МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ХЕРСОНСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ

Факультет комп'ютерних наук, фізики та математики

Кафедра комп'ютерних наук та програмної інженерії

Аналіз даних для прогнозування курсу акцій з використанням фінансових показників на основі бізнес-звітів

Кваліфікаційна робота (проєкт)

на здобуття ступеня вищої освіти «магістр»

Виконав: студент 2 курсу 261М групи
Спеціальності
126 «Інформаційні системи та технології»
(шифр, назва)
Освітньо-професійної програми:
«Інформаційні системи та технології»
(назва)
Іванов Олексій Юрійович
Керівник: доктор економічних наук, професор Кобець В.М.
Рецензент: Калініченко І.В.
Senior Softwarengineer
IT компанії Softserve

Івано Франківськ – 2023

ЗМІСТ

ВСТУП

Актуальність роботи. У сучасному світі фондовий ринок стає все більш популярним. Це пов'язано з кількома факторами, в тому числі збільшеною доступністю ринку для звичайних користувачів. Сьогодні будь-хто в світі може легко придбати акції будь-якої компанії. Необхідно тільки відкрити рахунок в брокерській фірмі. На цей момент однією з найпопулярніших брокерських фірм серед звичайних користувачів є Interactive Brokers [1] і Robinhood [2]. Обидві програми забезпечують широкий доступ до фондового ринку та інших фінансових інструментів, включаючи ETF, облігації тощо. Ґрунтуючись на статистиці використання Robinhood [2], визначено, що понад 15 мільйонів людей використовують торгівельні платформи. Розширення доступу до фондового ринку призводить до зростання інтересу до інвестування та збільшення кількості нових інвесторів. Однак інвестування на фондовому ринку не позбавлене ризиків і вимагає якісної підготовки та аналізу. Щоб захистити та збільшити свої інвестиції, інвестори мають ретельно проаналізувати компанію, у яку вони планують інвестувати. Якщо говорити про ринок США, який на цю мить є одним із найбільших ринків у світовій економіці, усі публічні компанії на цьому ринку регулюються Комісією з цінних паперів і бірж. (SEC), яка вимагає від компаній подавати щоквартальні звіти про свій бізнес. Ці звіти називаються формою 10-К [3]. Вони є публічними та доступними для інвесторів та аналізу, і на основі даних, наданих у цих звітах, можна оцінити стан справ у компанії, наскільки добре налагоджений їхній бізнес, і відповідно приймати рішення.

Об'єкт дослідження - методи аналізу даних фінансових звітів для дослідження фінансових показників компаній.

Предмет дослідження - прогнозування фінансових показників компаній на основі аналізу бізнес звітів за допомогою методів машинного навчання.

Мета дослідження – провести аналіз даних для прогнозування курсу акцій з використанням фінансових показників на основі бізнес-звітів засобами машинного навчання.

Завдання.

1. Огляд існуючих напрацювань у сфері прогнозування фінансових показників методами машинного навчання.
2. Дослідити та описати основні фінансові показники бізнес звітів.
3. Дослідити та описати методи машинного навчання для проведення аналізу бізнес звітів.
4. Розробити та побудувати модель машинного навчання для прогнозування курсу акцій на основі бізнес звітів.

Методи та технології дослідження. Алгоритми та методи машинного навчання.

Апробація.

Апробація роботи проведена на PhD симпозіумі Міжнародної конференції ICTERI 2023, за результатами якої підготовлена стаття у збірник CCIS видавництва Springer.

Структура роботи. Дослідження складається зі вступу, трьох розділів, висновку та списку використаних джерел. Перша частина аналізує фінансові показники у бізнес звітах. Друга частина описує методи машинного навчання для аналізу бізнес звітів. Третя частина демонструє технології та процес реалізації моделі для прогнозування курсу акцій для прийняття інвестиційних рішень.

РОЗДІЛ 1
ФІНАНСОВІ БІЗНЕС ЗВІТИ ТА ЇХ ПОКАЗНИКИ

1.1. Огляд досліджень у сфері прогнозування фінансових показників

У сфері аналізу фінансових даних та прогнозування фінансових показників за допомогою методів машинного навчання (ML) та аналізу даних було проведено численні дослідження. Деякі з них ми розглянемо в цьому розділі.

Zanc та ін. [4] досліджували застосування глибоких нейронних мереж для прогнозування фінансового ринку, вони використовували фондовий індекс з даними про валютний курс і продемонстрували, що глибокі нейронні мережі дозволяють забезпечити високу точність прогнозування. Вассербахер та ін. [5] досліджували застосування методів машинного навчання для аналізу та прогнозування фінансових даних. Їх дослідження використовувало дані про акції компаній і показало, що методи машинного навчання можна ефективно використовувати для прогнозування фінансових показників. Доряб та ін. [6] вивчали застосування регресійного аналізу для прогнозування прибутку від інвестицій. У своєму дослідженні вони використали дані про прибутковість акцій компанії та продемонстрували, що регресійний аналіз можна ефективно використовувати для прогнозування прибутку від інвестицій. Сніговий та ін. [7] прогнозували ціни на криптовалюту за допомогою різних алгоритмів ML. Усі ці дослідження показують, що методи машинного навчання та аналізу даних можна ефективно використовувати для прогнозування фінансових показників. На відміну від попередніх досліджень, у цій статті ми застосуємо ці методи для аналізу фінансових даних і прогнозування фінансових показників на основі бізнес-звітів.

Mushtaq та ін. [8] використовують обробку природної мови (NLP), підгалузь ШІ, щоб передбачити настрої інвесторів під час аналізу 3729 річних 10-К фінансових звітів компаній S&P 500 за 2002–2019 роки. Вони показали, що немає статистично значимого зв’язку між показниками фінансової

діяльності фірми та позитивністю звітів 10-К [8]. Ми вважаємо, що чим більше звітів, тим менша кореляція між звітами 10-К і фінансовими показниками, оскільки фондові ринки сильніше реагують на негативні результати звітів, ніж на позитивні, як описано Huang та ін. [9].

Річні звіти фірми допомагають інвесторам прийняти рішення щодо капіталовкладень в акції компаній. Як правило, інвестори аналізують фінансові дані, щоб передбачити ціни акцій і майбутні доходи, волатильність і ризики. При цьому фінансові показники можуть вплинути на обсяг та зміст 10-K звіту компанії10-K звіту компанії.

Звіти 10-K виступають сигналом, який може розкрити позитивні показники компанії, використовуючи складний і заплутаний виклад змісту [10]. Водночас фінансові показники можуть розкривати реальний стан компанії без маніпуляцій з боку агентів (керівників), які намагаються зберегти позитивний імідж компанії та власні позиції в компанії.

Коен та ін. [11] виявили, що звіти 10-K актуальні для фінансових показників фірми, таких як майбутні прибутки, прибутковість та оголошення новин, і можуть передбачити банкрутство на рівні фірми.

Регресійна модель із пояснюючими фінансовими та контрольними змінними може вимірювати їхній вплив на залежну змінну (наприклад, позитивні чи негативні новини як якісна змінна, ціна акцій як кількісна змінна тощо). Серед фінансових показників існуючі дослідження містять рентабельність активів (ROA), рентабельність капіталу (ROE), коефіцієнт Тобіна Tobin’s Q (TQ) і рентабельність інвестованого капіталу (ROIC). Контрольні змінні компанії складаються з розміру фірми, кількості активів фірми, ліквідності, фінансових потреб або дефіциту та фінансового важеля. Інвестори мають різні профілі ризику, що визначає їхню схильність до різних фінансових інструментів (ФІ) [12]. Кластери можуть об’єднувати інвесторів з однаковими перевагами, які зацікавлені в однакових бізнес-звітах [13].

Нейронні мережі використовуються для прогнозування цін на акції на основі великих даних [14].

1.2. Фінансові показники бізнес звіту 10-К

Звіт 10-К — це звіт компаній, зареєстрованих у Сполучених Штатах, акції яких торгуються на американських фондових біржах, які мають подати до Комісії з цінних паперів і бірж США (SEC). Звіт 10-К містить детальну інформацію про фінансовий стан, діяльність, управління, ризики та стратегію компанії.

Звіт 10-К містить таку інформацію:

1. Фінансові звіти: такі фінансові показники, як баланс, звіт про прибутки та збитки, звіт про рух грошових коштів і звіт про зміни у капіталі.
2. Огляд бізнесу: опис основних аспектів бізнесу, включаючи продукти, послуги, галузь, географічні ринки та основних конкурентів.
3. Фактори ризику: аналіз потенційних ризиків, які можуть негативно вплинути на ефективність компанії, включаючи конкуренцію, регуляторні ризики та технологічні зміни.
4. Обговорення та аналіз керівництва: аналіз компанією її фінансових результатів, стратегії, планів і факторів, які можуть вплинути на майбутні результати.
5. Корпоративне управління: інформація про раду директорів і виконавчих директорів компанії, а також інформацію про їхню винагороду, пакети акцій та опціони на акції.
6. Право власності на цінні папери та додаткові витрати: опис структури акціонерного капіталу, включаючи різні класи акцій та права акціонерів.
7. Судові процеси: інформація про будь-які значні судові процеси, в яких бере участь компанія.

8. Податкові питання: опис податкових зобов’язань компанії та будь-яких потенційних податкових проблем, які можуть виникнути.

9. Значні угоди та контракти: опис важливих контрактів і угод, які можуть бути суттєвими для діяльності компанії.

