення аномалій [6]

Завдяки зростанню різних технологій глибокого навчання, явища аномалії за допомогою машинного навчання (або МН) є практичним рішенням сьогодні. Алгоритми машинного навчання можуть бути розгорнуті для значення нормальних моделей даних і використання моделей ML для пошуку відхилень або аномалій.

Отже, як аналітик даних, можна впровадити визначення явища аномалії за допомогою машинного навчання що мають методи і переваги явища аномалії з використанням технологій глибокого навчання.

Крім того, є явища викиду, явище аномалії - це просто режим явища та ідентифікації аномальних даних у будь-якій події або спостереженні на основі даних, що суттєво відрізняється від решти даних. Аномальні дані можуть мати вирішальне значення при виявленні рідкісної моделі даних або потенційної проблеми у погляді фінансових махінацій, медичних умов або навіть несправного обладнання.

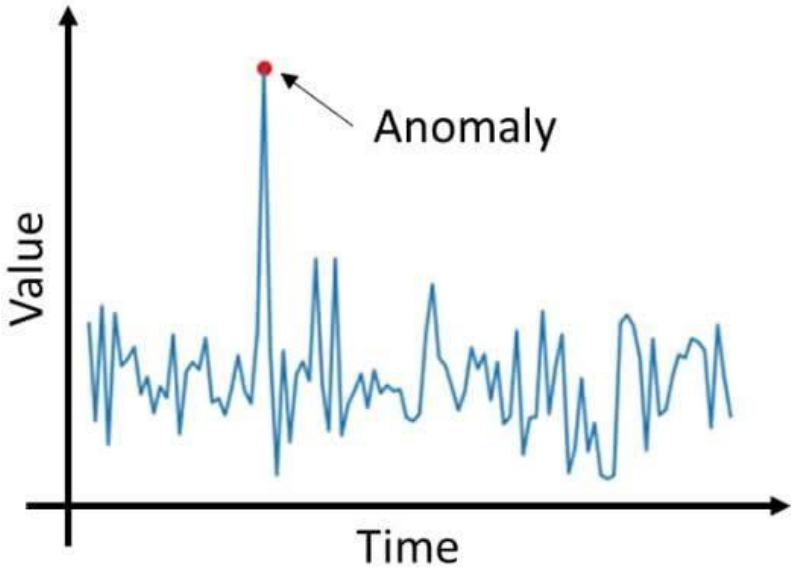


Рисунок 1.2 – Аномалії в даних [2]

Якщо йдеться про явища аномалії в даних. Розглянемо це за допомогою випадку явища аномалії за допомогою 2 змінних (X & Y). Розглянемо такі візуалізовані дані.

Розглянемо шаблони даних змінних 2 на основі графіків праворуч. Виходячи з цих точок даних, неможливо найти будь-яку аномалію (або викид). Однак, коли 2 змінні один проти одного (як показано на лівому рисунку), можемо чітко знайти аномалію.

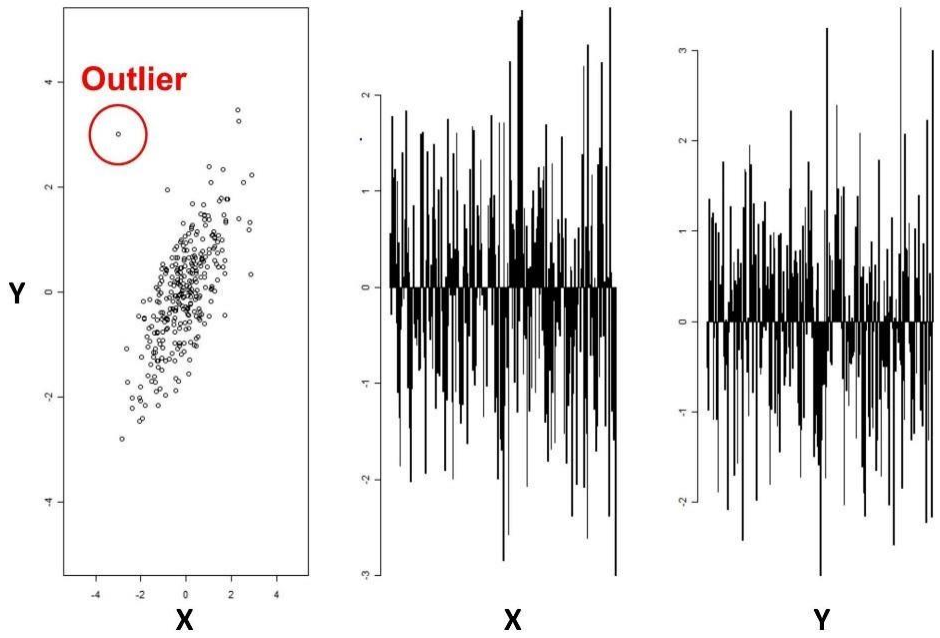


Рисунок 1.3 – Аномальне спостереження [2]

Чи підводить це нас до питання про те, чому потрібне машинне навчання при виявленні аномалії. Явища аномалій може бути дуже складним завданням, якщо плануєте не дві, а сотні таких змінних у реальних сценаріях.

1.4 Машинне навчання та аналіз викидів

Виявлення аномалії в високовимірних даних стає фундаментальною проблемою досліджень, яка має різні додатки в реальному світі. Тим не менш, багато існуючих методів виявлення аномалій не зберігають достатньої точності через так звані "великі дані", що характеризуються високим обсягом, і даними високої швидкості, що генеруються різними джерелами. Це явище наявності обох проблем разом можна віднесли до "прокляття великої вимірності", які впливають на існуючі методи з точки зору як продуктивності, так і точності. Щоб вирішити цю прогалину і зрозуміти основну проблему, необхідно визначити унікальні проблеми, які виникають через виявлення аномалії як з високою розмірною, так і з проблемами великих даних. Отже, це дослідження має на меті задокументувати стан виявлення аномалії у великих даних з

високими вимірами, представляючи унікальні виклики, використовуючи трикутну модель вершин: проблему (великий вимір), методи / алгоритми (виявлення аномалії) та інструменти (додатки для великих даних / фреймворки). Робота авторів, які потрапляють безпосередньо в будь-яку з вершин або тісно пов'язаних з ними, враховується для розгляду. Крім того, обговорюються обмеження традиційних підходів і поточних стратегій високовимірних даних разом з останніми методами і додатками на великих даних, необхідних для оптимізації виявлення аномалії.

Багато наборів даних, постійний потік з блогів, фінансових операцій, медичних записів і журналів спостереження, а також з бізнесу, телекомунікацій та біонаук. Цей розділ нещодавно став центром дослідження, який називається "великі дані", термін, який описує великий і розподілений характер наборів даних. Великі дані визначаються як великі обсяги, високу швидкість і набори даних високої різноманітності, які вимагають економічно ефективної нової аналітики даних для прийняття рішень і для створення корисних аналітичних даних. В останні роки широко налагоджені основні проблеми великих даних. Вони містяться в межах п'яти властивостей великих даних - значення, достовірність, різноманітність, швидкість і обсяг.

Викид ідентифікується як будь-який об'єкт даних або точка, яка значно відхиляється від решти точок даних. У сумі даних аутсорсинг зазвичай відкидається як виняток або просто шум. Тим не менш, те ж саме не може бути зроблено при виявленні аномалії, отже, акцент на аналізі викидів.

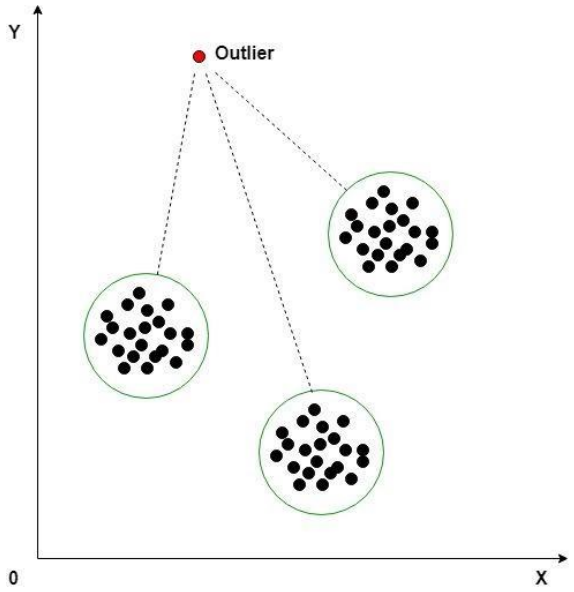


Рисунок 1.4 – Викид в даних [3]

Прикладом визначення явища аномалії за допомогою машинного навчання є метод кластеризації K-means. Цей метод використовується для явища викиду на основі їх відстані s від найближчого кластера.

K-means метод кластеризації передбачає формування декількох кластерів точок даних кожен з середнім значенням. Об'єкти у кластері мають найближче середнє значення. Будь-який об'єкт із граничним значенням, більшим за середнє значення найближчого кластера, ідентифікується як викид. Ось покроковий метод, який використовується в K-means кластеризацію:

– Обчислити середнє значення кожного кластера.

– Установіть початкове граничне значення.

– Під час тестування зазначте відстань кожної точки даних від середнього значення.

– Зазначте кластер, найближчий до точки даних перевірки.

– Якщо значення "відстань" більше, ніж значення "поріг", позначте його як більший елемент.

Далі розглянемо деякі інші методи визначення явища аномалії за допомогою машинного навчання.

На основі різних алгоритмів машинного навчання методи явища аномалій в першу чергу класифікуються на наступні два типи.

Контрольовані методи. Як пливає з назви, цей метод явища аномалії вимагає існування позначеного набору даних, який містить як звичайні, так і аномальні точки даних. Приклад контрольованих методів є явища аномалії за допомогою нейронних мереж, байєсівських мереж та методу K-найближчих сусідів (або k-NN).

Контрольовані методи забезпечують кращу якість визначення явища аномалії завдяки їх здатності кодувати будь-яку взаємозалежність між змінними і включення попередніх даних в будь-яку прогностичну модель.

Безконтрольні методи. Безконтрольні методи явища аномалії не залежать від будь-яких навчальних даних з ручним маркуванням. Ці методи засновані на статистичному припущенні, що більшість притоку даних є нормальними, і лише незначний відсоток буде аномальними даними. Ці методи також підрахували, що будь-які шкідливі дані будуть сильно відрізняється статистично від звичайних даних. Деякі з безнаглядових методів включають метод K-means, автокодери та аналіз на основі гіпотези.

У наступних розділах розглянемо деякі з переваг явища аномалії за допомогою машинного навчання.

Використовуючи можливості машинного навчання, явище аномалії має практичні застосування та переваги в різних сферах бізнес-операцій. Деякі переваги середовища явища аномалії включають:

Явища вторгнення. Будь-яка нечесна діяльність, яка може пошкодити інформаційну систему, може бути широко класифікована як вторгнення. явища аномалії може бути ефективним як у виявленні, так і у рішенні вторгнення будь-якого типу. Поширені вторгнення, пов'язано з даними, включаючи кібератаки, порушення даних або навіть дефекти даних.

Дані датчика мобільного зв'язку. Ще однією перевагою явища аномалії за допомогою машинного навчання є збір та аналіз даних мобільного датчика.

Зростаюче використання пристроїв IoT і зниження трат на збір даних за допомогою мобільних датчиків, безумовно, є рушійною силою цієї тенденції.

Наприклад, конкретне галузеве дослідження полягає в тому, що IBM Data Science Experience розробив інструмент для явища аномалії для захоплення даних датчиків з мобільних телефонів та підключених пристроїв IoT.

Крім того, явища аномалії можуть надати будь-які допоміжні дані, які можуть виначити першопричину проблеми.

Контроль статистичного процесу. Статистичний процес контролю (або SPC) є стандартом якості, який є поширеним у виробничому процесі. Дані, пов'язані з якістю продукту або процесу, завантажуються під час виконання виробництва та виводяться на графіку для моніторингу, якщо дані в межах настроєного елемента керування.

1.5 Постановка задачі

Метою дослідження є розробка методу аналізу та визначення нетипових даних з використанням штучного інтелекту на базі фінансових даних.

Для досягнення зазначеної мети поставлені наступні задачі:

– показати, що використання засобів машинного навчання дає змогу визначити натипові дані;

– провести дослідження визначення ознак виявлення аномалій в даних;

– провести порівняння застосовності методів машинного навчання базуючись на результатах досліджень.

При цьому передбачається розв'язок таких підзадач, як

– попередня обробка даних та їх очистка;

– вибір моделей, ділення ознак і застосування методів машинного навчання для отримування інформації про відповідні параметри;

– тестування методів на основі правил і з використанням машинного навчання та вибір оптимального методу для кожного з параметрів;

– програмна реалізація системи визначення аномалій.

Висновок до розділу 1

Майбутнє просування машинного навчання і технології глибокого навчання тільки додасть обсягу методів виявлення явища аномалії і їх значення для бізнес-даних. Зростаючий обсяг і складність даних можна використати на основні можливості аналізу цієї інформації для успіху бізнесу.

З моменту свого створення, та освоєння рішення машинного навчання в штучному інтелекті та машинного навчання для своїх даним дасть змогу виявити нестандартні дані. Якщо інвестували в інструменти машинного навчання, то зможемо допомогти побудувати бізнес-важелі з методів виявлення явища аномалії.

Явища аномалії застосовуються, щоб перевірити, чи будь-які дані виходять за межі контролю та зазначити першопричину. Коротше кажучи, явище аномалії може бути використане для знаходження будь-якого варіанту продукту або будь-якої проблеми, що виникає в процесі, яка повина бути негайно вирішена.

Розділ 2
Аномалії та викиди в даних
2.1 Типізація аномальних явищ

Явища аномалії відноситься до завдання пошуку спостережень, які не відповідають нормальній, очікуваній поведінці. Ці спостереження можуть бути названі як аномалії, викиди, новизна, винятки, сюрпризи в різних доменах додатків. Найпопулярнішими термінами, які зустрічаються найчастіше в літературі, є аномалії і викиди. Явища аномалії є актуальною проблемою в різних областях, таких як:

- явища вторгнення;
- явища шахрайства;
- явища зумисного пошкодження;
- медичне та громадське здоров'я;
- обробка зображень;
- явища аномалії в текстових даних;
- сенсорні мережі та інші домени.

На жаль, немає чіткого значення аномалії, тому вибрано наступне - аномалії є закономірності в даних, які не відповідають чітко зазначеній нормальній поведінці [1]. Можемо проілюструвати аномалії в простому двомірному просторі.

N_1 і N_2 регіони з нормальними даними, тому що більшість спостережень знаходяться в цих областях. Точки, які знаходяться далеко від нормальних районів, таких як точки O_1, O_2 і регіон O_3 є аномалія. Аномалії в даних можуть виникнути з різних причин. Зловмисна діяльність, шахрайство з кредитними картками, вторгнення, поломки системи тощо. Ці аномалії привабливі для аналітика даних. Тому явища аномалії є важливим процесом і розглядаються як перевага в різних системах прийняття рішень.