У звіті 10-K основні фінансові показники надають інформацію про фінансовий стан і результати діяльності компанії. Ці показники поділяються на кілька фінансових звітів, таких як баланс, звіт про фінансові результати та звіт про рух грошових коштів.

Звіт про прибутки та збитки (The Income Statement), також відомий як звіт про прибутки або звіт про результати діяльності, є підсумком доходів і витрат компанії за певний період часу, зазвичай рік або квартал. Він показує, як підприємство перетворює виручку від продажу товарів і послуг у чистий прибуток з урахуванням усіх витрат.

Розглянемо основні розділи та статті Звіту про фінансові результати:

1. Дохід (Revenue): дохід від продажу товарів або послуг. Дохід також можна назвати продажами або оборотом.

2. Собівартість проданих товарів (Cost of Goods Sold): витрати на виробництво або придбання товарів або послуг, які продає компанія. Собівартість включає витрати на матеріали, оплату праці та накладні витрати на виробництво.

3. Валовий прибуток (Gross Profit): різниця між доходом і вартістю проданих товарів. Валовий прибуток показує, скільки компанія заробляє після оплати прямих витрат на виробництво товарів або послуг.

4. Операційні витрати (Operating Expenses): витрати на управління компанією, не пов’язані з виробництвом товарів чи послуг. Операційні витрати включають заробітну плату працівникам, орендну плату, рекламу, амортизацію, дослідження та розробки та інші невиробничі витрати.

5. Операційний дохід (Operating Income): різниця між валовим прибутком і операційними витратами. Операційний прибуток показує, скільки компанія заробляє від свого основного бізнесу без урахування відсотків і податків.

6. Відсотки та інші фінансові витрати (Interest and Other Financial Expenses): витрати на відсотки за боргами та інші фінансові витрати, такі як плата за обслуговування боргів.

7. Податок на прибуток (Income Tax): сума податків, яку компанія має сплатити зі свого доходу.

8. Чистий прибуток (Net Income): відображає кінцевий прибуток після обліку всіх доходів, витрат, відсотків і податків за певний період часу, зазвичай рік або квартал. Чистий прибуток використовується для визначення успіху компанії та її здатності приносити прибуток акціонерам.

Баланс (The balance sheet) — це миттєвий знімок фінансового стану компанії на певну дату. Він містить активи (те, чим володіє компанія), зобов'язання (те, чим компанія заборгувала) і власний капітал (різниця між активами та зобов'язаннями). Складається з наступних розділів:

1. Активи (Assets): вони поділяються на поточні активи (наприклад, готівка, дебіторська заборгованість, запаси) і довгострокові активи (наприклад, обладнання, нерухомість, інтелектуальна власність).

2. Зобов'язання (Liabilities): До них належать поточні зобов'язання (наприклад, кредиторська заборгованість, короткострокова заборгованість) і довгострокові зобов'язання (наприклад, позичені кошти, пенсійні зобов'язання).

3. Власний капітал (Equity): це сума коштів, інвестованих акціонерами, і накопичений прибуток компанії.

Звіт про рух грошових коштів (Cash Flow Statement): цей звіт показує, як компанія генерує та використовує готівку протягом певного періоду часу, як

правило, за рік. Він детально описує зміни в грошових коштах та їх еквівалентах, розділених на три основні категорії:

1. Операційний грошовий потік (Operating Cash Flow): відображає чистий грошовий потік, отриманий від основної діяльності компанії, такої як продаж товарів і послуг, оплата постачальникам, зарплата співробітників і податки. Позитивний операційний грошовий потік свідчить про те, що компанія успішно конвертує свої прибутки в готівку.

2. Інвестиційний грошовий потік (Investing Cash Flow): відображає грошові потоки, пов’язані з інвестиціями в довгострокові активи, такі як купівля чи продаж обладнання, нерухомості, інтелектуальної власності, акцій інших компаній та боргових інструментів. Негативний інвестиційний грошовий потік може бути пов'язаний з інвестиціями у зростання і розвиток компанії.

3. Фінансовий грошовий потік (Financing Cash Flow): відображає грошові потоки, пов’язані з фінансуванням компанії, включаючи випуск і погашення акцій і боргів, виплату дивідендів акціонерам та інші операції, пов’язані з фінансуванням. Негативний фінансовий грошовий потік може свідчити про погашення боргів або дивідендів.

На основі даних цього звіту ми можемо розрахувати наступні важливі фінансові показники:

Валовий дохід (Gross Margin). Показує, скільки процентних пунктів доходу залишається після вирахування витрат, пов’язаних із виробництвом або продажем товарів і послуг. Чим вищий попит на акції компанії, тим вищий показник і ціна акцій. Розраховується за наступною формулою:

$$Gross\ Margin = (Revenue - Cost\ of\ Goods\ Sold) / Revenue$$

1. Операційний дохід (Operating Margin). Показує, скільки процентів доходу отримує компанія від операційної діяльності після вирахування витрат, пов’язаних з виробництвом, продажами, адміністративними витратами та податками. Чим більше прибутку на долар доходу від продажів, тим вищий показник. Розраховується за наступною формулою:

$$Operating\ Margin\ =\ Operating\ Income\ /\ Revenue$$

2. Маржа чистого прибутку (Net Profit Margin). Показує чистий прибуток компанії після вирахування всіх витрат, включаючи податки та відсотки за борг. Розраховується за наступною формулою:

$$Net\ Profit\ Margin\ =\ (Net\ Profit\ /\ Revenue)\ *\ 100\%$$

Коефіцієнт поточної ліквідності (Current Ratio). Оцінює здатність компанії виконувати свої поточні зобов'язання на основі її поточних активів.

Розраховується за наступною формулою:

$$Current\ Ratio\ =\ Current\ Assets\ /\ Current\ Liabilities$$

Співвідношення боргу до власного капіталу (Debt-to-Equity Ratio).

Порівнює загальний борг компанії з її загальним капіталом.

Розраховується за наступною формулою:

$$Debt\ to\ Equity\ Ratio\ =\ Total\ Debt\ /\ Total\ Equity$$

Рентабельність активів (Return on Assets). Вимірює, наскільки ефективно компанія використовує свої активи для отримання прибутку.

Розраховується за наступною формулою:

$$ROA\ =\ Net\ Income\ /\ Total\ Assets$$

3. Рентабельність власного капіталу (Return on Equity). Вимірює суму чистого прибутку, яку генерує компанія, як відсоток від загальної суми власного капіталу, інвестованого акціонерами. Розраховується за наступною формулою:

$$ROE\ =\ Net\ Income\ /\ Shareholders'\ Equity$$

4. Прибуток на акцію (Earnings per Share). Вимірює суму прибутку, яку компанія генерує на 1 акцію компанії. Розраховується за наступною формулою:

$$EPS = Net\ Income / Outstanding\ Shares$$

5. Співвідношення ціна-прибуток (Price-to-Earnings Ratio). Порівнює ціну акцій компанії з прибутком на акцію. Розраховується за наступною формулою:

$$P/E\ Ratio = Market\ Price\ per\ Share / Earnings\ per\ Share$$

6. Ціна/балансова вартість (Price-to-Book Ratio). Порівнює ринкову ціну компанії за акцію з її балансовою вартістю за акцію. Розраховується за наступною формулою:

$$P/B\ Ratio = Market\ Price\ per\ Share / Book\ Value\ per\ Share$$

Всі зазначені коефіцієнти є ключовими фінансовими показниками, які використовуються інвесторами, аналітиками та менеджерами для оцінки фінансового стану та результативності компанії. Вони надають важливу інформацію про рентабельність, ефективність управління активами та ризики інвестицій.

РОЗДІЛ 2
ВИКОРИСТАННЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ У ФІНАНСОВІЙ ГАЛУЗІ

2.1. Огляд застосування інтелектуального аналізу даних у фінансовій галузі

У фінансовій галузі активно поширюється використання інтелектуального аналізу даних (Data Mining). Відповідно до історії їх імплементації ці технології у фінансовій сфері сталі використовувати раніше, ніж додатки для мобільних банківських комп'ютерів.

За результатами опитування, проведеного на засіданні Bloomberg у 2017 році в Нью-Йорку, лише 16% фірм включили технології машинного навчання у свої інвестиційні стратегії. Тим часом решта досліджує способи як зробити це (24%), хотіли б дізнатися, як це зробити (26%), або навіть не міркували, як впровадити ці технології (32%). Керівник Bloomberg з машинного навчання Гері Казантев вважає, що незабаром машинне навчання прийде до кожної фірми.

Рис. 2.1. Основні фінансові сфери використання машинного навчання

Зараз в управлінні портфелем автоматизований фінансовий консультант або “Робо-едвайзер” є звичайним явищем у фінансовому ландшафті, але до 2008 року такого терміну навіть не існувало. Цей термін вводить в оману і взагалі не передбачає використання роботів. “Робо-едвайзери” таких компаній як Betterment, Wealthfront тощо насамперед є алгоритмами, побудованими для формування фінансового портфеля відповідно до цілей та ставлення інвесторів до ризику. Користувачі вводять свої цілі у відповідні поля робо-едвайзора (наприклад, виходять на пенсію у віці 60 років із заощадженнями в розмірі 1 млн. грн.), віку, доходу та поточних фінансових активів. Радник (якого доцільніше називати «розподільником») розміщує інвестиції між класами активів та фінансовими інструментами, щоб досягти цілей користувача. Система потім переглядає відповідно до змін цілей користувача та змін на ринку доходності фінансових інструментів у режимі реального часу, намагаючись завжди знаходити те, що найкраще відповідає початковим цілям користувача. Роботи консультанти стали затребуваними для інвесторів, які не потребують фізичного радника, щоб відчувати себе комфортно в процесі інвестування, а також менш спроможні здійснювати значно вищу оплату людським радникам.