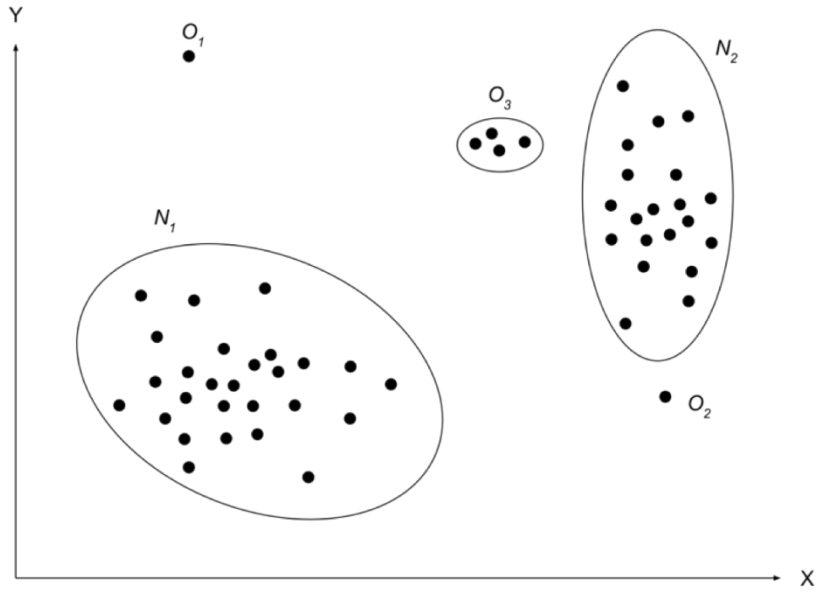


Рисунок 2.1 - Ілюстрація простих аномалій у двомірному просторі [3]

Типи аномалій. Аномалії можна класифікувати за трьома категорія:

Точку аномалії. Якщо один об'єкт можна спостерігати проти інших об'єктів як аномалія, це точка аномалії. Це найпростіша категорія аномалії і багато досліджень включають їх. Беручи до уваги приклад, представлений на рисунку 2.1 O_1 O_2 є точкою аномалії.

Контекстні аномалії. Якщо об'єкт є аномальним у значеному контексті. Тільки в цьому випадку це контекстна аномалія (також відома як умовна аномалія). На рисунку 2.2 можна побачити періодичний контекст. У цьому випадку точка O_1 аномалія, тому що вона відрізняється від періодичного контексту.

Колективні аномалії. Якщо деякі зв'язані об'єкти можна спостерігати проти інших об'єктів як аномалія. O_1 об'єкт не може бути аномальним в цьому випадку, тільки колекція об'єктів.

Значення відноситься до переваг, пов'язаних з аналізом даних; достовірність відноситься до точності даних; і різноманітність відноситься до багатьох типів даних, таких як структуровані, напівструктуровані або неструктуровані. Обсяг — це обсяг даних, що накопичується (тобто розмір

даних) — чим більша розмірність даних, тим більший обсяг. Розмірність – це кількість функцій, атрибутів або змінних у даних, доступних для аналізу. На відміну від цього, швидкість відноситься до "швидкості", з якою дані генеруються і можуть містити багато вимірів. Ці елементи поточного визначення великих даних вирішення основних проблем. Однак таке визначення ігнорує ще один важливий аспект: "розмірність", який відіграє вирішальну роль в реальному аналізі даних. Збільшення розмірів або функцій або атрибутів створює значні проблеми для виявлення аномалії у великих наборах даних.

У системах підтримки прийняття розв'язків важливе місце мають механізми прогнозного аналізу даних. Прогнозний аналіз даних є процесом формування суджень про майбутні факти на основі обробки та аналізу вихідного набору статистичних даних, що називають множиною навчання, або генеральною сукупністю. Результат навчання — аналітична модель, використовувана надалі при формуванні прогнозів. Серйозною перешкодою при побудові прогнозної моделі може бути наявність шумів у вихідних навчальних даних. Викликані шумом викривлення впливають на процес побудови прогнозної моделі, а також на якість її роботи, втілюються в точності розпізнавання об'єктів при прогнозуванні. У кінцевому рахунку викривлення у вихідних даних знижують ефективність роботи, впливаючи на розв'язки та керуючі оперативні впливи, формовані системою.

Задачу, яку ставимо перед собою, є дослідження та розробка методик явища аномалій у вихідних даних, на яких будуються прогнозні моделі. Існуючі підходи до розв'язку проблеми явища аномалій виявили методи які розбиті на кілька категорій по загальному характеру. Методи кожної категорії мають переваги та недоліки, повинні вибиратися залежно від специфіки предметної області окремо взятої задачі.

Процес побудови системи і, зокрема, прогнозній моделі аналізу даних починається з навчання моделі на вихідних даних, тому розглянемо методи, здатні працювати на етапі навчання моделі, а саме методи, засновані на широко

відомому підході *k* найближчих сусідів, коли об'єкти аналізують разом з іншими об'єкта, найближчими до них. Ключова проблема при виявленні аномалій — пошук відстаней між об'єкта даних, оскільки в системах прийняття розв'язків використовують не тільки числові дані, шкали вимірів яких часто заздалегідь відомі, але та категоріальні дані, виражені у вербальній формі, що утрудняє їхнє порівняння. Огляд істотно діючих критеріїв оцінки відстаней між значення категоріальних атрибутів даних проведений у роботі [4], також у цій роботі обраний оптимальний критерій оцінки відстані. Проблемою зазначеного критерію є його залежність від загального числа об'єктів даних. Це утрудняє розрахунки відстаней у динамічних системах, об'єкти даних у які можуть попадати в процесі роботи систем, а не тільки на початковому етапі формування моделі аналізу.

Ціль роботи — аналіз і опис методики обробки шуму в даних і заснованого на ній алгоритму розв'язків, що дозволяє здолати наступні проблеми, наявні в існуючих алгоритмах пошуку аномалій:

– проблема наявності різнорідних викривлень у даних;

– проблема вибору ефективної стратегії підщеннями якості даних. Розроблений алгоритм повинен знайти викривлення двох типів:

– аномальні значення атрибутів даних;

– відсутні значення.

Для обробки аномальних значень необхідно використовувати методи пошуку аномалій у даних, для обробки відсутніх значень — алгоритми заповнення пропусків у даних. У роботі показано, наскільки успішно можна застосовувати алгоритми пошуку викидів у прогнозних моделях.

Наукова новизна роботи полягає в розробці підходу до розв'язку задачі явища аномалій на етапі побудови моделі прийняття розв'язків за допомогою методики обробки шуму в даних. У відомих роботах не презентовано аналогічних або подібних підходів оптимізації побудови моделі.

Задача обробки вихідних даних з метою явища та корекції шуму має істотну актуальність, тому що кожний з типів шуму може впливати на процес побудови прогнозної моделі, особливо в областях, пов'язаних із забезпеченням безпеки людини.

Виявлення аномалії має на меті виявити аномальні закономірності, що відхиляться від решти даних, які називаються аномаліями або аутсайдерами. Висока розмірність створює труднощі для виявлення аномалії, оскільки, коли кількість атрибутів або функцій збільшується, обсяг даних, необхідних для точного узагальнення, також зростає, що призведе до розрідження даних, в якому точки даних більш розкидані та ізольовані. Ця розрідженість даних обумовлена непотрібними змінними або високим рівнем шуму декількох невідповідних атрибутів, які приховують справжні аномалії. Це питання широко визнається як "прокляття розмірності". Це перешкода для багатьох методів виявлення аномалій, що стосуються високої вимірності, які не зберігають ефективність звичайних підходів, таких як дистанційні, щільні та кластеризовані методи .

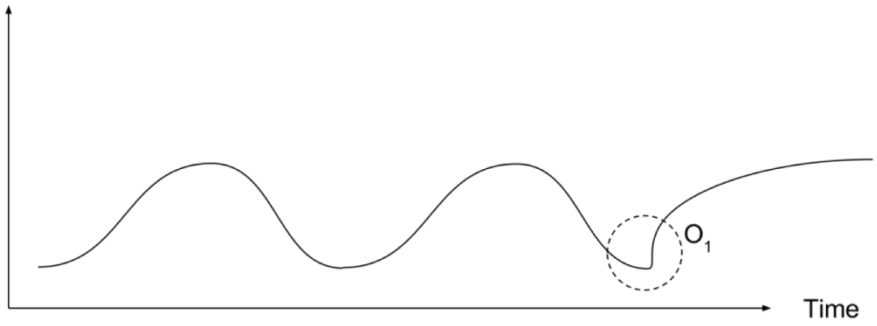


Рисунок 2.2 - Приклад контекстної аномалії [6]

Аномалії можуть бути пов'язані між собою. Точкові аномалії можуть стати контекстуальними, якщо застосуємо до неї контекст. Або точкові аномалії можуть стати колективними, якщо об'єднаємо кілька точкових аномалій разом.

На абстрактному рівні явища аномалій здаються простим завданням. Але це завдання може бути дуже складним. Ось деякі проблеми.

Визначити нормальні регіони дуже складно. У багатьох випадках межі між аномаліями і нормальними даними не є точними. У цьому випадку нормальні спостереження можна розглядати як аномалії і навпаки.

Якщо дія шкідлива, як шахрайство, це розглядається як аномалія. Дуже часто зловмисники намагаються адаптувати свої дії до нормальної поведінки. І знову ж таки, завдання знайти аномалії в цьому випадку не таке просте.

Що сьогодні вважається нормальним, не може бути нормальним у майбутньому. Більшість бізнес-систем змінюються в часі під впливом різних факторів.

Підходи до явища аномалії в одному полі часто не можна використовувати в іншому. Вони будуть неефективними в більшості випадків.

Навчання та перевірка даних, доступність для навчання моделі є великою проблемою.

Отже, не можна визначати аномалії в деяких випадках «на око». Навіть якщо застосовувати деякі статистичні методи, все одно це не легко.

2.2 Підхід до явищ аномалій

Підходи, які можна використовувати для пошуку аномалій, потрапляють до таких категорій [4]:

Контрольоване явища аномалії. Налаштування, де дані позначені в навчальних та тестових наборах даних; коли простий класифікатор може бути навчений, і застосовуватися. Цей випадок схожий на традиційне розпізнавання шаблонів за винятком класів, які в більшості випадків сильно незбалансовані. Не всі підходи класифікації підходять для цього завдання. Наприклад, деякі типи дерев рішень не можуть добре справлятись з незбалансованими даними. Використання векторних машин (SVM) або штучних нейронних мереж (ANN) повинна працювати краще. Однак, ця установка не є актуальним, тому що повинні знати всі аномалії і позначити ці дані правильно. Для багатьох випадків

аномалії не відомі заздалегідь або можуть виникнути як новинки на етапі тестування.

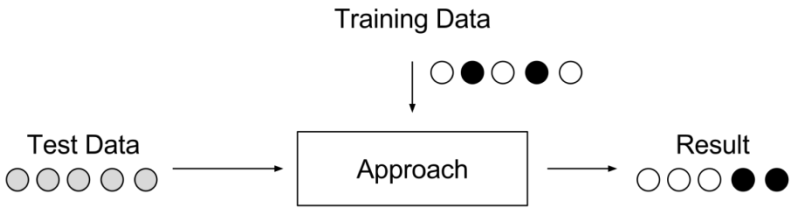


Рисунок 2.3 - Контрольоване явища аномалії [3]

Явища аномалії під наглядом. На початку, коли немає знань, збираємо їх з результатів навчання. Ця установка також використовує навчальні та тестові набори даних, де тільки навчальні дані складаються з нормальних даних без будь-яких аномалій. Ідея полягає в тому, що модель нормального класу вже дається і аномалії можуть бути виявлені шляхом відхилення від вченої моделі. Такий підхід також відомий як класифікація «однокласова». Відомі підходи є однотипні з SVM та автокодерами. Загалом, будь-який підхід до оцінки щільності може бути застосований для моделювання функції щільності ймовірності звичайних класів, таких як оцінка щільності ядра.

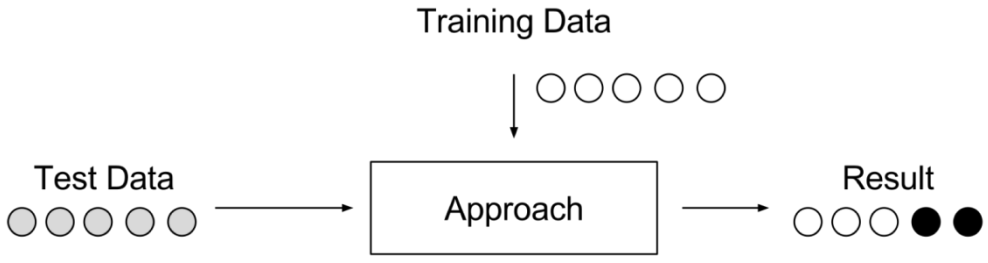


Рисунок 2.4 - Виявлення аномалій Supervised [3]

Явища аномалії без нагляду. Установка, коли не знаємо, що нормально в даних, а що ні. Це найбільш гнучка конфігурація, яка не вимагає будь-яких

етикеток. Також немає різниці між навчанням і тестом набором даних. Концепція полягає в тому, що без нагляду підходи до явища аномалії групують дані виключно на основі природних особливостей набору даних. Як правило, відстані або щільність використовуються, щоб дати оцінку, що є нормальним, а що є більшим на нормальне значення. Візуальне представлення можна знайти на рисунку 2.5.

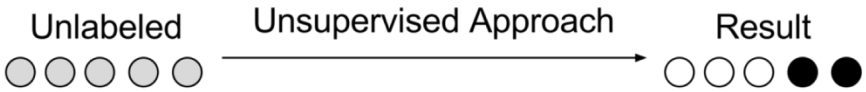


Рисунок 2.5 - Явища аномалії Non-Supervised [3]

Більшість підходів базується на маркуванні. Різниця між скорингом і маркуванням знаходиться в гнучкості. За допомогою методів підрахунку очок аналітик може брати значення, які більше підходять для проблемних областей. Після цього він може використовувати порогове значення для набору аномалій або просто брати верхні. Маркування є класифікацією. Ніякі підходи на сьогоднішній день не можуть бути використані у всіх доменах з однаковим успіхом без досліджень про домен і видалення функцій.

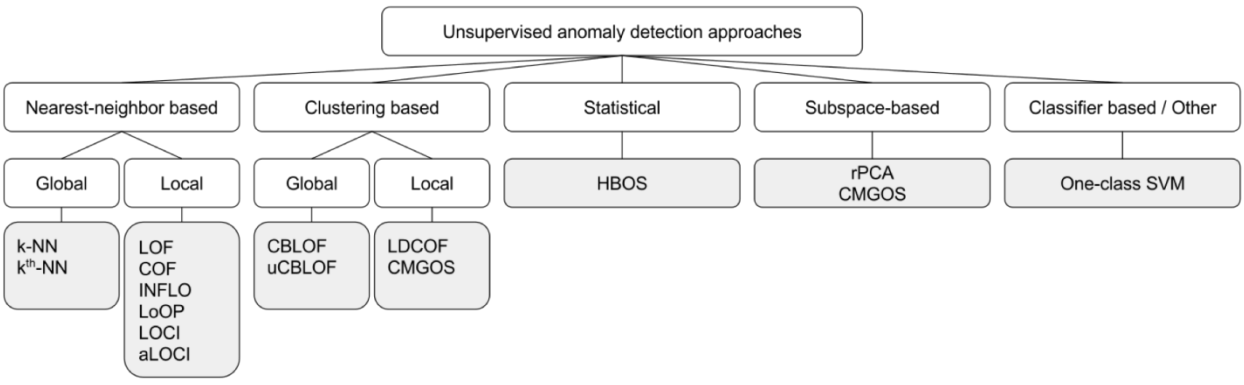


Рисунок 2.6 - Non-Supervised явища аномалії, схеми класифікації [5]

Реальне життя динамічне. Якщо поведінка спостережуваних об'єктів у зазначеному середовищі змінюється в часі, це середовище можна розглядати як динамічне середовище.

Динамічне середовище, таке як виробничий процес, мережа зв'язку, дорожній рух тощо, може містити величезну кількість інформації, що змінюється з часом, що є цінним ресурсом для розуміння загальної поведінки навколишнього середовища, явища регулярності та аномалій, що зараз відбуваються в навколишньому середовищі, контролю процесу еволюції та інтелектуального моделювання або управління навколишнім середовищем [3].