Також роботи консультанти активно використовуються для прийняття швидких рішень на ринку, що дозволяє отримувати прибуток на швидких угодах. Такий тип торгівлі отримав назву “Алгоритмічна торгівля”.

З витоками, що йдуть від 1970-х років, алгоритмічна торгівля або "Автоматизовані торгові системи", передбачає використання складних систем зі штучним інтелектом для прийняття надзвичайно швидких торгових рішень. Алгоритмічні системи часто роблять тисячі або мільйони торгів упродовж дня, тому ще вживається термін "високочастотна торгівля" (англійською High-frequency trading, HFT), який вважається підмножиною алгоритмічної торгівлі. Більшість хедж-фондів та фінансових установ публічно не розкривають підходи до обміну інформацією при торгівлі, але вважається, що видобуток даних і

глибоке навчання (Deep Learning) грають все більш важливу роль у прийнятті торгових рішень в реальному часі. Біржі інколи встановлюються обмеження на ексклюзивне використання машинного навчання в торговельних акціях і товарах.

Також машине навчання активно використовується у виявленні шахрайства. Кожен рік збільшується кількість цінних даних компанії, що зберігаються в Інтернеті, і це створює "ідеальний шторм" для ризику безпеки даних. Хоча попередні системи виявлення фінансових шахрайств значною мірою залежать від складних і надійних правил, сучасне виявлення шахрайства виходить за рамки переліку факторів ризику - штучний інтелект активно вивчає та відбирає нові потенційні (або реальні) загрози безпеці. Ті самі принципи справедливі й для інших проблем безпеки даних. Використовуючи машинне навчання, системи можуть виявити унікальні види діяльності або поведінку ("аномалії") та позначати їх для перевірки групою безпеки. Виклик для цих систем полягає в тому, щоб уникнути помилкових позицій - ситуацій, коли "ризиками" позначаються явища, які ніколи не були пов'язані з ризиком.

Наступною сферою застосування МН у фінансах є страхування, а саме таких процесів як андеррайтинг. Андеррайтинг - це процес, за допомогою якого страхова компанія визначає, чи приймати пропозицію страхувальника щодо укладання договору страхування, і якщо приймати, то на яких умовах.

Цей процес може бути описаний як ідеальна робота для машинного навчання в галузі фінансів, і, дійсно, в галузі є велике занепокоєння, оскільки машини можуть заміняти велику частку позицій андеррайтингу, тому, що за допомогою МН можна замінити рутинну роботу людей, які працюють сьогодні. Особливо у великих компаніях (великі банки або страхові компанії) за допомогою використання алгоритмів машинного навчання можна аналізувати мільйони прикладів споживчих даних (вік, робота, сімейний стан тощо) та результатів фінансового кредитування чи страхування (раніше це здійснювала

людина) щодо погашення кредиту у встановлений термін, ймовірність потрапити в автомобільну аварію тощо.

Основні тенденції, які можна оцінити за допомогою алгоритмів, постійно аналізуються, щоб виявляти тренди, які можуть вплинути на кредитування та страхування в майбутньому. Ці результати приносять величезний дохід компаніям, які мають ресурси для найму спеціалістів у сфері Big Data та величезних обсягів минулих і поточних даних для навчання своїх алгоритмів.

Перспективними галузями фінансової сфери, де може використовуватися Data Mining, є:

1. Обслуговування клієнтів
2. Аналіз новин та почуттів людини
3. Продажі / рекомендації фінансових продуктів

Чат-боти та розмовні інтерфейси - це стрімко зростаюча галузь венчурного інвестування та обслуговування бюджету клієнтів (за даними Techemergence у 2016 ця галузь була оцінена як є найперспективніша). Такі компанії, як Kasisto, вже створюють спеціальні боти-вказівники для фінансування, щоб допомогти клієнтам поставити питання через чат, серед яких: як "Скільки я витратив на товари поточного споживання минулого місяця" та "Який баланс особистого заощаджуваного рахунку був 60 днів тому?". Ці помічники мають бути побудовані завдяки надійним двигунам для обробки природної мови, а також фінансовим взаємовідносинам споживачів. Банки та фінансові установи, які дозволяють такі швидкі запити та взаємодію, можуть перехопити клієнтів з банків, які пропонують клієнтам користувалися традиційним онлайновим банківським порталом і виконувати все самостійно. Цей досвід спілкування в чаті (або в майбутньому - голос) не є сьогодні нормою в банківській справі або фінансуванні, але може стати життєздатним варіантом для мільйонів банківських клієнтів упродовж найближчих 5 років. Ця програма виходить за межі навчання в галузі фінансів, і, ймовірно, з’явиться як спеціалізовані чат-боти у різних галузях.

Передбачається, що більша частина майбутніх застосувань машинного навчання полягає в розумінні соціальних медіа, тенденцій в новинах та інших джерел даних, а не лише цін на акції та облігації. Фондовий ринок рухається під впливом множини факторів, пов'язаних з людьми. Очікується, що МН зможе відтворити і посилити людську "інтуїцію" фінансової діяльності, виявляючи нові тенденції та сигнали.

На сьогодні існують програми продажу автоматизованих фінансових продуктів, деякі з яких можуть не містити машинні навчання (а, скоріше, інші системи, керовані правилами). Робо-консультант може запропонувати зміну інвестиційного портфелю (ребалансування), оскільки існує множина сайтів із рекомендаціями щодо страхування, тому тут може використовувати деякий ступінь штучного інтелекту, щоб запропонувати конкретний план страхування автомобіля або будинку. У майбутньому все більш персоналізовані додатки та особисті помічники можуть бути сприйняті як більш надійні та об'єктивні, ніж особисті людські радники. Так само як Amazon і Netflix можуть рекомендувати книги та фільми краще за будь-якого живого "експерта", поточні розмови з фінансовими персональними помічниками можуть зробити те ж саме для фінансових продуктів. Як видно, ці процеси починаються і в галузі страхування.

Для того, щоб розуміти як використовуються цей інструмент аналізу даних, треба розібрати що таке інтелектуальний аналіз даних, і його джерела.

2.2. Суть інтелектуального аналізу даних

Термін Data Mining перекладається як «інтелектуальний аналіз даних» також можна перекласти як «видобуток даних». Також існує два терміни які використовуються поряд із зазначеним терміном – це виявлення знань (knowledge discovery) та сховище даних (data warehouse). Зазначені терміни є важливою частиною Data Mining, і пов'язані з новим етапом розвитку засобів і

методів обробки даних. Метою Data Mining є виявлення корисних даних та закономірностей у великих та неструктурованих даних.

Велику кількість даних мозок людини неспроможний сприйняти, та опрацювати, а ще побачити важливі закономірності (взаємозв'язки) в цій кількості даних. Мозок більшої частини людства, за дуже рідкісними винятками, не здатен визначити більше ніж 3 взаємозв'язки навіть у не дуже великих вибірках. Навіть традиційна статистика, що давно претендує на роль як основний інструмент аналізу даних, нерідко дає збої при вирішенні реальних завдань з реальними даними. Статистика оперує характеристиками вибірки, котрі нерідко є фальсифікованими величинами (середня платоспроможність клієнта, в залежності від функцій ризику або втрат треба спрогнозувати можливості й наміри клієнта та інше).

Тому корисні методи математичної статистики надаються для перевірки заздалегідь вже наявної гіпотези, а не для виявлення нової. За допомогою МН перепрацьовують інформацію, щоб автоматизувати пошук та виявлення патернів (шаблонів), характерних для будь-яких неоднорідних багатовимірних даних.

В інтелектуальному аналізу даних тягар виявлення незвичайних шаблонів та формування гіпотез тепер на обчислювальних потужностях комп'ютера, а не на людині, на відмінну від статистики. Інтелектуальний аналіз даних (data mining) - це сукупність великої кількості різних методів виявлення знань. Від типу наявних даних та очікуваного результату часто залежить вибір методу обробки. Серед цих методів: кластеризація, асоціація (об'єднання), аналіз часових рядів, класифікація і прогнозування та ін.

Розглянемо властивості виявлених знань, які видобуваються з даних:

1. Отриманні знання мають містити новизною, тобто які раніше були невідомі, оскільки зусилля, витрачені на отримання даних, не окупаються, якщо дані вже відомі.

2. Нові знання мають бути неочевидні. Результати аналізу мають відображати нетривіальні, несподівані закономірності в даних, тобто приховані знання, які були раніше неочевидні. Щоб ці дані неможна було отримати легшим способом, не використовуючи методи МН, наприклад візуальною перевіркою, або за допомогою стандартних методів математичної статистики.

3. Отримані знання мають приносити практичну користь у використанні. Отримана інформація повинна бути інформативною і вірогідним джерелом при подальшому аналізі нових даних з високим рівнем достовірності.

4. Знання мають бути зрозумілі людині і мають бути логічно та доступно пояснені, оскільки що існує ймовірність, що нові дані є випадковими. Нові отримані дані мають бути отримані у зрозумілому для людини форматі.

Для того, щоб представити отримані знання, в МН застосовуються моделі. Моделі поділяються на види та залежать від методів, за допомогою яких були отримані ці дані. Найпоширенішими є: кластери, правила, математичні функції й дерева рішень.

Сфера застосування видобутку даних не обмежена - видобуток даних має попит в усіх галузях, де є велика кількість даних, яку не може обробити людина.

Згідно з звітом компанії Accenture та Frontier Economics, використання штучного інтелекту (AI) та видобутку даних може збільшити продуктивність у промисловому секторі на до 40% [15]. Крім того, видобутку даних може допомогти підприємствам виявляти нові ринки, збільшувати продажі, покращувати взаємодію з клієнтами та забезпечувати більш точне прогнозування попиту [15]. Згідно зі звітом McKinsey, організації, "які використовують механізми зростання обсягів продажів B2B на основі даних,

повідомляють про зростання, що перевищує ринок, і збільшення EBITDA в діапазоні від 15% до 25%. [16]

Керівники компаній усвідомлюють, що за допомогою використання методів видобутку даних у ході конкуренції вони можуть отримувати відчутні переваги.

2.3. Класифікація завдань інтелектуального аналізу даних

Більшість завдань, з якими доводиться стикатись аналітику, можна вирішити за допомогою інтелектуального аналізу даних. З них основними є: кластеризація, регресія, пошук асоціативних правил та класифікація. Нижче описано детальніше кожного завдання які вирішує інтелектуальний аналіз.

1. Завдання кластеризації зводиться к пошуку незалежних груп і їх характеристик у всій множині аналізованих даних. Об’єднання груп однорідних об'єктів дає змогу скоротити їх число, а тому поліпшує аналіз кластеризованих даних.

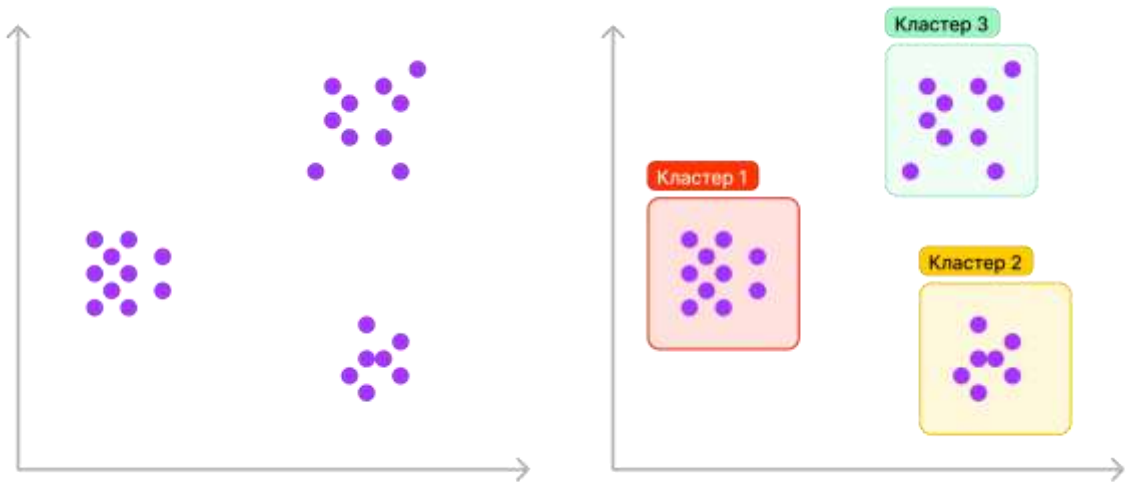

Рис. 2.3.1. Відображено 3 кластери, сформовані з набору даних

2. Завдання регресії дозволяє отримати значення деякого параметра за допомогою існуючих характеристик (предикторів) об'єкта. На відміну від

завдання класифікації, значеннями параметра є не кінцева множина класів, а множина дійсних чисел.

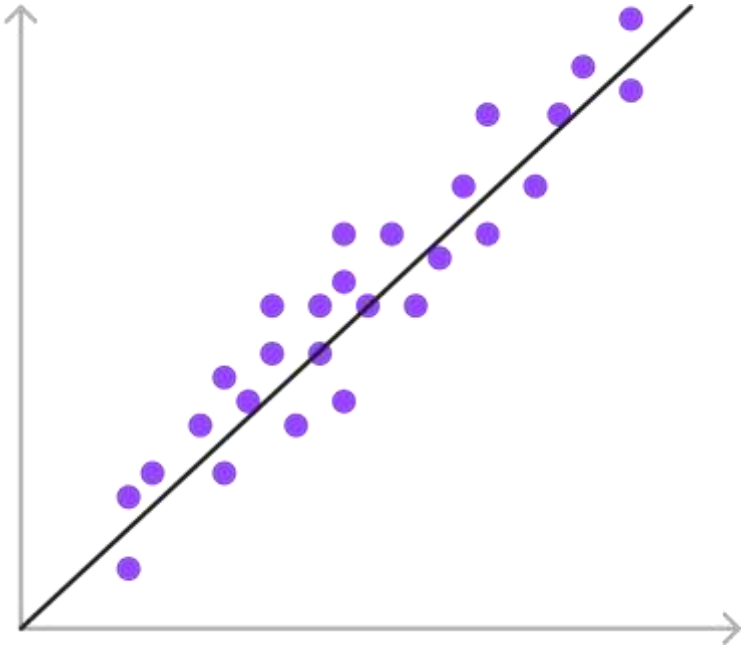

Рис. 2.3.2. Графік лінійної залежності даних

3. Завдання асоціації, основною метою якого є пошук періодичних асоціацій (або залежностей) між сутностями та подіями. Залежності, які були знайдені, представляють у вигляді правил для кращого усвідомлення даних, які були проаналізовані, а також для передбачення різних явищ або подій.
4. Завдання класифікації полягає у визначенні класу об'єкта за даними характеристиками. В цьому завданні відома множина класів, класом якого може бути даний об'єкт.

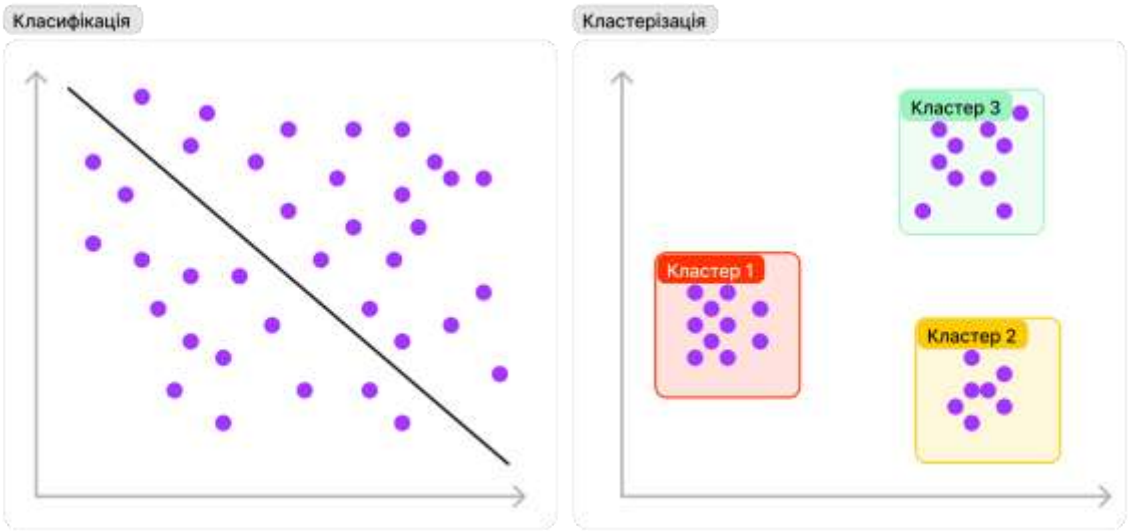

Рис. 2.3.3. Порівняння класифікації і кластеризації

5. Послідовні шаблони - встановлення залежностей між пов'язаними в часі подіями, тобто виявлення залежностей, що якщо відбувається подія X, то у майбутньому через деякий час повинна відбутися подія Y.Аналіз відхилення - виявлення непередбачуваних або відмінних від звичних шаблонів від звичайних шаблонів чи поведінки даних. Основна мета аналізу відхилень полягає в виявленні аномалій, або викидів, що можуть вказувати на важливі події, помилки або аномалії у процесі, які потребують уваги або подальшого дослідження.

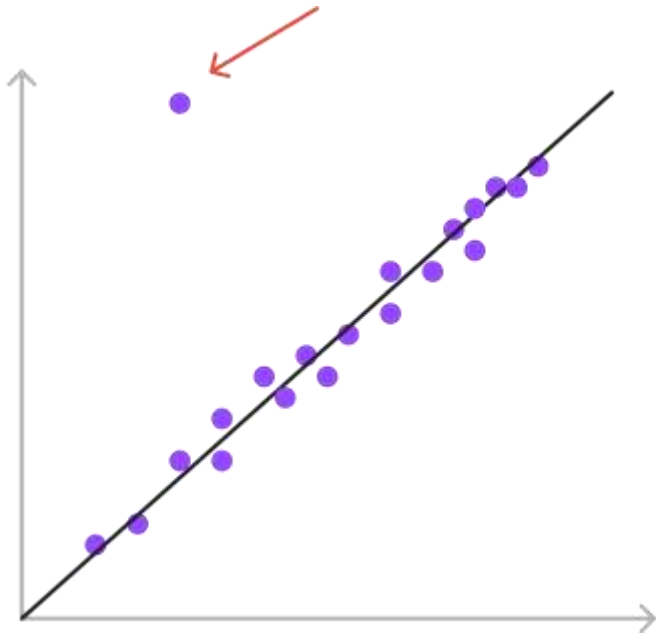

Рис. 2.3.4. Приклад відхилення даних

Описані завдання можуть бути поділенні за призначенням на прогнозні та описові, пов’язані з інтерпретацією результатів.