Наприклад: інтернет-магазин, який переживає зростання. Зміна обсягу трафіку та транзакцій. Те, що було нормально раніше, не буде зараз, і цю зміну можна розглядати тільки як аномалію. Це одна з проблем, виявлених раніше. Іншим прикладом є система, яка аналізує поведінку клієнтів. Клієнти можуть бути класифіковані спочатку. Але клієнти також можуть перейти чомусь з одного класу в інший в часі. Системи розвиваються з плином часу, коли програмне забезпечення оновлюється або змінюється поведінка. По-перше, для ефективного явища аномалій потрібні системи безперервного навчання. По-друге, щоб знайти аномалії на ранніх стадіях, не можна чекати, поки метрика буде явно за межами. Раннє явище потребує здатності виявляти тонкі зміни в шаблонах, які не є очевидними. Крім того, оскільки аномалії за своєю природою несподівані, ефективна система явища повинна бути в змозі визначити, чи є аномальні події новими, не покладаючись на заздалегідь запрограмовані пороги.

Явища аномалій в динамічному середовищі здається важким завданням. Найбільша проблема, з якою можна зіткнутися, полягає в тому, що в більшості компаній не буде даних для навчання. Більшість підходів до використання в такому випадку повинні бути без нагляду.

Можливі рішення. Рішення, які в даний час прогресивні і дають прийнятні реезультати включають нейронні мережі (NN), деякі з них в поєднанні зі статистичними підходами.

Популярною NN, яка використовується зараз для тайм-послідовності, є довгострокова пам'ять (LSTM) — вона отримала популярність у сфері Deep Learning через її можливість запам'ятати тривалі залежності.

Ще одним цікавим видом мережі, яка є новою, і не згадується часто, є ієрархічна тимчасова пам'ять (HTM) - це нейронна мережа, яка заснована на дослідженнях в нейронауці, яка також має здатність вчитися і запам'ятовувати тривалі залежності.

2.3 Ієрархічна тимчасова пам'ять

Ієрархічна тимчасова пам'ять (HTM) Non-Supervised відноситься до напівнаглядного методу машинного навчання, розробленого Джеффом Хокінсом і Ділепом Джорджем з Numenta, Inc. Це технологія, яка спрямована на захоплення структурних і алгоритмічних властивостей неокортексу. HTM є теорією неокортексу і має кілька подібних структур мозку. Неокортекс складається з близько 75% об'єму людського мозку, і це ядро більшості того, що вважаємо інтелектом. HTM є біологічною теорією, тобто вона походить від нейроанатомії та нейрофізіології і пояснює, як працює біологічний неокортекс. Кажуть, що теорія HTM «біологічно обмежена» — терміну, який використовується в машинному навчанні.

Метод навчання HTM призначений для роботи з динамічним середовищем як датчиком і моторними даними. Вхід датчика може змінюватися в залежності від різних перешкод, як показники із сервера. Замість цього вхідні дані можуть змінюватися, тому що датчик рухається так само, як людські очі дивляться на статичну картинку.

Системи HTM вчаться он-лайн. Коли будь-які зміни у вхідних даних, оновлюється пам'ять системи HTM. Існує стандартний підхід до машинного навчання, як розділення даних у тестових та навчальних наборах даних, а також навчяння мережі для досягнення зазначеного рівня точності. Однак є питання:

"Якщо немає маркування, як воно знає, що воно працює правильно. Як це може виправити його поведінку.". У будь-який момент часу НТМ робить прогноз того, що очікує, що буде відбуватися далі. Іншими слова, НТМ будує прогностичну модель світу. Потім прогноз порівнюється з тим, що сталося, і цей результат є основою для навчання. Як класичні методи машинного навчання, система намагається звести до мінімуму межу прогнозів. НТМ постійно вчиться. Це невід'ємна якість для біологічного організму, щоб жити. НТМ був побудований навколо цієї ідеї, що інтелектуальні системи повинні постійно вчитися як біологічні організми. Є додатки, де системам не потрібно вчитися он-лайн, але це не норма, а винятки. НТМ видає себе за універсальний метод навчання для будь-якого завдання, яке має біологічну поведінку. Якщо найшли новий підхід, як довести, що він працює.

Багато дослідників, досліджуючи методи без нагляду явища аномалії, стикаються з проблемою — немає зазначених рамок для оцінки їх ефективності. Багато вчених роблять лише теоретичні припущення. В даний час з'явилися деякі спроби створити такі рамки.

Маркус Голдштейн і Сейчі Учіда у своїй роботі [5] пропонують, як слід робити оцінку підходів до явища аномалії. Але методологія їх оцінки орієнтована на методи, які працюють з табличними даними.

На 14-й Міжнародній конференції з машинного навчання та застосування (IEEE ICMLA) 2015 Олександр Лавін і Субутай Ахмад запропонували свою методологію оцінки процедур явища аномалій в реальному часі — Орієнтир аномалії Numenta (NAB)[6].

Це перший орієнтир для процедур явища аномалії серії в часі. NAB - це рамка з відкритим вихідним кодом, яка спрямована на оцінку процедур явища аномалії в режимі реального часу через контрольоване та повторюване середовище. Корпус даних NAB складається з 58 реальних наборів даних, де позначені вікна аномалії. Кожен файл складається з від 2000 до 22000 разів з

даними, агреговані в 5-хвилинні інтервали. Він має в цілому 365551 точок. Ці набори даних є синтетичними та реальними:

- іт-показники;
- соціальні медіа;
- промислові датчики.

Механізм підрахунку очок значення вікна явища аномалії для вимірювання раннього явища. Він підвищує рейтинг алгоритмам, які:

- знайти всі теперішні аномалії;
- знайти їх якомога швидше;
- знайшли кілька з хибно позитивних явищ;
- робота з даними в реальному часі;
- автоматизовані в усіх наборах даних.

Він використовує профілі застосунків для оцінки продуктивності для різних сценаріїв.

Багато компаній, які є основним продуктом якоїсь бази даних, продумують цю тему на конференціях і показують , як можна знайти аномалії за допомогою своїх інструментів. Але проблема полягає в тому, що показаний підхід хороший тільки для цього конкретного набору даних і випадку використання.

Явище аномалій вимагає відповідних значень від аналітика. Існують різні типи аномалій. Існують різні методи пошуку конкретних типів аномалій. Завжди намагаються звузити проблему. І перевіряють свій підхід за допомогою доступних тестів або деяких наборів даних (якщо доступно) для проблеми.

Розглянуті проблеми, що никають при побудові моделей у прогнозному аналізі даних з урахуванням наявності у них аномальних викидів. Обґрунтований вибір методу явища аномалій і його застосування в алгоритмі побудови прогнозної моделі дерева розв'язків. Описані етапи роботи цього алгоритму, методика пошуку аномалій у даних. Наведений опис параметрів настроювання пошуку і їх принципове вплив на результат роботи методики.

Представлені результати сполучення методики пошуку аномалій з алгоритмом побудови моделі дерева розв'язків, виражені в підвищенні точності прогнозної моделі за рахунок підвищення стійкості до викидів у даних.

2.4 Задача явища аномалій при побудові прогнозної моделі прийняття розв'язків

Зазначимо, що використовуються в роботі поняття вихідних даних і шуму. Є вихідна множина інформаційних об'єктів (об'єктів даних) $X = \{X_1, X_2, .., X_n\}$ і множина атрибутів $A = \{A_1, A_2, \ldots, A_k\}$. Кожний об'єкт є кортежем значень атрибутів $X_i = \{a_{i1}, a_{i2}, .., a_{ik}\}$.

Шумом називають перекручені значення атрибутів об'єктів. Об'єкт X_i вважають перекрученим об'єктом, тобто маючим шум, якщо існує такий атрибут A_j , $j = 1, k$ значення a_{ij} якого є перекрученим (маючим шум). Розглянемо шум двох типів:

1) відсутність значень;

2) аномальні значення.

Шум типу «відсутність значення» позначимо як a_{ij} = null. Якщо деякі об'єкти даних мають пропуски в значеннях яких-небудь атрибутів, вважаємо, що ці пропуски не несуть фізичного змісту та маркіруються як шум. Викривлення типу «аномальні значення» можуть мати або не мати фізичного змісту.

В роботі розглянуті аномалії в даних. Цей тип викривлень становлять інтерес у зв'язку зі складністю цього явища в порівнянні із пропуску або викривлення, які легко знайти перебором словника. Аномалії можуть нести фізичний зміст і не бути фактично значимии в даних. Однак при побудові моделі в прогнозному аналізі опираються на фундаментальне припущення про збереження тренда: події або явища, що мали місце в минулому, збережуть імовірність їх появи в майбутньому. При цьому аномалії або викиди в даних розглядають як викривлення, які необхідно знайти та очистити.

Об'єкти генеральної сукупності являють собою екземпляри сутностей, що володіють однаковим набором атрибутів. Значення цих атрибутів аналізують для явища закономірностей усіх об'єктів генеральної сукупності.

Викидами, або аномалія називають такі об'єкти даних, які не задовольняють параметрам, характерним для більшості інших об'єктів генеральної сукупності.

Оскільки кожний об'єкт даних має рядом атрибутів, можна затверджувати про ступінь схожості об'єктів, ґрунтуючись на порівнянні всіх значень відповідних атрибутів цих об'єктів.

Більшість методів пошуку викидів у даних засновані на обчисленні відстаней між об'єкта даних [3]. Метод пошуку викидів, заснований на методі розрахунках показника локальної аномальності *LOF* [7], описаний у роботах [2, 8]. Одна з важливих переваг методу — здатність давати деяку імовірнісну оцінку приналежності кожного об'єкта даних до аномалій.

Це дозволяє більш гнучко оцінювати результат аналізу, на відміну від методів об'єктів, що однозначно означають приналежність до аномалій. У той же час, необхідні інструменти для керування зазначеними перевагами методу розрахунків *LOF*, а саме потрібне створення набору правил оцінки результатів роботи методу. Слід ввести деякі додаткові критерії викидів які ідентифікуються.

Метод розрахунків *LOF* заснований на відомому методі *k* найближчих сусідів, у зв'язку із чим виникає задача вибору параметра *k*. Загальні рекомендації з вибору параметра *k* наведені в роботі [7], у якій запропоновано вибирати параметр *k* окремо для кожної задачі з урахуванням специфіки аналізованих даних, їхньої кількості, прогнозованого числа можливих викидів і т.д.

Висновки до розділу 2

Застосування розглянутої методики при побудові прогнозних моделей дозволяє ефективно визначити викривлення в даних і знижувати вплив шуму на результат роботи систем підтримки прийняття розв'язків.

Одна з найбільш відомих і ефективних моделей у прогнозному аналізі — дерево розв'язків. Ця модель відноситься до виду алгоритмів навчання із учителем, тобто для побудови моделі використовують деяку вибірку інформаційних об'єктів, яку називають навчальною *вибіркою*. Дерева розв'язків організовані у вигляді ієрархічної структури, що складається з вузлів прийняття розв'язків по оцінці певних змінних для прогнозування результуючого значення. Будь-яке дерево розв'язків визначає прогнозоване значення, отримані результати оцінки деяких вхідних атрибутів. Кожний рівень у дереві можна розглядати як одне з розв'язків. Вузол дерева забезпечує перевірку умови, а кожне ребро позначає один з можливих варіантів. Вузли прийняття розв'язків містять критерії вибору, а ребра виражають взаємовиключні результати перевірки відповідності цим критеріям.

Розділ 3

Розробка інформаційної технології сегментованого аналізу

3.1 Метод побудови моделі дерева рішень

Алгоритм побудови моделі дерева рішення наведений у роботі [2]. На першому етапі відбувається вибір стратегії підвищення якості даних відповідно до показників, запропонованих в роботі [9]. На другому етапі відбувається підвищення якості даних по цьому алгоритму заповнення відсутніх атрибутів даних, а також по розглянутому в роботі [8] алгоритму явища аномалій. Далі будується дерево розв'язків за допомогою алгоритму.

Алгоритм явища аномалій, що працює на другому етапі методу у рамках процесу підвищення якості даних, у свою чергу, проводить обробку викидів у два етапи. На першому етапі викиди у даних необхідно ідентифікувати. Для ідентифікації аномалій застосовують метод розрахунків *LOF*. На другому етапі виявлені об'єкти підлягають обробці.

3.2 Методика явища та обробки аномалій

Застосування цієї методики роботи алгоритму побудови прогнозних моделей обумовлене використанням алгоритму пошуку аномалій LOF, який, як було відзначено вище, мають переваги в порівнянні з аналогічними алгоритмами, але вимагає інтерпретації результатів роботи.

Методика заснована на понятті ядра об'єктів навчального множини [11]. Об'єкти аналізу — об'єкти навчального множини, які являють собою кортежі атрибутів даних. Атрибути можуть бути як дискретними, так і безперервними, і в сукупності являють собою кортеж, який розглядається у рамках методики як єдиний об'єкт аналізу. Кожний атрибут кортежами є одиницею даних, а весь кортеж — об'єктом, що мають інформаційну значимість, або вага. Очевидно, що

різні об'єкти аналізу будуть мати різну інформаційна вага, що зменшується з появою шуму в атрибутах цих об'єктів.

Якщо представити об'єкти аналізу сферичними тіла, то можна визначити частоту $f_n\ (a_i)$ появи значення a_i атрибута A_n в об'єктах генеральної сукупності як масу сфери. Чим частіше значення атрибута з'являється серед об'єктів генеральної сукупності, тем «важливіше» дане значення. Дійсно, шум у даних хаотичний випадок, що має здебільшого, виключний характер, представляється як інформаційно більш «легкий» об'єкт.

Введемо параметр ρ, що характеризує щільність об'єктів. Приймемо, що щільність усіх об'єктів однакова. Таке припущення правомірне, оскільки відсутня апріорна інформація про ймовірність виникнення шуму в яких-небудь конкретних атрибутах кортежу даних. Тоді, змінюючи значення, можна регулювати об'єм тіл і, відповідно, займану ними площу в загальному інформаційному просторі W, створеному множиною об'єктів генеральної сукупності у деякому просторі W не порожньо: Якщо перетинання об'єктів a_i, a_j $a_i \cap a_j \neq \varnothing$, то приймемо, що об'єкти належать множині C: $a_i \in$X і $j \in$X. Множина C усіх об'єктів, що мають перетинання, називають ядром у просторі W:

$$C = \left\{ a_1, a_2, \ldots, a_k \middle| \left(\bigcup_{i=1}^{k} \bigcup_{j=1}^{k} (a_i \cap a_j) \right) \neq \varnothing \right\}. \qquad (3.1)$$

Методику явища аномалій виконують у три етапи. На першому етапі розраховують відстані між усіма об'єкта аналізу по формулі, запропонованій у роботі [12]:

$$\mathrm{dist}_{A_n}(a_i, a_j) = \sqrt{\frac{f_n(a_i) + f_n(a_j)}{f_n(a_i)\, f_n(a_j)}}, \qquad (3.2)$$

де A_n - атрибут, який бере значення $D(A_n) = \{a_1, .., a_p\}$; $f_n(a_i)$ - величина, яка визначається прямим підрахунком числа значень a_i атрибута A_n з об'єктів генеральної сукупності.

Обчислюють показники локальної аномальності *LOF* для кожного об'єкта. На другому етапі відбувається автоматичний аналіз середнього показника *LOF* об'єктів ядра:

$$\overline{LOF} = \frac{\sum_{i=1}^{|C|} LOF(x_i)}{|C|}. \quad (3.3)$$

Тут C — потужність множини C, тобто число об'єктів ядра. Якщо предстати множину C на площині, то $S \leq (X)$ — площа фігури C. Визначають відношення площі фігури ядра до загальної площі фігур об'єктів

$$S_{rel} = \frac{S(C)}{S(D(A_n))}. \quad (3.4)$$

Параметр щільності об'єктів ρ зменшується із заданим кроком, який автоматично коректують у міру просування процесу аналізу. При зменшенні щільності площа об'єктів збільшується, нові об'єкти попадають у перетинання, стаючи частиною ядра. Потім знаходять середній показник *LOF* по формулі (3.3) і відношення площ по формулі (3.4). Щільність ρ зменшується доти, поки всі об'єкти не потраплять у ядро, тобто стане справедлива рівність $S_{rel} = 1$.