Прогнозування даних включає декілька етапів:

1. За допомогою набору даних та відомих результатів будується модель на основі навчальних даних.

2. Проводиться тестування моделі на основі тестових даних.

3. Відбувається передбачення нових результатів з використанням нових наборів даних.

Потрібно щоб створені моделі працювали з високим рівнем точності (зокрема, мінімальною середньою стандартною похибкою). Цей вид завдань охоплює побудову регресії та класифікацію. Також сюди можуть бути віднесені пошук асоціативних правил, коли результати розв’язку цього завдання будуть використовуватися для передбачення появи певних подій.

Описові завдання більш спрямовані на поліпшення розуміння даних для аналізу. Ключовими моментами в цих моделях є легкість в інтерпретації отриманих результатів для людини. Скоріше за все, виявлені залежності будуть мати специфічні риси конкретних даних, які досліджували, і будуть

зустрічатися лише у цих даних, але ці дані мають корисність і тому мають бути відомі. Такими завданнями є пошук асоціативних правил та кластеризація.

Завдання також поділяються за способами вирішення на навчання без вчителя (англ. unsupervised learning) та навчання з учителем (англ. supervised learning). Ця назва походить від терміну машинне навчання (англ. machine learning), яке найчастіше використовується в англомовній літературі й означає всі технології машинного навчання.

Для навчання з викладачем завдання зводиться до двох етапів. Перший – це класифікатор («розмітка даних»), тобто за допомогою алгоритму видобутку даних створюється модель аналізованих даних. Другий етап містить навчання цього класифікатора. Тобто перевіряється, наскільки правильно він працює, і якщо він працює некоректно, то здійснюють додаткове навчання створеного класифікатора. Така процедура тримає, доки класифікатор не досягне необхідного рівня якості, або з'явиться розуміння, що обраний алгоритм не відповідає цим даним або працює не коректно, або дані мають нечітку структуру, яку можна було б виявити. Завданнями, які належать до цього типу, є регресія та класифікація.

Навчання без вчителя об’єднує завдання, які визначають описові моделі, як, наприклад залежності у покупках покупців великих магазинів. Якщо ці закономірності існують, модель має представити їх у зрозумілому для інтерпретації вигляді. Перевагою є те, що вони надають можливість отримати результат без попередніх знань про дані для аналізу. До цих завдань відносять пошук асоціативних правил та кластеризацію.

2.3.1. Завдання регресії та класифікації

Під час проведення аналізу найчастіше необхідно встановити, до якого з вже відомих типів або категорій належать об'єкти, що досліджуються, тобто провести їх класифікацію. Наприклад, коли особа звертається до банку за отриманням кредиту, спеціальний алгоритм або працівник банку має прийняти

рішення: чи може цей потенційний клієнт отримати кредит. Таке рішення базується на даних про досліджуваний об'єкт (у цьому випадку - особу): її місце роботи, рівень доходів, вік, сімейний стан, кредитну історію тощо. Після аналізу цих даних алгоритм або працівник банку визначає, до якого з двох відомих класів належить особа - "кредитоспроможний" або "некредитоспроможний".

В загальному випадку кількість класів може бути понад два. Наприклад, у завданні з розпізнавання образів кількість можливих класів дорівнює 10 (відповідно до кількості цифр у десятковій системі числення). У цьому завданні об'єктом класифікації є матриця, що складається з пікселів, при цьому кожен піксель має свій власний колір, який є характеристикою аналізованого об'єкта.

У сфері Data Mining завдання класифікації сприймається як процес визначення значення одного з параметрів аналізованого об'єкта на основі значень інших параметрів. Цей параметр зазвичай називається залежною змінною, а параметри, які впливають на його значення, - незалежними змінними. У вказаних прикладах незалежними змінними були:

- Кредитоспроможність клієнта, яка може мати значення "так" або "ні".
- Тип повідомлення, який може бути класифікований як "spam" або "mail".
- Цифра образу, яка може приймати значення від 0 до 9.

Так, у всіх наведених прикладах незалежна змінна має обмежену множину значень, які можуть бути представлені як {так, ні}, {spam, mail}, {0, 1, ..., 9}. Завдання, в яких незалежна змінна приймає дійсні числові значення, називають завданнями регресії. Прикладом задачі регресії може бути визначення суми кредиту, яку банк може видати клієнту на підставі деяких незалежних змінних, таких як дохід, кредитна історія тощо.

Завдання регресії і класифікації вирішується в декілька кроків. На першому виділяється навчальна вибірка. До неї входять об'єкти, для яких відомі

значення як незалежних, так і залежних змінних. В описаних раніше прикладах такими навчальними вибірками можуть бути:

- інформація про клієнтів, яким раніше видавалися кредити на різні суми, і інформація про їх погашення;
- повідомлення, що класифіковані вручну як спам або як лист;
- розпізнані раніше матриці образів цифр.

Будуючи модель для визначення значення залежної змінної на основі навчальної вибірки, часто використовують терміни "функція класифікації" або "функція регресії". Для створення найбільш точної моделі, яка відповідає навчальній вибірці, визначаються наступні ключові вимоги:

- Щоб забезпечити точність моделі класифікації або регресії, кількість об'єктів у вибірці повинна бути значною. Чим більше об'єктів враховано, тим більш точною буде функція класифікації або регресії, побудована на їх основі.
- У вибірку необхідно включити об'єкти, які охоплюють всі можливі класи у випадку задачі класифікації, або представляють всю область значень, якщо мова йде про задачу регресії.
- Для кожного класу у задачах класифікації або для кожного інтервалу у просторі значень у завданнях регресії вибірка повинна містити достатню кількість об'єктів.

На другому етапі отриману модель застосовують до об'єктів, для яких значення залежної змінної не відомо. Задачі регресії та класифікації можуть мати геометричне тлумачення [17].

Основні труднощі, з якими стикаються при розв'язанні задач класифікації та регресії, полягають в незадовільній якості вихідних даних. Це включає помилкові дані, пропущені значення, різні типи атрибутів (числові й категоричні), а також нерівну значущість атрибутів. Додатково, виникають проблеми перенавчання (overfitting) і недонавчання (underfitting). Перенавчання відбувається, коли класифікаційна функція занадто точно підлаштовується під

навчальну вибірку, намагаючись інтерпретувати помилки і аномальні значення як частину внутрішньої структури даних. Під недонавчанням розуміють ситуацію, коли на навчальній вибірці виявляється велика кількість помилок, що вказує на відсутність чітких закономірностей або потребу в іншому методі їх виявлення [18].

2.3.2. Завдання пошуку асоціативних правил

Пошук асоціативних правил представляє собою один із найпопулярніших застосунків Data Mining. Сутність цього завдання полягає у виявленні часто зустрічаючихся наборів об'єктів. Ця задача є окремим випадком задачі класифікації і спочатку вирішувалася під час аналізу поведінки покупців у супермаркетах. Дані про покупки, які покупці складають у свої кошики або візки, піддавалися аналізу. Під час цього аналізу важливою є інформація про те, які товари часто купуються разом, в якій послідовності, які категорії споживачів, які товари популярніше, а також інформація про періоди часу. Ця аналітика дозволяє ефективніше планувати закупівлю товарів, проведення рекламних кампаній тощо.

Наприклад, з аналізу набору покупок у магазині можна виявити такі часті комбінації товарів:

- {Чипси, пиво};
- {Вода, горіхи}.

Так, на підставі знань про часті комбінації покупок можна зробити висновок, що наявність чипсів часто співвідноситься з пивом, а горіхів - з водою. Використовуючи ці знання, можна розмістити ці товари поруч, об'єднати їх у спільний пакет зі знижкою, стимулюючи покупця здійснити придбання [19].

Задача виявлення асоціативних правил актуальна не лише в сфері торгівлі. Наприклад, у сфері обслуговування інтерес представляє, якими послугами клієнти прагнуть користуватися в сукупності.

Для отримання цієї інформації проводиться аналіз даних щодо послуг, якими користується клієнт протягом певного періоду часу, такого як місяці чи роки. Це дозволяє визначити оптимальні пакети послуг, що можуть бути надані клієнту з метою максимізації його користі.

У медицині аналізуються симптоми та хвороби, які спостерігаються у пацієнтів. Знання про найбільш типові зв'язки між хворобами та симптомами допомагає лікарям правильно ставити діагноз та призначати відповідне лікування.

Точно, в аналізі даних велике значення має вивчення послідовності подій, що відбуваються. Виявлення закономірностей в цих послідовностях дозволяє з певною ймовірністю прогнозувати майбутні події, що забезпечує більш точне та обґрунтоване прийняття рішень. Це завдання є важливим різновидом аналізу асоціативних правил і відоме як послідовний аналіз.

Так, у послідовному аналізі основною відмінністю є встановлення відношень порядку між досліджуваними наборами. Це відношення може бути визначено різними способами. При аналізі послідовності подій в часі, об'єктами таких наборів є події, а відношення порядку відповідає хронології їх появи.

Послідовний аналіз, або сіквенційний аналіз, знаходить широке застосування, включаючи телекомунікаційні компанії. В цьому випадку він використовується для аналізу даних щодо аварій на різних вузлах мережі. Аналіз послідовності аварій може допомогти виявити несправності та запобігти новим аваріям.

Так, знаючи це, можна розробляти профілактичні заходи, спрямовані на усунення причин виникнення збоїв. У разі наявності інформації про інтервали часу між збоями можна передбачити не лише факт їх появи, а й приблизний час виникнення, що часто є не менш важливим.

2.3.3. Завдання кластеризації

Завдання кластеризації передбачає розподіл досліджуваної множини об'єктів на групи "схожих" елементів, які називаються кластерами. Сам термін "кластер" англійського походження (cluster) означає згусток, групу або пучок. У літературі також використовуються споріднені поняття, такі як клас, таксон або згущення. Часто розв'язання задачі розбиття множини елементів на кластери називають кластерним аналізом.

Так, кластеризація знаходить застосування практично у будь-якій області, де потрібне дослідження експериментальних або статистичних даних. Давайте розглянемо приклад з області маркетингу, де ця задача відома як сегментація.

Концепція сегментації базується на припущенні, що всі споживачі відрізняються один від одного. Вони мають різні потреби та вимоги до товарів, а також виявляють відмінне споживання під час вибору, придбання, використання та реакції на товар. Це стимулює різноманітні підходи до роботи зі споживачами, включаючи пропозицію різноманітних товарів, їхнє різноманітне просування й продаж. Сегментація споживачів вимагає визначення унікальних відмінностей між ними та розуміння, як ці відмінності впливають на вимоги до товарів.

У маркетингу для сегментації використовуються критерії (характеристики), такі як географічне розташування, соціально-демографічні ознаки, мотиви здійснення покупки та інші.

На основі результатів сегментації маркетолог може визначити характеристики сегментів ринку, такі як фактична й потенційна величина сегмента, групи споживачів, чиї потреби не повністю задовольняються жодним виробником, що працює на даному сегменті ринку. За цими параметрами маркетолог може зробити висновок щодо привабливості для компанії роботи в кожному з виокремлених сегментів ринку [20].

Дослідження результатів кластеризації, зокрема виявлення причин, що об'єднують об'єкти в групи, може відкривати нові перспективні напрямки у

наукових дослідженнях. Традиційним прикладом такого дослідження є створення періодичної таблиці елементів. У 1869 році Дмитро Менделєєв класифікував 60 відомих на той час хімічних елементів на кластери або періоди. Елементи, які потрапили в одну групу, мали схожі характеристики. Вивчення причин, що призвели до розбиття елементів на кластери, істотно вплинуло на пріоритетні напрямки наукових досліджень на наступні десятиліття. Проте, лише через 50 років квантова фізика надає переконливі пояснення періодичній системі.

Кластеризація відрізняється від класифікації тим, що для проведення аналізу не потрібно мати виділену залежну змінну. З цієї точки зору вона відноситься до безконтрольного навчання (unsupervised learning). Це завдання зазвичай вирішується на початкових етапах дослідження, коли про дані відомо небагато інформації. Вирішення цього завдання допомагає краще зрозуміти дані, тому з цієї точки зору завдання кластеризації є описовим.

Для завдання кластеризації характерна відсутність будь-яких попередніх відомостей про відмінності між змінними або об'єктами. Натомість, фокус здійснюється на пошуку груп найбільш близьких, схожих об'єктів. Методи автоматичного розбиття на кластери зазвичай використовуються спільно з іншими методами Data Mining, щоб дослідити, що конкретно підтримує таке розбиття та які ознаки воно відображає.

Кластерний аналіз дозволяє розглядати досить великий обсяг інформації й різко скорочувати, стискати великі масиви інформації, робити їх компактними й наочними (Рис. 2.3.1.). Відзначимо ряд особливостей, властивих задачі кластеризації.

По-перше, рішення, пов'язані з кластеризацією, суттєво залежать від природи об'єктів даних і їхніх атрибутів. З одного боку, це можуть бути об'єкти, що мають однозначно визначені та чітко вимірювані атрибути, а з іншого - об'єкти, для яких характеристики мають невизначений розподіл або нечіткий опис.

По-друге, рішення також значно залежить від уявлення про кластери та передбачувані відносини між об'єктами даних та кластерами. Тому важливо враховувати такі властивості, як можливість або неможливість належності об'єктів до декількох кластерів. Необхідне також чітке визначення поняття приналежності до кластера: однозначна (належить / не належить), ймовірнісна (ймовірність приналежності) або нечітка (ступінь приналежності) [21].

РОЗДІЛ 3
РОЗРОБКА МОДЕЛІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ БІЗНЕС ЗВІТІВ

3.1. Побудова моделі

Ми отримали історичні дані за останні 20 років (1997-2022) для Amazon Inc. [22] за допомогою зовнішнього сервісу Financial Modeling [23], який зберігає та надає річні та квартальні звіти компаній через API (рис. 3.1).

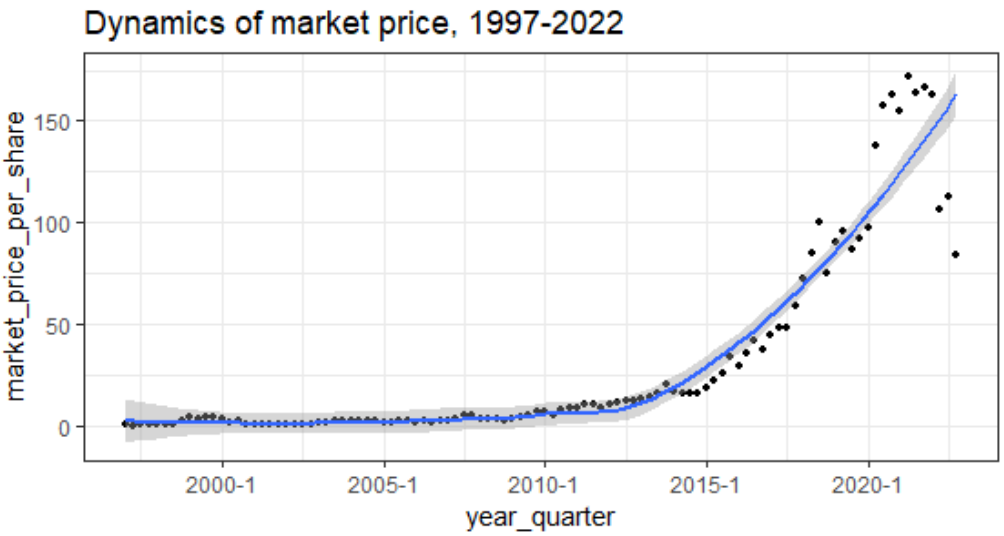

Рис. 3.1. Динаміка ринкової ціни за акцію Amazon, Inc., 1997-2022 рр.

З цих даних ми вилучили такі важливі показники, як дохід, собівартість доходу, операційний прибуток, чистий прибуток, поточні активи, поточні зобов'язання, загальний борг, загальний капітал, акціонерний капітал, акції в обігу, ринкова ціна на акцію, прибуток на акцію та Балансова вартість на акцію.

Використовуючи мову програмування Python, ми обробили ці дані та створили квартальний набір даних (Таблиця 3.1). Ми аналізуємо цей набір даних і визначаємо зв'язки між ціною акцій і фінансовими показниками, отриманими з квартального звіту, використовуючи алгоритм випадкового лісу і алгоритми множинної лінійної регресії.

Таблиця 3.1. Зведена статистика

Variable						
market_price_per_share,$	0	1.985	6	30	36.	1
Capitalization, bln $	0	16.36	5	29	34	1
revenue, bln $	0	1324	6	26	33.	1
cost_of_revenue, bln $	0	0.996	5	20	25.	1
operating_income, bln $	-	0.011	0	0.	0.7	8
net_income, bln $	-	-0.00	0	0.	0.3	1
current_assets, bln $	0	1.497	8	28	35.	1
current_liabilities, bln $	0	0.920	5	26	33.	1
total_debt, bln $	0	1.258	2	17	8.2	1
total_equity, bln $	-	0.035	5	20	16.	1
shareholders_equity,bln $	-	0.035	5	20	16.	1
outstanding_shares bln $	0	8.269	9	8.	9.6	1
earning_per_share	-	-0.00	0	0.	0.0	1
book_value_per_share	-	0.011	0	2.	1.7	1

Ми обрали два алгоритми, щоб порівняти їх ефективність, прогнозуючи ціни на акції на основі фінансових показників, отриманих із бізнес-звітів. Random Forest використовує кілька дерев рішень, щоб робити точні прогнози, запобігаючи переобладнанню шляхом усереднення результатів. Множинна лінійна регресія аналізує вплив різних змінних на ціни акцій. Порівняння цих алгоритмів допоможе нам оцінити внесок кожної змінної для пояснення зміни курсу акцій (рис. 3.2).

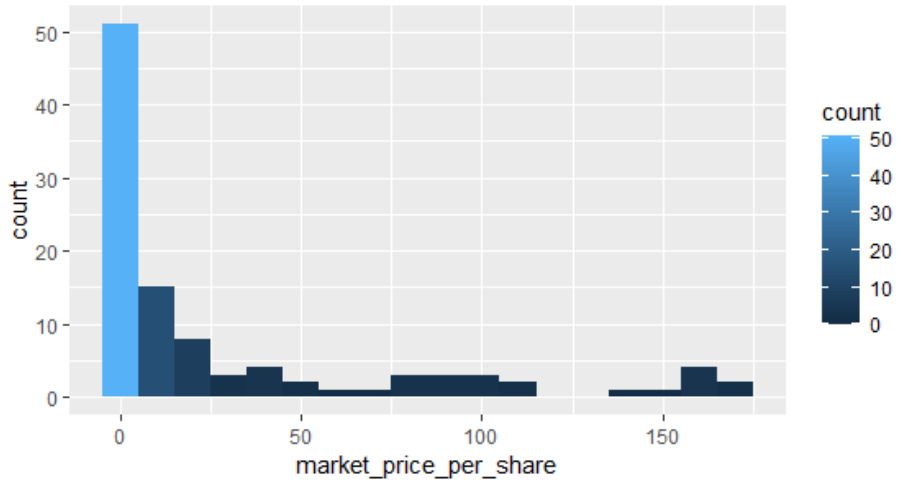

Рис. 3.2. Гістограма ринкової ціни за акцію Amazon, Inc., 1997-2022 рр.