3.3 Статистичний аналіз вибіркової сукупності

Виявити наявність серед вихідних даних значень, що різко виділяються, («викидів» даних) з метою виключення з вибірки аномальних одиниць спостереження.

Розрахувати узагальнюючі статистичні показники сукупності по досліджуваних ознаках: середню арифметичну ($\tilde{\mathbf{x}}$), моду (Мо), медіану (Ме), розмах варіації (R), дисперсію($\boldsymbol{\sigma}_{\mathbf{n}}^{\mathbf{2}}$), середні відхилення – лінійне ($\bar{d}$) і квадратичне (σ_n), коефіцієнт варіації (V_σ), структурний коефіцієнт асиметрії Пірсона (A_{sn}).

На основі розрахованих показників у припущенні, що розподіл одиниць по обом ознакам близькі до нормального, оцінити:

а) ступінь коливальності значень ознак у сукупності;

б) ступінь однорідності сукупності по досліджуваних ознаках;

в) стійкість індивідуальних значень ознак;

г) кількість влучень індивідуальних значень ознак у діапазони ($\tilde{x} \pm \sigma$), ($\tilde{x} \pm 2\sigma$), ($\tilde{x} \pm 3\sigma$).

Дати порівняльну характеристику розподілів одиниць сукупності по двом досліджуваним ознакам на основі аналізу:

а) варіації ознак;

б) кількісної однорідності одиниць;

в) надійності (типовості) середніх значень ознак;

г) симетричності розподілів у центральній частині ряду.

Побудувати інтервальний варіаційний ряд і гістограму розподілу одиниць сукупності за ознакою середньорічна вартість і встановити характер (тип) цього розподілу. Розрахувати моду Мо отриманого інтервального ряду та зрівняти її з показником Мо незгрупованого ряду даних.

Статистичний аналіз генеральної сукупності. Розрахувати генеральну дисперсію σ_N^2, генеральне середнє квадратичне відхилення σ_N та очікуваний розмах варіації ознак RN. Співставити значення цих показників для генеральної та вибіркової дисперсій.

Для досліджуваних ознак розрахувати:

а) середню помилку вибірки;

б) граничні помилки вибірки для рівнів надійності P=0,683, P=0,954, P=0,997 і границі, у яких будуть перебувати середні значення ознаки генеральної сукупності при заданих рівнях надійності.

Розрахувати коефіцієнти асиметрії As і ексцесу Ek. На основі отриманих оцінок зробити висновок про ступінь близькості розподілу одиниць генеральної сукупності до нормального розподілу.

3.4 Метод пошуку аномалій

Однієї з основних задач є задача пошуку функції корисності або ефективності роботи розглянутої складної системи. Є множина способів знаходження її для конкретних класів. На жаль, практично всі вони засновані на емпіричному доборі. Загальний спосіб теоретичного пошуку цієї функції невідомий.

Оскільки ввели досить багато понять, що стосуються складних систем, встановимо, виходячи з них, вимоги, яким повинна задовольняти найбільш ефективна складна система.

Вимоги ці очевидні:

- складна система повинна досягати поставлені перед нею мету;
- величина роботи, виробленою складною системою, повинна бути мінімальна (різниця між витраченою та корисною роботою повинна прагнути до нуля);

– у процесі здійснення складною системою корисної роботи повинні бути використані максимально всі її можливі ресурси.

Для зручності проведення аналізу роботи складної системи необхідно, щоб показники, що описують ефективність її роботи, задовольняли наступним вимогам:

– були б порівнянні між собою;

– описували б роботу не тільки всієї складної системи в цілому, але та кожного її елемента;

– були б нормованими.

Усім поставленим вище вимогам, на наш погляд, відповідає лише одна можлива конструкція показника ефективності роботи складної нею, що представляє відношення корисної роботи, виконаною самою складною системою, до величини максимальної роботи, яку вони здатні зробити (модифікована оцінка коефіцієнта корисної дії) за параметрами аномалії.

⋮

$$K_{ij} = \frac{\sum A^{n}_{ij..m}}{\sum_{..m} A^{\max}_{ij..m}}; \tag{3.8}$$

$$K_{i} = \frac{\sum A^{n}_{ij..m}}{\sum_{j..m} A^{\max}_{ij..m}}; \tag{3.9}$$

$$K_{j} = \frac{\sum A^{n}_{ij..m}}{\sum_{i..m} A^{\max}_{ij..m}}; \tag{3.10}$$

$$K = \frac{A^{n}}{A^{\max}} = \frac{\sum A^{n}_{ij..m}}{\sum_{ij..m} A^{\max}_{ij..m}}; \tag{3.11}$$

де К, A^{n}, A^{max} - відповідно ефективність, корисна та максимально можлива робота СС;

$K_{ij\ \ m}$, A_{ij}^{n} і A_{ij}^{max} - ефективність, корисна та максимально можлива робота Ефа, обумовленого ознаками i,j m.;

K_i,$K_{j,,}$ K_{ij}, $K_{ij\ \ m}$ - ефективність роботи, що мають признаки i, j, ij, ij…m.

Величину роботи, зробленої будь-яким елементом складної системи, можна визначати через модуль різниці його енергій у початковому та кінцевому стані, тобто

$$A = |Э_н - Э_к|; \qquad (3.12)$$

де А – його робота,

$Э_н$ і $Э_к$ – енергія початкового та кінцевого стану.

Знак модуля використаний для того, щоб вважати роботу завжди позитивною, незалежно від того, зменшується або збільшується енергія при переході з початкового стану в кінцеве.

Величину енергії в загальному виді можна записати у вигляді наступного вираження:

$$Э = \varphi(\lambda_1 \ \ \lambda_n)\psi(\lambda_1 \ \ \lambda_n, \beta_1 \ \ \beta_m)C+B ; \qquad (3.13)$$

де Э - показник результатів роботи,

$\beta_{1...}$ βm - параметри,

$\lambda_{1...}$ λ_n - параметри складної системи,

В - постійна, що залежить від системи відліку,

φ и ψ - деякі функції параметрів та складної системи, причому на функцію.

Завдяки останньому обмеженню надалі нам у більшості випадків не потрібно буде розраховувати повні вираження для енергій усіх елементів при розрахунках эффективностей роботи складних систем, тому що у виразах (3.8)-(3.11) ці функції скорочуються.

З рівняння (3.13) видно, що та константа В при розрахунках величини роботи Ефа виключається, тому тут, і далі не будемо робити ніяких припущень про її природу. Пропорційність енергій величині показника роботи складної системи очевидна, якщо розглядати її як здатність робити роботу та ураховувати, що при виробництві роботи для досягнення будь-якого одиничного результату необхідні витрати однакової кількості енергії. З обліком вищевикладеного вираження (3.12) прийме вид:

$$A = \varphi(\alpha_1 ... \alpha_n)\psi(\alpha_1 ... \alpha_n, \beta_1 ... \beta_m)\left|C^{н} - C^{к}\right|; \qquad (3.14)$$

де $C^{н}$ і $C^{к}$ значення показника роботи в початковому та кінцевому станах.

Порівнюючи отримане вираження із широко відомим у механіці виразом dA = F dq, де F - узагальнена сила, а dq - зміна узагальненої координати, можна затверджувати, що показники, що описують результати роботи складних систем, утворюють фазовий простір її узагальнених координат, а добуток функцій параметрів φ і ψ дають нам величини узагальнених сил, що діють у системі. Надалі, оскільки величини φ, входячи у вирази (3.8)-(3.11), будуть скорочуватися, нас буде цікавити тільки величини ψ, що є скороченими узагальненими силами, що діють із боку відповідних елементівна виробництво ними результатів, або величинами, пропорційними питомим роботам, необхідним для виробництва одиниці відповідного результату (скороченими питомими роботами).

3.5 Етапи роботи методики визначення аномалій

На третьому етапі формується залежність середнього показника локальної аномальності об'єктів ядра *LOF* (S_{rel}) від відношення площ фігури ядра до загальної площі об'єктів. Уся процедура повторюється кілька раз для різних

значень параметра *k*, що характеризує число найближчих об'єктів при розрахунку показника *LOF*.

Описані ще етапи роботи методики можна більш формально записати у вигляді наступної послідовності кроків.

Крок 1. Вихідні дані являють собою набір значень деякого окремо взятого категоріального атрибута, що є підмножиною генеральної сукупності.

Крок 2. По формулах (3.1), (3.2), (3.4) проводять аналіз значень категоріального атрибута. При цьому початкова щільність повинна бути задана з тих міркувань, щоб у момент початку аналізу не існувало перетинань об'єктів (ядро було порожнім). Далі щільність автоматично регулюється в процесі аналізу.

Крок 3. За результатами аналізу даних будують залежність середнього показника *LOF* ядра від відношення площі ядра до сумарної площі всіх об'єктів.

Крок 4. Кроки 2, 3 повторюють кілька раз для різних значень параметра у діапазоні $[1,\ p-1]$, де p — число унікальних значень розглянутого категоріального атрибута. Таким чином, одержують набір залежностей середнього його показника *LOF* ядра від його відносної площі.

Крок 5. У залежності, відповідної до обраного значення параметра *k*, визначають точку *X* початку зростання функції. Викидами вважають точки, що не ввійшли в ядро в точці *X*.

Побудована залежність середнього показника локальної аномальності об'єктів ядра *LOF* від точок ядра до відносної площі фігури ядра.

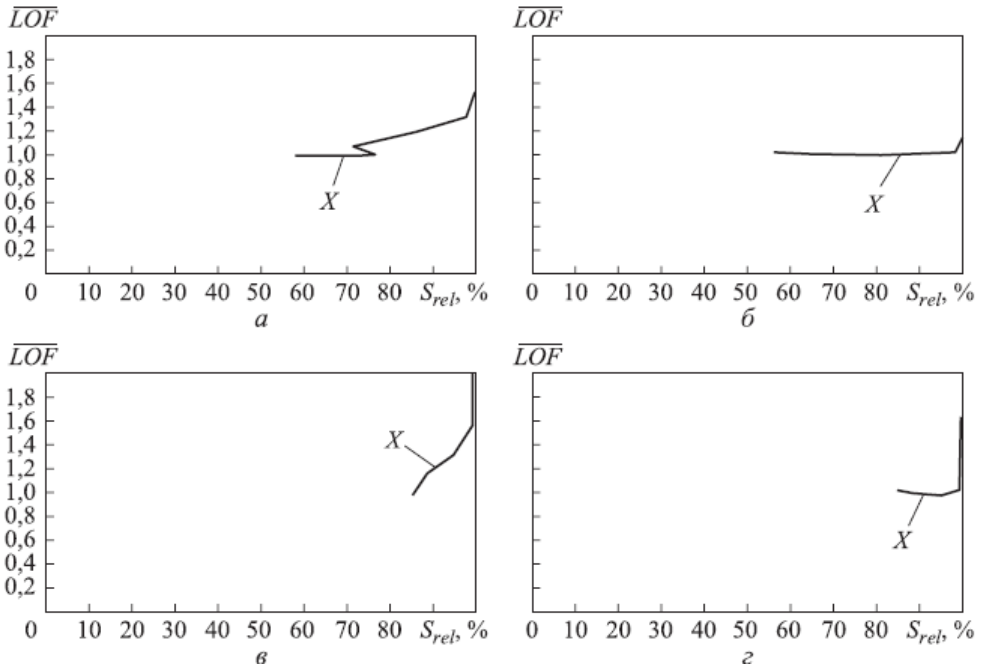


Рисунок 3.1 - Залежність показника локальної аномальності об'єктів ядра *LOF* (S_{rel}) при *k* = 5 (*а*), 10 (*б)*, 2 (*в)* і 4 (*г)* [10]

Аналіз проведений для різних значень параметра *k*. Приклад залежності наведений на рисунку 3.1, додаткові залежності — у роботі [11]. Результати аналізу атрибутів представлені в таблиці.

Таблиця 3.1 - Результати аналізу атрибутів

k	**Показник $\overline{\text{LOF}}$**	**Розкид точок ядра ΔLOF**	**Число викидів**
5	1,003	0,056	9
10	0,999	0,035	4
2	1,000	0,033	4
4	1,001	0,078	2

Для кожної залежності експерт визначає деяку точку *X*, у якій починається зростання функції, показник *LOF* ядра в точці *X*, а також розкид

ΔLOF точок ядра. Точки, що не ввійшли в ядро в точці *X*, вважалися викидами при даному значенні *k*. Результати експериментів підтверджують, що при збільшенні параметра *k* залежність середнього показника локальної аномальності об'єктів ядра від відносної площі фігури ядра стає більш пологою, сигнал про появу викидів з'являється пізніше, тобто більше число об'єктів попадає в ядро та менше точок ідентифікуються як викиди. Таким чином, параметр *k* можна розглядати як «регулятор» ступеня жорсткості ідентифікації викидів. Чим більше значення параметра *k*, тим «м'якше» аналіз і менше об'єктів будуть віднесені до викидів.

Точки, що входять у ядро, мають розкид показника *LOF* у межах однієї десятої. При розширенні границь ядра в нього починають попадати точки, які являються викидами. У даним момент відношення середнього показника *LOF* ядра до його відносної площі починає збільшуватися, що розцінюється побудованою моделлю як сигнал про влучення в ядро потенційного викиду. Отримані висновки дозволяють інтерпретувати значення показника *LOF*, а також гнучко вибирати значення параметра *k* на основі суб'єктивних очікувань експерта засобами нечіткої логіки [11].

Після збору та очищення аномалій побудована по алгоритму ID3O модель перевіряється. Для перевірки точності класифікації використовують підготовлену заздалегідь тестову вибірку, об'єкти якої вже класифіковані експертом. Для оцінки точності застосовують критерій *Errratio*, названий коефіцієнтом помилки класифікатора . Цей критерій визначають як відношення числа невірно класифікованих об’єктів до загального числа об'єктів помилково класифікованих побудовано. моделлю дерева розв'язків.

$$ErrRatio = \frac{|X_f|}{|X|}. \qquad (3.5)$$

Тут X - множина об'єктів в тестовій вибірці; X_f – множина об'єктів, помилково класифікованих побудованої моделлю дерева рішень.

Висновки до розділу 3

У результаті застосування методики явища аномалій вдалось сполучити ефективний метод пошуку викидів у даних *LOF* з алгоритмом побудови моделі дерева розв'язків. Це забезпечило останньому високу стійкість до викривлень у даних одночасно зі значним збільшенням продуктивності системи при побудові моделі. Стійкість до викривлень визначена як зниження точності класифікації при різних рівнях шуму в даних, яке при використанні запропонованої методики виялося суттєво менше зниження точності при застосуванні інших алгоритмів. Збільшення продуктивності становить p /2 раз для кожного атрибута об'єкта даних, де p — число значень атрибута, перевіряемого на аномальність.

За допомогою аналізу порівнюють результати класифікації тестової вибірки, сформованою прогнозною моделлю, з результатами класифікації експертів, які вважаються еталонними.

Розділ 4

Дослідження ефективності визначення аномалій

4.1 Явища аномалії для багатоваріатних даних

Явища аномалії це процес явища неочікуваних елементів або подій у наборах даних, які відрізняються від нор. І явища аномалії часто застосовується на необроблених даних, які відомі як без нагляду явища аномалії. Явища аномалії має два основних припущення:

– аномалії зустрічаються тільки дуже рідко в даних;

– їх особливості суттєво відрізняються від звичайних екземплярів.

Перш ніж дійдемо до явища багатоваріантної аномалії, необхідно працювати над простим прикладом методу явища аномалії, в якому виявляємо викиди з розподілу значень в одному просторі функцій.