Після аналізу кореляції ми вилучили корельовані пояснюючі змінні: earning_per_share, total_debt, total_assets, net_profit_margin, return_on_assets,

price_to_book_ratio. Ми також видалили три додаткові змінні: debt_to_equity_ratio, current_ratio та return_on_equity. Ми помітили, що після видалення цих змінних відсоток відхилення для нашої моделі значно покращився. Покращений відсоток відхилення показав, що прогнози моделі були ближчими до фактичних цін на акції. Тепер у нас є лише 5 незалежних змінних: net_income, price_to_earning_ratio, total_equity, operating_margin і gross_margin. Щоб побудувати моделі множинної регресії та моделі випадкового лісу, ми використали бібліотеку sklearn [24] і pandas для створення набору даних на основі попередньо оброблених даних (рис. 3.3).

	net_income	...	price_to_book_ratio
net_income	1.000000	...	0.045761
total_equity	0.600405	...	0.069058
total_debt	0.590331	...	0.060904
total_assets	0.608594	...	0.071829
gross_margin	0.122052	...	0.056929
operating_margin	0.245492	...	0.069011
net_profit_margin	0.294221	...	0.075584
current_ratio	0.172363	...	0.091200
debt_to_equity_ratio	0.039714	...	0.995438
return_on_assets	0.245485	...	0.047480
return_on_equity	0.056283	...	0.124220
earning_per_share	0.962601	...	0.040639
price_to_earning_ratio	0.019554	...	0.008598
price_to_book_ratio	0.045761	...	1.000000

Рис. 3.3. Кореляційний аналіз для пояснювальних змінних

Ми розділили вихідний набір даних на навчальні та тестові підмножини, щоб навчити модель і оцінити її ефективність. Ми дотримувалися загальноприйнятого підходу, коли дані розділялися у співвідношенні 80:20, де 80% даних використовувалися для навчання моделі, а решта 20% використовувалися для оцінки її ефективності. Це допомогло нам оцінити точність і надійність моделі на основі нових даних і запобігти переобладнанню. Розбиття набору даних на навчальні та тестові підмножини було виконано за допомогою функції “train_test_split” у бібліотеці sci-kit-learn. Код показано на рис. 3.4.

```
df = pd.read_csv('data_AMZN/AMZN_v4.csv')
columns = ['price_to_earning_ratio', 'total_equity', 'gross_margin',
           'operating_margin', 'net_income']
x = df[columns]
y = df['market_price_per_share']
x_train, x_test, y_train, y_test = train_test_split(
    x, y, test_size=0.2, random_state=42
)
linear_model = LinearRegression().fit(x_train, y_train)
forest_model = RandomForestRegressor(random_state=42).fit(x_train, y_train)
linear_predictions = linear_model.predict(x_test)
forest_predictions = forest_model.predict(x_test)

def calculate_deviation_percentage(actual, predicted):
    deviation = predicted - actual
    return (deviation / actual) * 100

def calculate_metrics(actual, predicted):
    rmse = mean_squared_error(actual, predicted, squared=False)
    r_squared = r2_score(actual, predicted)
    return rmse, r_squared

results = pd.DataFrame(columns=['Algorithm', 'RMSE', 'Deviation Percentage', 'R-squared'])
algorithms = {
    'Linear Regression': linear_predictions,
    'Random Forest': forest_predictions,
}
results = []
for algorithm, predictions in algorithms.items():
    rmse, r_squared = calculate_metrics(y_test, predictions)
    deviation_percentage = calculate_deviation_percentage(y_test, predictions)
    average_deviation_percentage = sum(deviation_percentage) / len(deviation_percentage)
    results.append(
        [algorithm, rmse, f"{average_deviation_percentage:.2f}%", f"{r_squared *
100:)2f}%"]
print(tabulate(results,
               headers=["Algorithm", "RMSE", "Deviation Percentage", "R-squared"],
               tablefmt='psql', numalign='center'))
```

Рис. 3.4. Оцінка вибраних алгоритмів

3.2. Аналіз отриманих результатів

Для порівняння наших моделей ми використовували RMSE, відсоток відхилення та коефіцієнт детермінації R-квадрат. RMSE вимірює точність, відсоток відхилення вимірює відносну похибку, а R-квадрат вимірює, наскільки добре модель відповідає даним. Разом ці показники забезпечують комплексну оцінку адекватності моделі реальній дійсності, оскільки їх можна

інтерпретувати, ретельно оцінювати та охоплювати різні аспекти точності прогнозування та відповідності моделі. Загалом вони дозволяють нам оцінювати та повідомляти про ефективність наших моделей у прогнозуванні цін на акції. Результати наших моделей наведено в таблиці 3.2.

Таблиця 3.2. Порівняння коефіцієнтів

Model	RMSE	Deviation Percentage	R-squared
Linear Regression	27.87	394.24%	76.22%
Random Forest	9.53	34.61%	97.22%

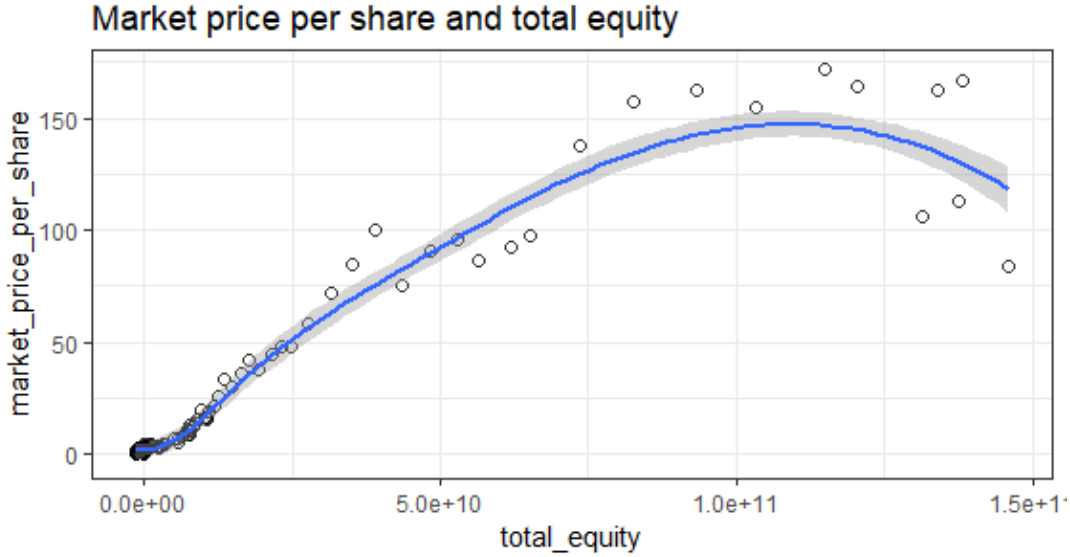

Рис. 3.5. Ринкова ціна акції та загальний капітал (графік)

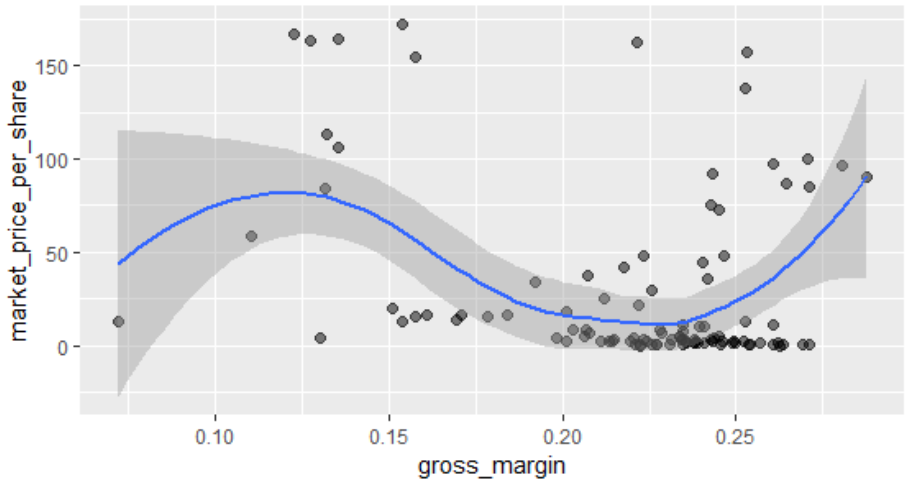

Рис. 3.6. Ринкова ціна за акцію та валовий прибуток (поліноміальна залежність)

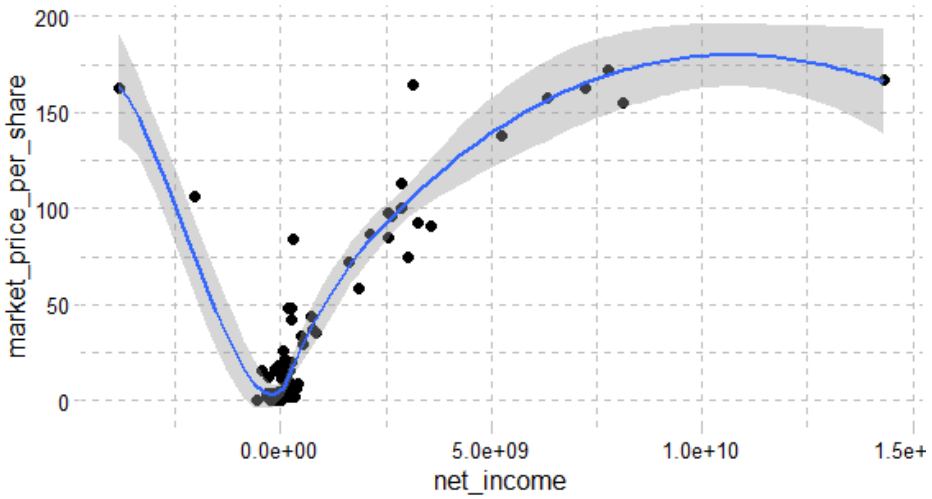

Рис. 3.7. Ринкова ціна акції та чистий прибуток (пряма та поліноміальна залежності)

На основі результатів оцінки модель Random Forest краще прогнозує ціни акцій, використовуючи фінансові показники, ніж модель лінійної регресії. Модель Random Forest має нижчий RMSE 9,54 (порівняно з 27,87 для лінійної регресії) і відсоток відхилення 34,61% (порівняно з 394,24% для лінійної регресії), що означає, що вона має кращу точність прогнозування та менше помилок прогнозування. Крім того, модель Random Forest має вище значення R-квадрат 97,22% (порівняно з 76,22% для лінійної регресії), що вказує на те, що вона має кращу загальну пояснювальну силу для фіксації мінливості цін на акції. Однак слід бути обережним, переоцінюючи модель, оскільки додаткові фактори також можуть вплинути на результат. Можна виділити пряму (рис. 3.5) і поліноміальну (рис. 3.6-3.7) залежності між ринковою ціною акції та статистично значущими пояснюючими змінними.

ВИСНОВКИ

Було проведено аналіз ряду попередніх досліджень у сфері аналізу фінансових даних та прогнозування фінансових показників, що підкреслив значущий потенціал використання методів машинного навчання у фінансовій сфері. Зазначені дослідження свідчать про ефективність глибоких нейронних мереж, регресійного аналізу та інших методів машинного навчання у прогнозуванні цін на акції, фінансових ринків та інших ключових фінансових показників. Також, дослідження в галузі обробки природної мови акцентують на важливості аналізу настроїв інвесторів для оцінки їх реакції на фінансові звіти компаній.

Також було досліджено аналіз ключових показників які можно отримати з бізнес звіту 10-K, такі як валовий дохід, операційний дохід, маржа чистого прибутку, коефіцієнт поточної ліквідності, співвідношення боргу до власного капіталу, рентабельність активів, рентабельність власного капіталу, прибуток на акцію та інші ключові фінансові показники.

Також ми дослідили і описали існуючі методи машиного навчання та інтелектуального видобутку даних та їх використання у фінансовій сфері.

На основі досліджених даних ми побудували та порівняли дві моделі використовуючі різні алгоритми Random Forest та множинна лінійна регресія. Random Forest продемонстрував кращу якість у прогнозуванні цін на акції порівняно з множинною лінійною регресією. Він досяг значно нижчих значень як для RMSE (9,53), так і для відсотка відхилення (34,61%), що вказує на покращену точність і зменшення помилок передбачення. Крім того, модель Random Forest продемонструвала вище значення R-квадрат 97,22%, підкреслюючи її винятковий пояснювальний вплив. Це означає, що фінансові показники, які використовуються в моделі, роблячи її більш надійною, можуть пояснити приблизно 97,22% мінливості цін на акції. Відсоток відхилення в 34,61% свідчить про те, що все ще є місце для покращення точності моделі.

Попри те, що Random Forest перевершив множинну лінійну регресію, важливо враховувати інші фактори, які можуть вплинути на ціни акцій, крім вибраних фінансових показників.

Загалом результати дослідження показали, що Random Forest є ефективним інструментом для аналізу залежності курсу акцій від різних змінних. Однак наша модель не враховує всі фактори, що впливають на ціни акцій, тому інвестори мають додатково аналізувати ринок для кожного окремого фінансового інструменту, щоб приймати правильні інвестиційні рішення.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. «Interactive Brokers». – [Електронний ресурс]. – Режим доступу: https://www.interactivebrokers.com/en/home.php
2. «Robinhood». – [Електронний ресурс]. – Режим доступу: https://robinhood.com
3. «Rules and Use of Form 10-K». – [Електронний ресурс]. – Режим доступу: https://www.sec.gov/files/form10-k.pdf
4. Zanc, R., Cioara, T., Anghel, I.: Forecasting Financial Markets using Deep Learning, 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, 2019, pp. 459-466, doi: 10.1109/ICCP48234.2019.8959715.
5. Wasserbacher, H., Spindler, M. Machine learning for financial forecasting, planning and analysis: recent developments and pitfalls. Digit Finance 4, 63–88 (2022). doi: 10.1007/s42521-021-00046-2.
6. Doryab, B., Salehi, M.: Modeling and forecasting abnormal stock returns using the nonlinear Grey Bernoulli model. Journal of Economics, Finance and Administrative Science 23 (44), 95-112 (2018). doi: 10.1108/JEFAS-06-2017-0075.
7. Snihovyi, O., Ivanov, O., Kobets, V.: Cryptocurrencies prices forecasting with anaconda tool using machine learning techniques. In: Ermolayev, V. et al. (eds.) Proceedings of 14th International Conference ICTERI, pp. 453–456. CEUR-WS 2105, Aachen University (2018). https://ceur-ws.org/Vol-2105/10000453.pdf
8. Mushtaq, R., Gull, A.A., Shahab, Y., Derouiche, I.: Do financial performance indicators predict 10-K text sentiments? An application of artificial intelligence. Research in International Business and Finance 61, 101679 (2022). doi: 10.1016/j.ribaf.2022.101679

9. Huang, A.H., Zang, A.Y., Zheng, R., 2014. Evidence on the information content of text in analyst reports. Acc. Rev. 89 (6), 2151–2180 (2014). doi: 10.2308/accr-50833.
10. De Souza, J.A.S., Rissatti, J.C., Rover, S., Borba, J.A.: The linguistic complexities of narrative accounting disclosure on financial statements: An analysis based on readability characteristics. Res. Int. Bus. Finance 48, 59–74 (2019). DOI: 10.1016/j.ribaf.2018.12.008.
11. Cohen, L., Malloy, C., Nguyen, Q.: Lazy prices. J. Finance 75 (3), 1371–1415 (2020). doi: 10.1111/jofi.12885.
12. Kobets, V., Yatsenko, V., Mazur, A., Zubrii, M. Data analysis of personalized investment decision making using robo-advisers (2020) Science and Innovation, 16 (2), pp. 80-93. DOI: 10.15407/SCINE16.02.080.
13. Kilinich, D., Kobets, V. Support of investors' decision making in economic experiments using software tools (2019) CEUR Workshop Proceedings, 2393, pp. 277-288. https://ceur-ws.org/Vol-2393/paper_273.pdf
14. Snihovyi, O., Ivanov, O., Kobets, V. Implementation of Robo-Advisors Using Neural Networks for Different Risk Attitude Investment Decisions (2018) 9th International Conference on Intelligent Systems 2018: Theory, Research and Innovation in Applications, IS 2018 - Proceedings, art. no. 8710559, pp. 332-336. DOI: 10.1109/IS.2018.8710559.
15. «What Is Data Mining? How It Works, Benefits, Techniques, and Examples». – [Електронний ресурс]. - Режим доступу: https://www.investopedia.com/terms/d/datamining.asp
16. «AI And Data Mining: Do You Have The Keys To The Castle?». – [Електронний ресурс]. - Режим доступу: https://www.forbes.com/sites/forbestechcouncil/2022/07/26/ai-and-data-mining-do-you-have-the-keys-to-the-castle/?sh=1b4b679a5cc5

17. Pang-Ning Tan, M. Steinbach, A. Karpatne, V. Kumar: Introduction to Data Mining
18. «Вирішення задач класифікації та регресії». – [Електронний ресурс]. – Режим доступу: http://moodle.chdu.edu.ua/mod/resource/view.php?id=450
19. «Пошук асоціативних правил. Алгоритм Apriori». – [Електронний ресурс]. – Режим доступу: http://moodle.chdu.edu.ua/mod/resource/view.php?id=452
20. «Інтелектуальний аналіз даних». – [Електронний ресурс]. – Режим доступу: https://prezi.com/dqroov5n3wet/presentation/
21. «Інтелектуальний аналіз даних. Класифікація і регресія В». – [Електронний ресурс]. – Режим доступу: https://ukrbukva.net/print:page,1,44734-Intellektual-nyiy-analiz-dannyh-Klassifikaciya-i-regressiya.html
22. «Amazon.com, Inc. (AMZN)». – [Електронний ресурс]. – Режим доступу: https://finance.yahoo.com/quote/AMZN/
23. «Financial Modeling». – [Електронний ресурс]. – Режим доступу: https://site.financialmodelingprep.com/
24. «sklearn». – [Електронний ресурс]. – Режим доступу: https://scikit-learn.org