Використовуємо набір даних Super Store Sales, який можна завантажити, і збираємося знайти шаблони в продажах і прибутку окремо, які не відповідають очікуваній поведінці. Тобто, явища аутсайдерів по одній змінній за раз.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib
from sklearn.ensemble import IsolationForest
```

Розподіл продажів

```
df = pd.read_excel("Superstore.xls")
df['Sales'].describe()
```

```
count     9994.000000
mean       229.858001
std        623.245101
min          0.444000
25%         17.280000
50%         54.490000
75%        209.940000
max      22638.480000
Name: Sales, dtype: float64
```

Рисунок 4.1 – Розподіл за значеннями

```
plt.scatter(range(df.shape[0]), np.sort(df['Sales'].values))
plt.xlabel('index')
plt.ylabel('Sales')
plt.title("Sales distribution")
sns.despine()
```

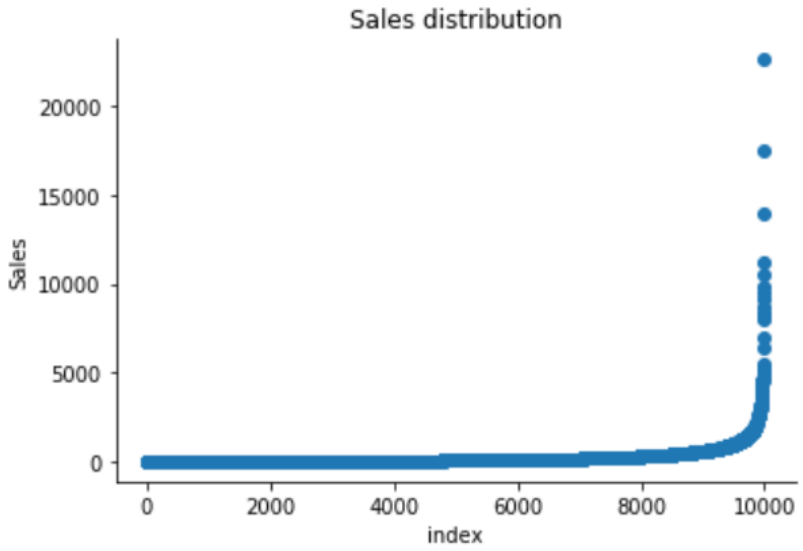


Рисунок 4.2 – Розподіл за продажами

```
sns.distplot(df['Sales'])
plt.title("Distribution of Sales")
sns.despine()
```

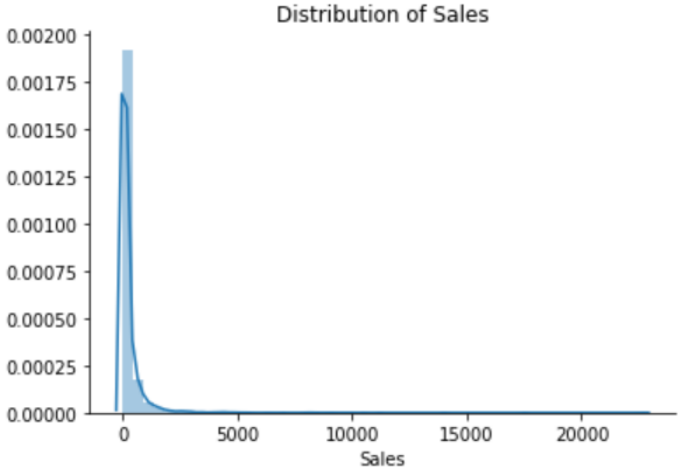


Рисунок 4.3 – Відносний розподіл за продажами

```
print("Skewness: %f" % df['Sales'].skew())
print("Kurtosis: %f" % df['Sales'].kurt())
```

```
Skewness: 12.972752
Kurtosis: 305.311753
```

Розподіл продажів Superstore далекий від звичайного розподілу, і він має позитивний довгий тонкий хвіст, маса розподілу зосереджена зліва від фігури. А розподіл продажів хвоста набагато перевищує хвости нормального розподілу.

Є один регіон, де дані мають низьку ймовірність появи, який знаходиться на правій стороні розподілу.

Розподіл прибутку

```
df['Prt'].desibe()
```

```
count    9994.000000
mean       28.656896
std       234.260108
min     -6599.978000
25%         1.728750
50%         8.666500
75%        29.364000
max      8399.976000
Name: Profit, dtype: float64
```

Рисунок 4.4 – Значення прибутку

```
plt.scatter(range(df.shape[0]), np.sort(df['Profit'].values))
plt.xlabel('index')
plt.ylabel('Profit')
plt.title("Profit distribution")
sns.despine()
```

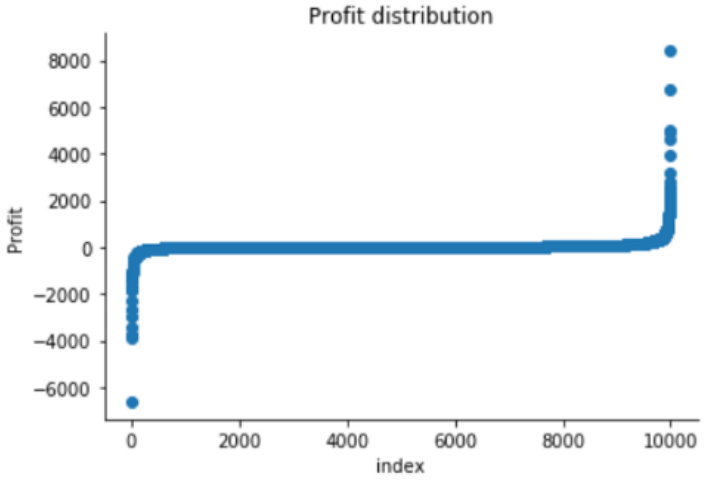


Рисунок 4.5 – Розподіл прибутку

```
sns.distplot(df['Profit'])
plt.title("Distribution of Profit")
sns.despine()
```

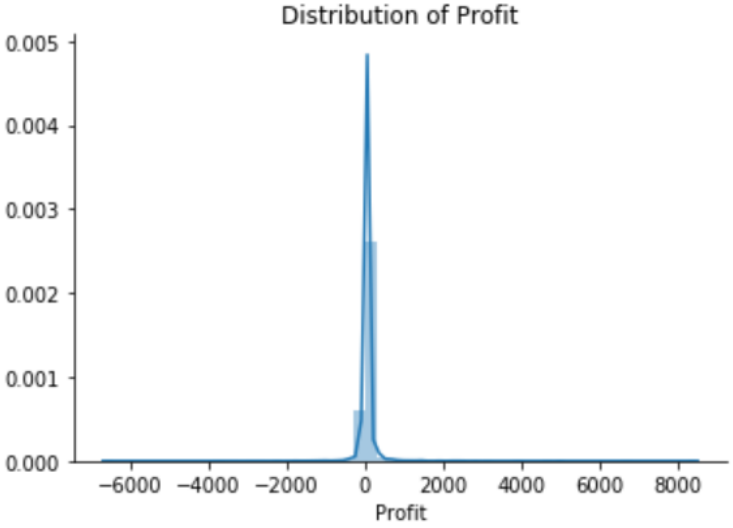


Рисунок 4.6 – Відносний розподіл прибутку

```
print("Skewness: %f" % df['Profit'].skew())
print("Kurtosis: %f" % df['Profit'].kurt())
```

```
Skewness: 7.561432
Kurtosis: 397.188515
```

Розподіл прибутку Superstore має як позитивний, так і негативний хвіст. Однак позитивний хвіст довший за негативний. Отже, розподіл є позитивним перекосом, а дані важкохвості або велика кількість викідів.

Є два регіони, де дані мають низьку ймовірність знаходження: один з правого боку розподілу, інший зліва.

4.2 Явища аномалії на фінансових даних

Ізоляція лісу є алгоритмом для явища викидів, який повертає аномальний бал кожного зразка за допомогою алгоритму IsolationForest, який заснований на тому, що аномалії є точками даних, які є відмінними. Ізольований ліс є моделлю дерева. У цих деревах розділи створюються шляхом випадкового вибору функції, а потім вибору випадкового значення розділення між мінімальним і максимальним значенням вибраної функції.

Наступний процес показує, як IsolationForest поводиться у випадку продажу Susperstore, і алгоритм був реалізований в Sklearn.

Навчений ліс ізоляції за допомогою даних збуту. Продажі в масиві NumPy для використання в моделях.

Обчислили оцінку аномалії для кожного спостереження. Аномальна оцінка вхідного зразка обчислюється як середній показник аномалії дерев у лісі.

Класифікується кожне спостереження як елітне або не-викид. Візуалізація виділяє регіони, де виходять назовні.

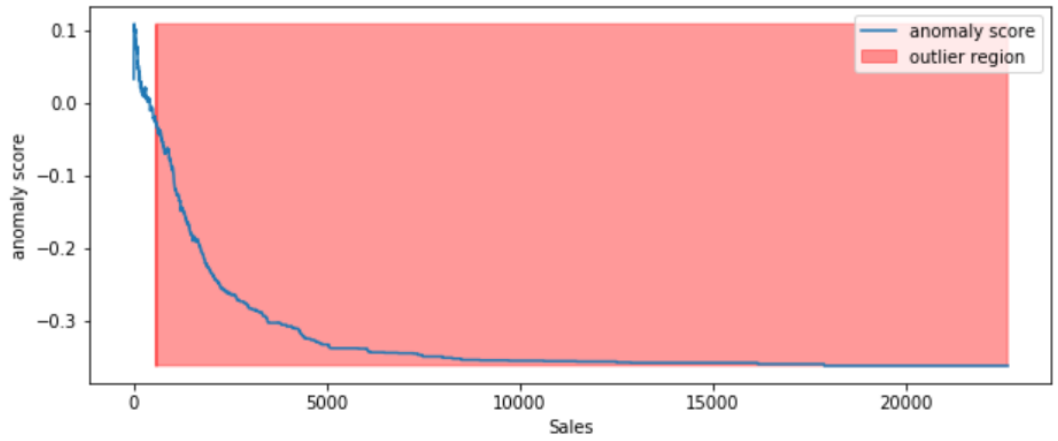


Рисунок 4.7 – Відносні аномалії

Згідно з вищевказаними результатами і візуалізацією, здається, що продажі, які перевищують 1000, безумовно, будуть розглядатися як викид.

Візуально дослідити одну аномалію

```
Row ID                                          11
Order ID                                CA-2014-115812
Order Date                        2014-06-09 00:00:00
Ship Date                         2014-06-14 00:00:00
Ship Mode                            Standard Class
Customer ID                                BH-11710
Customer Name                       Brosina Hoffman
Segment                                    Consumer
Country                               United States
City                                    Los Angeles
State                                    California
Postal Code                                 90032
Region                                       West
Product ID                          FUR-TA-10001539
Category                                 Furniture
Sub-Category                                Tables
Product Name     Chromcraft Rectangular Conference Tables
Sales                                       1706.18
Quantity                                          9
Discount                                        0.2
Profit                                      85.3092
Name: 10, dtype: object
```

Рисунок 4.8 – Дослідження аномалії

Ця покупка здається нормальною та очікуваною, це був більший обсяг продажів в порівнянні з іншими замовлення в даних.

4.3 Вплив явища аномалії на прибуток

Збережемо прибуток в масиві NumPy для використання в наших моделях пізніше.

Обчислемо оцінку аномалії для кожного спостереження. Аномальна оцінка вхідного зразка обчислюється як середній показник аномалії дерев у лісі.

Класифікується кожне спостереження як елітне або не-викид.

Візуалізація виділяє регіони, де виходять назовні дані.

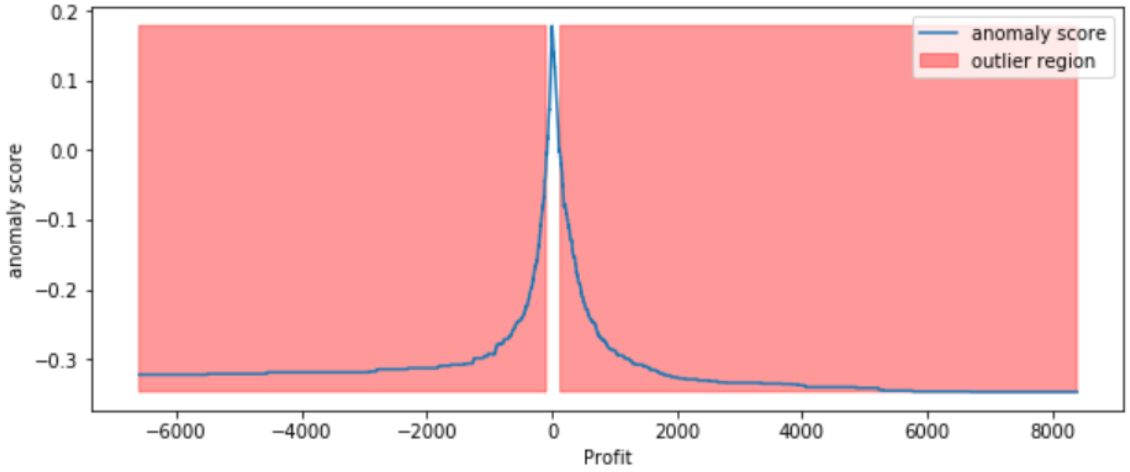


Рисунок 4.9 – Вплив явища аномалії на прибуток

4.4 Візуальне дослідження аномалій

Згідно з вищевказаними результатами і візуалізацією, здається, що прибуток, який нижче -100 або перевищує 100 буде розглядатися як більш аномальний, візуально розглянемо один приклад кожного, який визначається нашою моделлю і щоб побачити, чи мають вони сенс.

```
Row ID                                                        4
Order ID                                           US-2015-108966
Order Date                                    2015-10-11 00:00:00
Ship Date                                     2015-10-18 00:00:00
Ship Mode                                          Standard Class
Customer ID                                              SO-20335
Customer Name                                     Sean O'Donnell
Segment                                                  Consumer
Country                                           United States
City                                             Fort Lauderdale
State                                                    Florida
Postal Code                                                33311
Region                                                     South
Product ID                                       FUR-TA-10000577
Category                                               Furniture
Sub-Category                                              Tables
Product Name     Bretford CR4500 Series Slim Rectangular Table
Sales                                                    957.577
Quantity                                                       5
Discount                                                    0.45
Profit                                                   -383.031
outlier                                                        0
Name: 3, dtype: object
```

Рисунок 4.10 – Значення аномалій

Будь-який негативний прибуток буде аномалія і повинні бути подальші дослідження, що само собою зрозуміло.

```
Row ID                                                        2
Order ID                                           CA-2016-152156
Order Date                                    2016-11-08 00:00:00
Ship Date                                     2016-11-11 00:00:00
Ship Mode                                         Second Class
Customer ID                                              CG-12520
Customer Name                                       Claire Gute
Segment                                                  Consumer
Country                                           United States
City                                                  Henderson
State                                                    Kentucky
Postal Code                                                42420
Region                                                     South
Product ID                                       FUR-CH-10000454
Category                                               Furniture
Sub-Category                                              Chairs
Product Name     Hon Deluxe Fabric Upholstered Stacking Chairs,...
Sales                                                    731.94
Quantity                                                       3
Discount                                                       0
Profit                                                   219.582
Name: 1, dtype: object
```

Рисунок 4.11 – Значення аномалій

В нашій моделі було зазначено, що замовлення з великим прибутком – це аномалія. Однак, коли досліджуємо це замовлення, це може бути просто продукт, який має відносно високий запас.

Наведені ще дві візуалізації показують оцінки аномалії та виділяють регіони, де знаходяться аутсайдери. Як і очікувалося, показник аномалії відображає форму основного розподілу, а більш плоскі регіони відповідають областям низької ймовірності.

Тим не менш, аналіз може отримати тільки по цих даних. Можемо усвідомити, що деякі з цих аномалій, які визначаються наши, не є тими аномаліями, які очікували. Коли дані багатозначні, на відміну від невиявлених**,** підходи до явища аномалії стають більш обчислювально інтенсивними і більш математично складними.

4.5 Багатоваріантні явища аномалії

Більшість аналізу, який в кінцевому підсумку робимо, багатоваріатні через складність даних. У багатоваріатній аномалії явища, викид є комбінованим незвичайним показником принаймні на двох змінних.

Таким чином, використовуючи змінні продажів і прибутку, збираємося побудувати багатоваріатний метод явища аномалії на основі декількох моделей без нагляду.

Використовуємо PyOD, яка є бібліотекою Python для явища аномалій у багатоваріантних даних.

Продажі та прибуток. Очікуємо, що продажі та прибуток позитивно корелюють. Якщо деякі точки даних продажів і точки даних прибутку не є позитивними, вони будуть розглядатися як викиди і повинні бути додатково досліджені.

```
sns.regplot(x="Sales", y="Profit", data=df)
```

```
sns.despine();
```

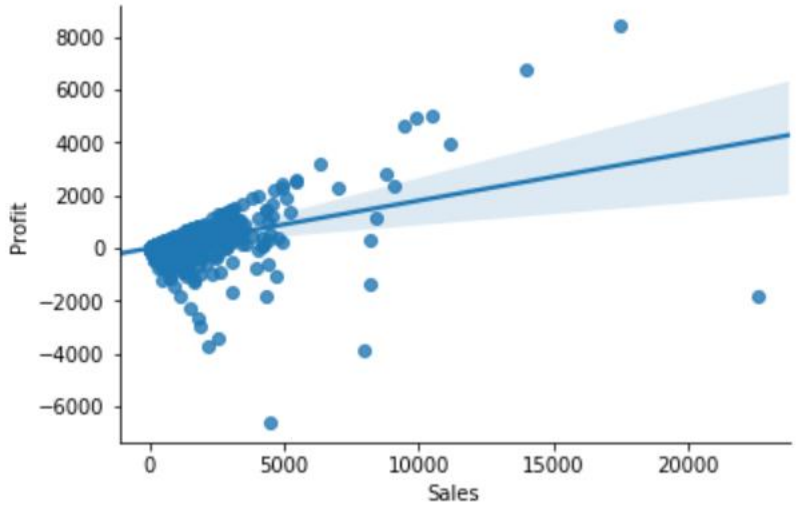


Рисунок 4.12 – Розподіл даних

З наведеної ще діаграми кореляції бачимо, що деякі точки даних є очевидними викидами, такими як екстремально низькі та екстремальні високі значення.

Кластерний локальний фактор викидів (КЛФВ) обчислює показник викид на основі кластерного локального фактора. Оцінка аномалії обчислюється відстанню кожного екземпляра до його кластерного центру, помноженого на екземпляри, що належать до його кластера.

Масштабування продажів і прибутку між нулем і одиницею. Довільно викладені викиди фракції як 1% на основі оціночного розгляду

Припасувати дані до моделі КЛФВ і прогнозувати результати можна таким чином:

– використовувати порогове значення, щоб вважати точку даних більш ін'єктивною .

– використовувати функцію рішення для розрахунку оцінки аномалії для кожної точки.

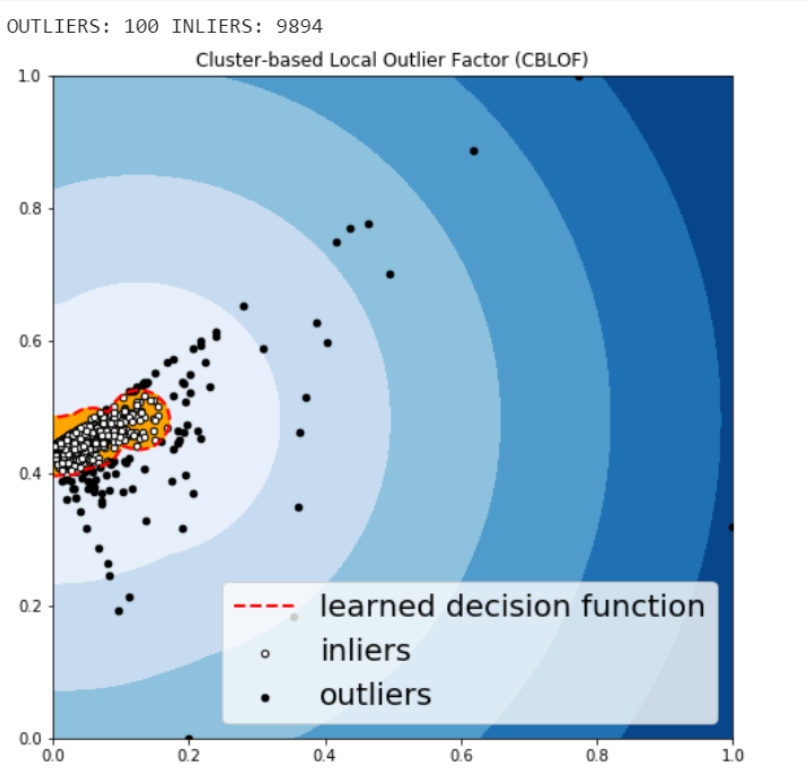


Рисунок 4.13 – Кластерний локальний фактор викидів

Явище аномалій на основі гістограми бере на себе незалежність функції і розраховує ступінь аномалій шляхом побудови гістограм. При багатоваріантних аномалії явища гістограма для кожної окремої функції може бути обчислювана, звизначена індивідуально і об'єднана в кінці.

Ізоляція лісу схожа в принципі на випадковий ліс і побудована на основі рішень дерев. Ізольований ліс ізолює спостереження, випадково обираючи функцію, а потім випадково обираючи розділене значення між максимальним і мінімальним значення цієї обраної функції.

Модуль Isolation Forest є обгорткою Scikit-learn Isolation Forest з більшими функціональними можливостями.

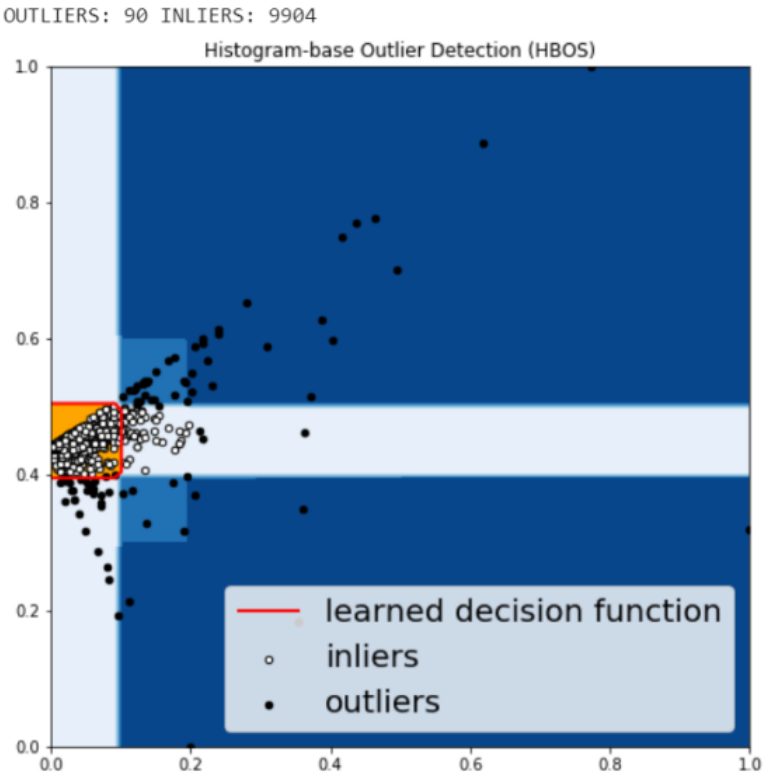


Рисунок 4.14 – Явища аномалій на основі гістограми

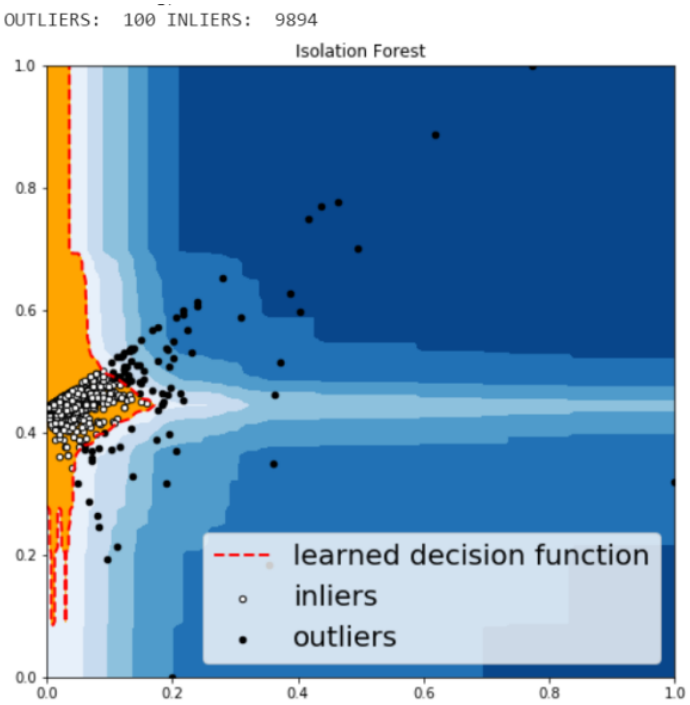


Рисунок 4.15 – Ізольований ліс

KNN є одним з найпростіших методів у виявленні аномалії. Для точки даних його відстань до найближчого сусіда можна розглядати як більшу оцінку аномальності.

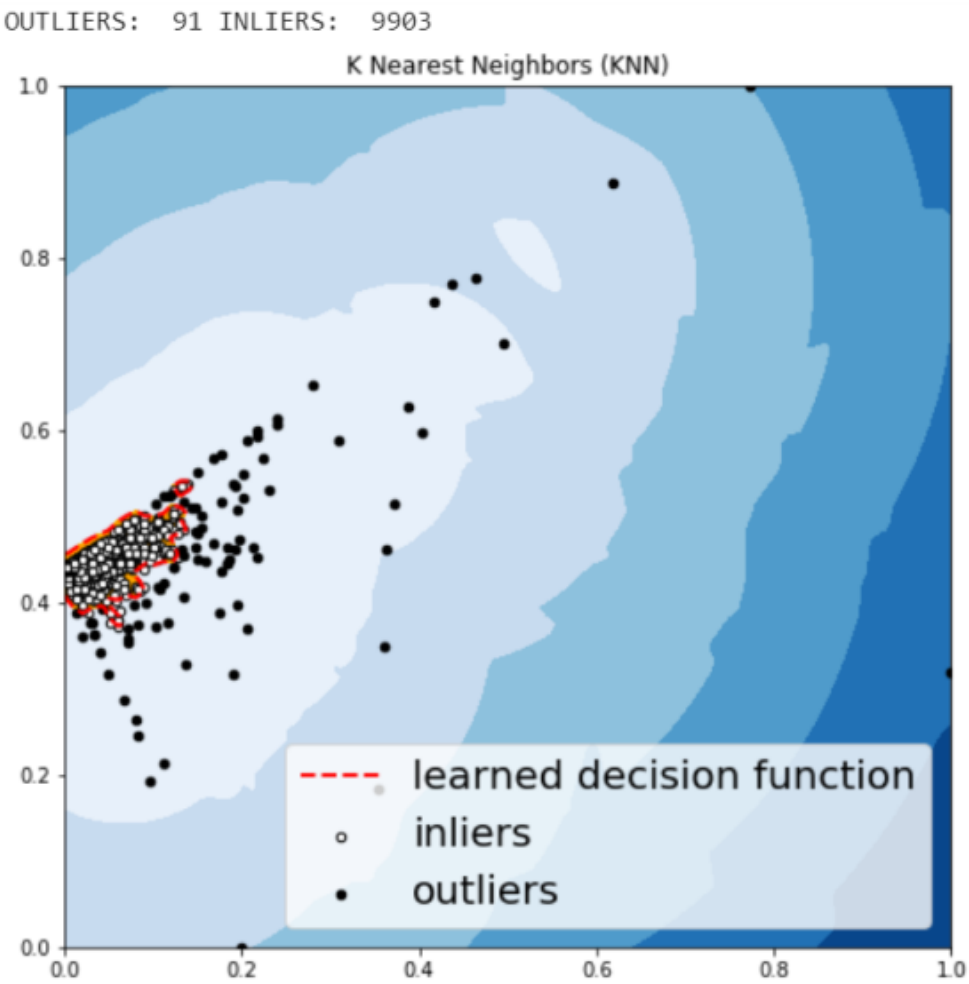


Рисунок 4.16 – K - Найближчі сусіди (KNN)

Аномалії, передбачені вищезазначеними чотирма алгоритмами, не сильно відрізнялися.

Дослідимо кожен з викідів, визначених нашою моделлю, наприклад, детально розглянемо пару викідів, визначених KNN, і дослідимо, що робить їх аномаліями.

```
Row ID                                                      1996
Order ID                                          US-2017-147221
Order Date                                   2017-12-02 00:00:00
Ship Date                                    2017-12-04 00:00:00
Ship Mode                                          Second Class
Customer ID                                            JS-16030
Customer Name                                        Joy Smith
Segment                                               Consumer
Country                                          United States
City                                                   Houston
State                                                    Texas
Postal Code                                              77036
Region                                                 Central
Product ID                                     OFF-AP-10002534
Category                                       Office Supplies
Sub-Category                                        Appliances
Product Name     3.6 Cubic Foot Counter Height Office Refrigerator
Sales                                                   294.62
Quantity                                                     5
Discount                                                   0.8
Profit                                                  -766.012
Name: 1995, dtype: object
```

Рисунок 4.17 – Значення аномалій

Для цього конкретного замовлення клієнт придбав 5 товарів із загальною ціною на 294,62 і прибуток нижче -766, зі знижкою 80%. Необхідно знати про втрату для кожного продукту, який продаємо.

```
Row ID                                                      9650
Order ID                                          CA-2016-107104
Order Date                                   2016-11-26 00:00:00
Ship Date                                    2016-11-30 00:00:00
Ship Mode                                        Standard Class
Customer ID                                            MS-17365
Customer Name                                Maribeth Schnelling
Segment                                               Consumer
Country                                          United States
City                                               Los Angeles
State                                               California
Postal Code                                              90045
Region                                                    West
Product ID                                     FUR-BO-10002213
Category                                             Furniture
Sub-Category                                         Bookcases
Product Name     DMI Eclipse Executive Suite Bookcases
Sales                                                  3406.66
Quantity                                                     8
Discount                                                  0.15
Profit                                                  160.314
Name: 9649, dtype: object
```

Рисунок 4.18 – Значення аномалій

Для цієї покупки здається, що прибуток на рівні близько 4,7% занадто малий і модель зазначила, що це замовлення є аномалією.

```
Row ID                                                9271
Order ID                                    US-2017-102183
Order Date                              2017-08-21 00:00:00
Ship Date                               2017-08-28 00:00:00
Ship Mode                                  Standard Class
Customer ID                                      PK-19075
Customer Name                                Pete Kriz
Segment                                         Consumer
Country                                    United States
City                                       New York City
State                                           New York
Postal Code                                        10035
Region                                              East
Product ID                                 OFF-BI-10001359
Category                                   Office Supplies
Sub-Category                                       Binders
Product Name     GBC DocuBind TL300 Electric Binding System
Sales                                            4305.55
Quantity                                               6
Discount                                             0.2
Profit                                           1453.12
Name: 9270, dtype: object
```

Рисунок 4.19 – Значення аномалій

Для вищевказаного замовлення клієнт придбав 6 товарів загальною ціною 4305, після знижки 20% все одно отримуємо понад 33% прибутку.

Висновки до розділу 4

Розроблювальна система пропонує системний підхід з визначення та аналізу даних та встановлення аномалій в даних. Такі системи можуть допомогти у встановленні нетипових даних та визначення даних, які можуть носити важливий характер. Цей напрямок в дослідженнях важливий з точки зору пошуку нетипових даних особливо з точки зору машинного навчання.

Система показала прийнятний рівень розпізнання аномалій в даних та встановлення зон типових груп даних і формування сталого положення даних.

Для задоволення вимог сучасного бізнесу потрібне автоматичне виявлення аномалій, яке може надавати точну інформацію в режимі реального часу незалежно від того, скільки метрик потрібно відстежувати. Дійсно автоматизовані системи виявлення аномалій повинні включати виявлення,

ранжування та групування даних, усуваючи потребу у великих групах аналітиків.

Загальні висновки

Запропонована методика визначення аномалйх. Отримані аналітичні співвідношення для імовірнісних характеристик відповідних випадкових полів.

Синтезовані алгоритми виявлення аномалій при наявності випадкових перешкод в умовах невідомих параметрів.

Проведений аналіз ефективності виявлення запропонованих і відомих алгоритмів, що дозволяє виробити рекомендації з вибору необхідних значень параметрів для забезпечення заданих характеристик виявлення.

У результаті застосування методики явища аномалій вдалось сполучити ефективний метод пошуку викидів у даних *LOF* з алгоритмами машинного навчання. Це забезпечило останнім високу стійкість до викривлень у даних одночасно зі значним збільшенням продуктивності системи при побудові моделі. Стійкість до викривлень визначена як зниження точності класифікації при різних рівнях шуму в даних, яке при використанні запропонованої методики виявилося суттєво менше зниження точності при застосуванні інших алгоритмів. Збільшення продуктивності становить p /2 раз для кожного атрибута об'єкта даних, де p — число значень атрибута, перевіряемого на аномальність.

З часом світ стає все більш керованим даними і без загального підходу до виявлення аномалії великих даних, проблема високої вимірності неминуча в багатьох областях застосування. Визначення аномальних точок даних у великих наборах даних з проблемами високої розмірності є проблемою дослідження. Це дослідження забезпечило певний огляд методів виявлення аномалій, пов'язаних з великими особливостями обсягу і швидкості, і має; розглянуті стратегії вирішення проблеми високої вимірності. Очевидно, що необхідні подальші дослідження та оцінка стратегій виявлення аномалій великих даних, які умовнюють проблему високої розмірності.

Перелік посилань

1. Song, X., Wu, M. & Jermaine, C., 2007. Conditional Anomaly Detection. IEEE Transactions on Knowledge and Data Engineering, 26 March, 19(5), pp. 631–645.

2. Kawano, H., Nishio, S., Han, J. & Hasegawa, T., 1994. How does knowledge discovery cooperate with active database techniques in controlling dynamic environment.. Athens, Greece, Springer-Verlag Berlin Heidelberg.

3. Hawkins, J., Ahmad, S. & Lavin, A., 2016. Biological and Machine Intelligence.

4. Goldstein, M. & Uchida, S., 2016. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. PLoS ONE, 11(4), p. 31.

5. Lavin, A. & Ahmad, S., 2015. Evaluating Real-time Anomaly Detection Algorithms — the Numenta Anomaly Benchmark. Miami, IEEE.

6. ACFE: Report to the Nations on Occupational Fraud and Abuse, The 2016 GlobalFraud Study. Association of Certified Fraud Examiners (ACFE) (2016),

7. AICPA: Consideration of Fraud in a Financial Statement Audit. American Insti-tute of Certified Public Accountants (AICPA) (2002)

8. Amani, F.A., Fadlalla, A.M.: Data mining applications in accounting: A reviewof the literature and organizing framework. International Journal of AccountingInformation Systems24, 32–58 (2017)

9. An, J.: Variational Autoencoder based Anomaly Detection using Reconstruction Probability. Tech. rep. (2015)

10. Andrews, J.T.A., Morton, E.J., Griffin, L.D.: Detectng Anomalous Data UsingAuto-Encoders. International Journal of Machine Learning and Computing 6(1), 21–26 (2016)

11. Argyrou, A.: Auditing Journal Entries Using Self-Organizing Map. In: Proceedingsof the Eighteenth Americas Conference on Information Systems (AMCIS). pp. 1–10. No. 16, Seattle, Washington (2012)

12. Argyrou, A.: Auditing Journal Entries Using Extreme Vale Theory. Proceedingsof the 21st European Conference on Information Systems1(2013) (2013)

13. Bay, S., Kumaraswamy, K., Anderle, M.G., Kumar, R., Steier, D.M., Blvd, A., Jose,S.: Large Scale Detection of Irregularities in Accounting Data. In: Data Mining,2006. ICDM'06. Sixth International Conference on. pp. 75–86. IEEE (2006)

14. Benford, F.: The Law of Anomalous Numbers. Proceedings of the American Philo-sophical Society78(4), 551–572 (1938)

15. Bengio, Y., Yao, L., Alain, G., Vincent, P.: Generalized denoising auto-encodersas generative models. In: Advances in Neural Information Processing Systems, pp.899–907 (2013)

16. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: Identifying Density-BasedLocal викидs. In: Proceedings of the 2000 Acm Sigmod International Conferenceon Management of Data. pp. 1–12 (2000)

17. Campello, R.J.G.B., Moulavi, D., Sander, J.: Density-Based Clustering Based onHierarchical Density Estimates. Tech. rep. (2013)

18. Cozzolino, D., Verdoliva, L.: Single-image splicing localization throughautoencoder-based anomaly detection. In: 8th IEEE International Workshop onInformation Forensics and Security, WIFS 2016. pp. 1–6 (2017)

19. Das, K., Schneider, J.: Detecting anomalous records in categorical datasets. ACMSIGKDD International conference on Knowledge discovery and data mining pp.220–229 (2007)

20. Dau, H.A., Ciesielski, V., Song, A.: Anomaly Detection Using Replicator Neu-ral Networks Trained on Examples of One Class. In: Asia-Pacific Conference onSimulated Evolution and Learning. pp. 311–322 (2014)

21. D’Avino, D., Cozzolino, D., Poggi, G., Verdoliva, L.: Autoencoder with recurrentneural networks for video forgery detection. arXiv preprint (March), 1–8 (2017)

22. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A Density-Based Algorithm for Dis-covering Clusters in Large Spatial Databases with Noise. In: Proceedings of the 2ndInternational Conference on Knowledge Discovery and Data Mining. pp. 226–231(1996)

23. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforwardneural networks. Proceedings of the 13th International Conference on ArtificialIntelligence and Statistics (AISTATS)9, 249–256 (2010)

24. Hawkins, S., He, H., Williams, G., Baxter, R.: викид Detection Using ReplicatorNeural Networks. In: International Conference on Data Warehousing and Knowl-edge Discovery. pp. 170–180. No. September, Springer Berlin Heidelberg (2002)

25. Hinton, G.E., Salakhutdinov, R.R.: Reducing the Dimensionality of Data withNeural Networks. Science 313(5786), 504–507 (2006)

26. IFAC: International Standards on Auditing 240, The Auditor’s ResponsibilitiesRelating to Fraud in an Audit of Financial Statements. International Federationof Accountants (IFAC) (2009)

27. Schreyer, Sattarov et al.23. Islam, A.K., Corney, M., Mohay, G., Clark, A., Bracher, S., Raub, T., Flegel,U.: Fraud detection in ERP systems using Scenario matching. IFIP Advances inInformation and Communication Technology330, 112–123 (2010)

28. Jans, M., Lybaert, N., Vanhoof, K.: Internal fraud risk reduction: Results of adata mining case study. International Journal of Accounting Information Systems11(1), 17–41 (2010)

29. Jans, M., Van Der Werf, J.M., Lybaert, N., Vanhoof, K.: A business process miningapplication for internal transaction fraud mitigation. Expert Systems with Appli-cations38(10), 13351–13359 (2011)

30. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv preprintpp. 1–15 (2014)

31. Lauly, S., Larochelle, H., Khapra, M.: An Autoencoder Approach to LearningBilingual Word Representations. In: Advances in Neural Information ProcessingSystems. pp. 1853–1861 (2014)

32. LeCun, Y., Bengio, Y., Hinton, G., Y., L., Y., B., G., H.: Deep learning. Nature521(7553), 436–444 (2015)

33. Paula, E.L., Ladeira, M., Carvalho, R.N., Marzag˜ao, T.: Deep learning anomalydetection as suppor fraud investigation in Brazilian exports and anti-money laun-dering. In: Proceedings - 2016 15th IEEE International Conference on MachineLearning and Applications, ICMLA 2016. pp. 954–960 (2017)

34. Wells, J.T.: Corporate Fraud Handbook: Prevention and Detection. John Wiley &Sons (2017)

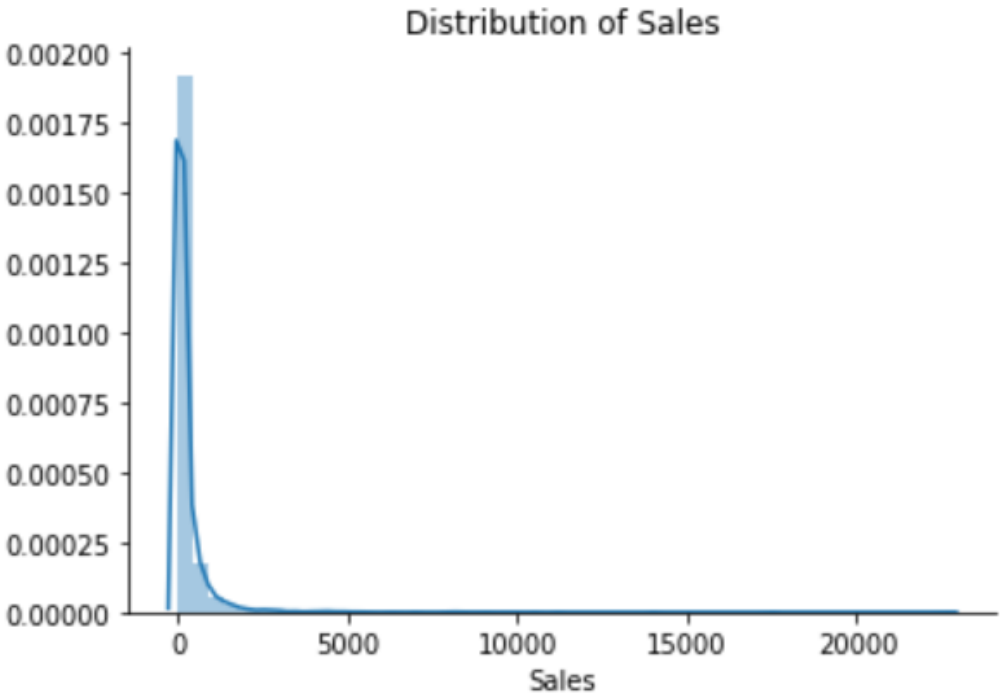
35. Xu, B., Wang, N., Chen, T., Li, M.: Empirical Evaluation of Rectified Activationsin Convolution Network. ICML Deep Learning Workshop pp. 1–5 (2015)

36. Zhai, S., Cheng, Y., Lu, W., Zhang, Z.: Deep Structured Energy Based Models forAnomaly Detection. In: International Conference on Machine Learning. vol. 48,pp. 1100–1109 (2016)

37. Zhou, C.: Anomaly Detection with Robust Deep Auto-encoders. In: KDD ’17 Pro-ceedings of the 23rd ACM SIGKDD International Conference on Knowledge Dis-covery and Data Mining. pp. 665–674 (2017)

38. Chandola, V., Banerjee, A. & Kumar, V., 2009. Anomaly Detection: A Survey. ACM Computing Surveys, July. 41(3).

Додатки




```
print("Skewness: %f" % df['Sales'].skew())
print("Kurtosis: %f" % df['Sales'].kurt())
```

УДК 004.4

Гордійчук Б. Г., Манзюк Е. А., Скрипник Т. К.

Хмельницький національний університет

ВИЯВЛЕННЯ АНОМАЛІЙ В ДАНИХ

Розроблений і реалізований метод виявлення аномалій та викидів в даних фінансової сфери. Розроблювальна система пропонує системний підхідз визначення та аналізу даних та встановлення аномалій в даних Такі системи можуть допомогти у встановлення нетипових даних та визначення даних, які можуть носить важливий характер. Цей напрямок в дослідженнях важливий з точки зору пошуку нетипових даних особливо з точки зору машинного навчання.

A method for detecting anomalies and emissions in financial data has been developed and implemented. The development system offers a systematic approach to identifying and analyzing data and identifying anomalies in data. Such systems can help identify atypical data and identify data that may be important. This direction in research is important from the point of view of search of atypical data especially from the point of view of machine learning.

У процесі виробництва на всіх стадіях і циклах ділової активності працівники підприємства (навмисно або не навмисно) можуть прибігати до викривлень і фальсифікації. У результаті таких дій підприємства можуть мати більші матеріальні й моральні втрати, тому однієї з основних задач аудита є явища помилок і фактів шахрайства, а також здійснення необхідних заходів щодо попередження можливих втрат підприємства. . Для ефективної роботи з явища й усуненню помилок і зловживань в 1982 р. був розроблений і затверджений міжнародний норматив аудита "Обман і помилка" Комітету з аудиторської практики і інструкція "Відповідальність аудиторів у зв'язку зі зловживання, іншими аномаліями й помилками". З 1 січня 1999 р. в Україні придбав силу національний норматив аудита № 7 " Про помилки й шахрайстві". Метою цього нормативу є "зобов'язання тлумачення й використання термінів "шахрайство" і "помилка" з позицій підготовки аудиторського висновку, значення ризику аудита й впливу шахрайства й помилок на вірогідність фінансової звітності клієнта".

Отже, як аналітик даних, можете впровадити явища аномалії за допомогою машинного навчання. А які методи і переваги явища аномалії з використанням технологій глибокого навчання.

Викид ідентифікується як будь-який об'єкт даних або точка, яка значно відхиляється від решти точок даних. У сумі даних аутсорсинг зазвичай відкидається як виняток або просто шум. Тим не менш, те ж саме не може бути зроблено при виявленні аномалії, отже, акцент на аналізі викид.

Типи аномалій. Аномалії можна класифікувати за трьома категорія:

Точку аномалії. Якщо один об'єкт можна спостерігати проти інших об'єктів як аномалія, це точка аномалії. Це найпростіша категорія аномалії і багато досліджень включають їх. Беручи до уваги приклад, представлений на рисунку 2.1 O_1 O_2 є точкою аномалії.

Контекстні аномалії. Якщо об'єкт є аномальним у значеному контексті. Тільки в цьому випадку це контекстна аномалія (також відома як умовна аномалія). На рисунку 2.2 можна побачити періодичний контекст. У цьому випадку точка O_1 аномалія, тому що вона відрізняється від періодичного контексту.

Колективні аномалії. Якщо деякі зв'язані об'єкти можна спостерігати проти інших об'єктів як аномалія. О1 об'єкт не може бути аномальним в цьому випадку, тільки колекція об'єктів.

Розглянуті проблеми, що никають при побудові моделей у прогнозному аналізі даних з урахуванням наявності у них аномальних викидів. Обґрунтований вибір методу явища аномалій і його застосування в алгоритмі побудови прогнозної моделі дерева розв'язків. Описані етапи роботи цього алгоритму, методика пошуку аномалій у даних. Наведений опис параметрів настроювання пошуку і їх принципове вплив на результат роботи методики. Представлені результати сполучення методики пошуку аномалій з алгоритмом побудови моделі дерева розв'язків, виражені в підвищенні точності прогнозної моделі за рахунок підвищення стійкості до викидів у даних.

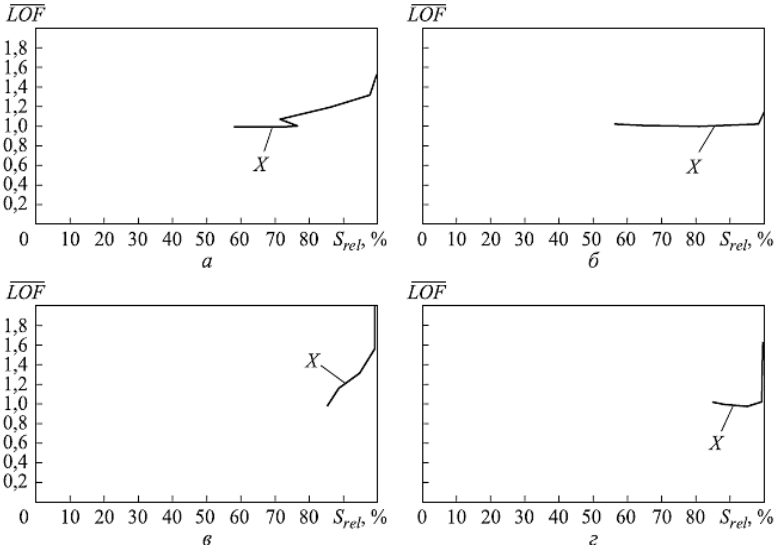


Рисунок 1 – Залежність показника локальної аномальності об'єктів ядра *LOF* (S_{rel}) при k = 5 (*а*), 10 (*б)*, 2 (*в)* і 4 (*г)*

Одна з найбільш відомих і ефективних моделей у прогнозному аналізі — дерево розв'язків. Ця модель відноситься до виду алгоритмів навчання із учителем, тобто для побудови моделі використовують деяку вибірку інформаційних об'єктів, називану навчальною вибіркою. Дерева розв'язків організовані у вигляді ієрархічної структури, що складається з вузлів прийняття розв'язків по оцінці певних змінних для прогнозування результуючого значення.

У результаті застосування методики явища аномалій вдалось сполучити ефективний метод пошуку викидів у даних LOF з алгоритмом побудови моделі дерева розв'язків. Це забезпечило останньому високу стійкість до викривлень у даних одночасно зі значним збільшенням продуктивності системи при побудові моделі.

Перелік посилань

1. Song, X., Wu, M. & Jermaine, C., 2007. Conditional Anomaly Detection. IEEE Transactions on Knowledge and Data Engineering, 26 March, 19(5), pp. 631–645.
2. Kawano, H., Nishio, S., Han, J. & Hasegawa, T., 1994. How does knowledge discovery cooperate with active database techniques in controlling dynamic environment.. Athens, Greece, Springer-Verlag Berlin Heidelberg.
3. Hawkins, J., Ahmad, S. & Lavin, A., 2016. Biological and Machine Intelligence.
4. Goldstein, M. & Uchida, S., 2016. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. PLoS ONE, 11(4), p. 31.

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ДИПЛОМНА РОБОТА МАГІСТРА

Виявлення аномалії в бухгалтерському звіті на базі штучного інтелекту

Розробив ст. гр. КНм-19-1:
Гордійчук Б.Г.

Хмельницький - 2020

В магістерській роботі розроблений і реалізований метод виявлення аномалій та викидів в даних фінансової сфери.
Розроблювальна система пропонує системний підхід з визначення та аналізу даних та встановлення аномалій в даних.
Такі системи можуть допомогти у встановленні нетипових даних та визначення даних, які можуть носити важливий характер.
Цей напрямок в дослідженнях важливий з точки зору пошуку нетипових даних особливо з точки зору машинного навчання.

Метою дослідження є розробка методу аналізу та визначення нетипових даних з використанням штучного інтелекту на базі фінансових даних.

Для досягнення зазначеної мети поставлені наступні **задачі**:

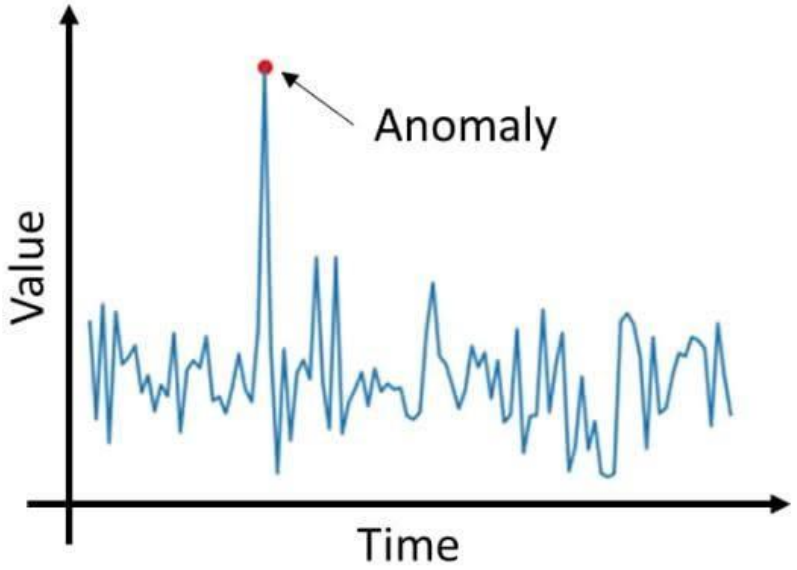
- показати, що використання засобів машинного навчання дає змогу визначити натипові дані;
- провести дослідження визначення ознак виявлення аномалій в даних;
- провести порівняння застосовності методів машинного навчання базуючись на результатах досліджень.

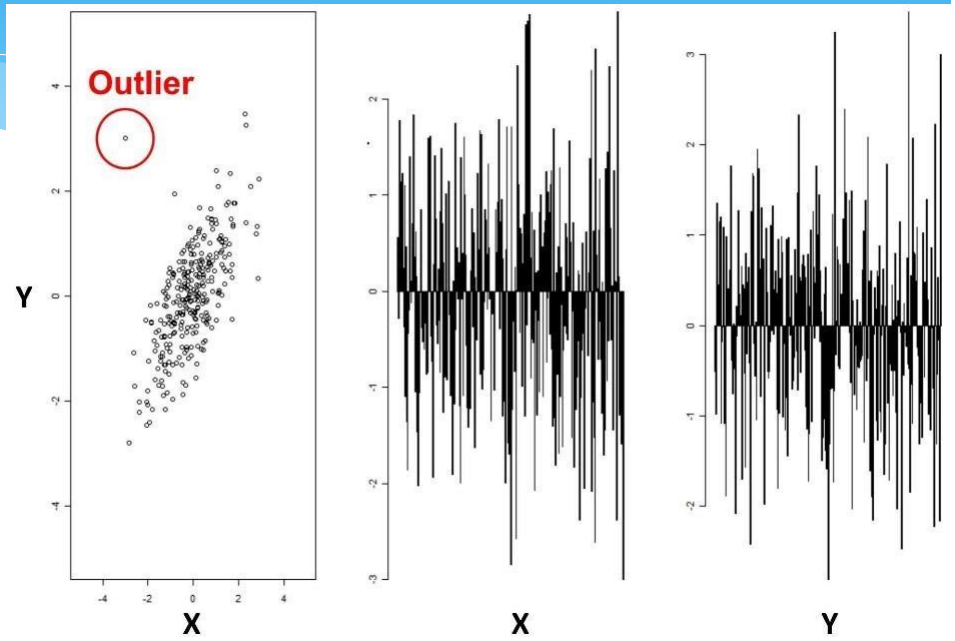
Об'єктом дослідження є порівняння методів визначення нетипових даних на аномалії та викиди.

Предметом дослідження є набір даних фінансового характеру а також ознаки та класифікатори даних машинного навчання

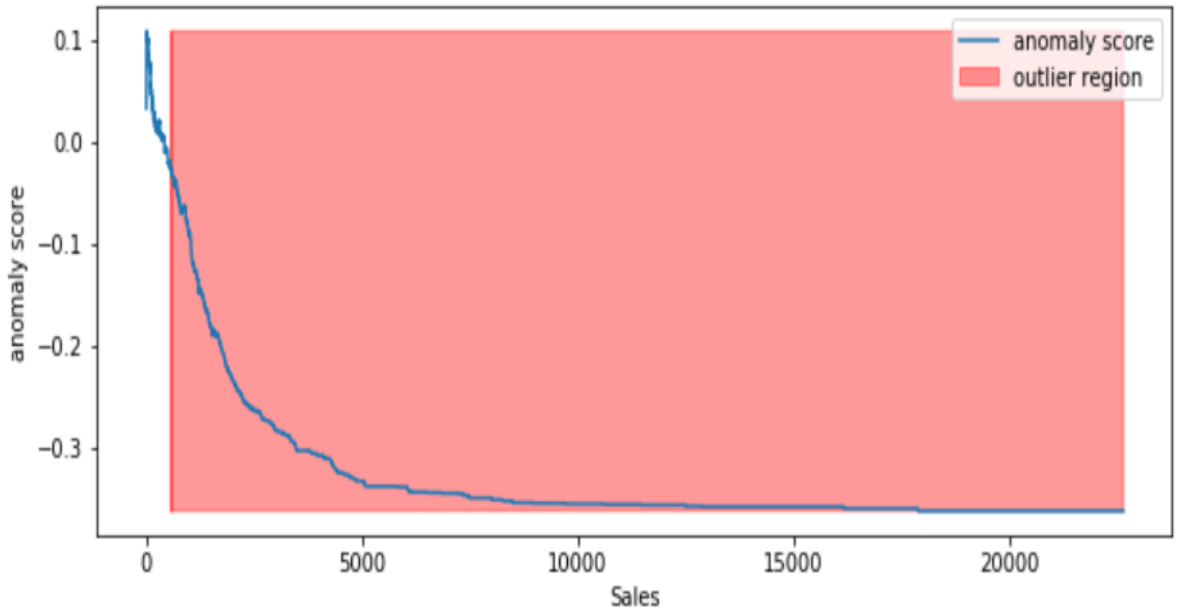
Існуючі системи розпізнавання та визначення нетипових даних розвинуті в недостатній мірі. Тому проведені дослідження з визначенням найбільш ефективних методів та підходів по визначенню та інтерпретації даних. Дані можуть бути як аномаліями, викидами так і новими даними. Визначення причин та факторів впливу є важливим фактором

Практична значимість дослідження полягає в тому, що описані методи та отримані результати застосовуються для виявлення аномалій в даних. Сфера практичного застосування може бути розповсюджена на дослідження даних будь-якого типу.

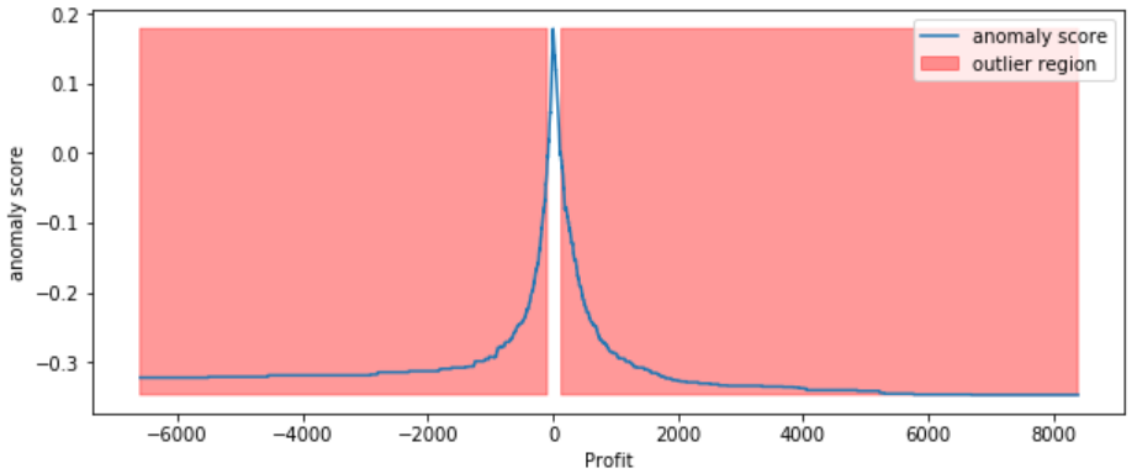




Аномалії в даних



Відносні аномалії



Вплив явища аномалії на прибуток

Висновки

Розроблювальна система пропонує системний підхід з визначення та аналізу даних та встановлення аномалій в даних. Такі системи можуть допомогти у встановленні нетипових даних та визначення даних, які можуть носити важливий характер. Цей напрямок в дослідженнях важливий з точки зору пошуку нетипових даних особливо з точки зору машинного навчання. Система показала прийнятний рівень розпізнання аномалій в даних та встановлення зон типових груп даних і формування сталого положення даних.

Для задоволення вимог сучасного бізнесу потрібне автоматичне виявлення аномалій, яке може надавати точну інформацію в режимі реального часу незалежно від того, скільки метрик потрібно відстежувати.

Дякую за увагу

Mon Nov 30 13:03:54 EET 2020, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 1.0%

Словники перевірки: en_US, ru_RU, ua_UA. **Помилок в документах: 4%**

	Документ		Сумарний збіг по Базі Даних	
ID: 81684 Назва: Виявлення аномалів в бухгалтерському звіті на базі штучного інтелекту Додано в БД: 2020-11-30 Автора: Гордійчук Богдан Геннадійович Керівники: Манзюк Е.А. Консультанти: Опоненти:	Символи	Лексеми	Символи	Лексеми
	71705	595	1229 (2%)	13 (2%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

РІШЕННЯ КАФЕДРИ **КОМП'ЮТЕРНИХ НАУК ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ** ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Підтверджуємо ознайомлення з результатом звіту подібності щодо роботи, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: **Виявлення аномалії в бухгалтерському звіті на базі штучного інтелекту**
Автор: **Гордійчук Б. Г.**
Спеціальність: **122 Комп'ютерні науки**
Науковий керівник: **к.т.н. доцент Манзюк Е.А.**

Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом (далі – зазначаються підстави віднесення запозичень до правомірних). Робота приймається до захисту.	**відповідає**
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи (далі – зазначаються детальні та аргументовані підстави віднесення запозичень до правомірних). Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи	-
3	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	-
4	Інше:	-

Підтвердження: Виявленні запозичення не є плагіатом так як не описують безпосередньо авторське дослідження і є широко вживаними поняттями предметної області і складають 4.8%.

01.11.2020
Дата

Підпис керівника

Підпис завідувача кафедри

Хмельницький національний університет
Факультет програмування та комп'ютерних і телекомунікаційних систем
Кафедра комп'ютерних наук та інформаційних технологій

ВІДГУК ОПОНЕНТА

на дипломну роботу магістра

Магістра *гр. КНм-19-1* Гордійчука Богдана Геннадійовича

На тему: Виявлення аномалії в бухгалтерському звіті на базі штучного інтелекту

1. Актуальність і значення теми

В роботі розроблений і реалізований метод виявлення аномалій та викидів в даних фінансової сфери. Розроблювальна система пропонує системний підхід з визначення та аналізу даних і встановлення аномалій в даних. Таким чином, представляється досить актуальною задача розробки та дослідження алгоритмів виявлення аномалій в умовах апріорної невизначеності відносних даних.

2.Оцінка якості та достовірності проведених досліджень.

Достовірність результатів забезпечується проведенням всебічного оцінювання та порівняння ефективності різних методів.

3. Оцінка запропонованих заходів та пропозицій, практичної цінності та ефективності.

Практична значимість дослідження полягає в тому, що розглядається застосування адаптивних алгоритмів компенсації коррельованих перешкод і алгоритмів виявлення аномалій. У результаті застосування методики явища аномалій вдалось сполучити ефективний метод пошуку викидів у даних з алгоритмами навчання.

4. Загальний висновок та оцінка

Робота виконана в повному обсязі. Пояснювальна записка оформлена в відповідності з нормами. Відмічені недоліки не знижують цінності дипломної роботи. За своєю структурою, практичними цінностями, поставленій меті та вирішеними задачами робота відповідає вимогам вищої школи і вимогам, що пред'являються до освітньо-кваліфікаційного рівня «магістр», а її автор заслуговує присвоєння кваліфікації магістра з комп'ютерних наук та інформаційних технологій.

Робота заслуговує на оцінку « *задовільно* »

Опонент *Мартинюк В.В., д.т.н., проф.* _____