НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені ІГОРЯ СІКОРСЬКОГО»
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ
ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

На правах рукопису
УДК 004.62

До захисту допущено
Завідувач кафедри ММСА
_____ Оксана ТИМОЩУК
«___» _____ 2024 р.

Магістерська дисертація
на здобуття ступеня магістра
за освітньо-професійною програмою «Системний аналіз фінансового ринку»
зі спеціальності 124 «Системний аналіз»
на тему: «Моделі інтелектуального аналізу даних для оцінювання фінансових даних»

Виконав:
Студент 2 курсу, групи КА-22мп
Коваленко Олександр Максимович _____

Науковий керівник:
Старший викладач кафедри ШІ, д.ф.
Гуськова Віра Геннадіївна _____

Рецензент:
Старший викладач кафедри ШІ, д.т.н.
Шаповал Наталія Віталіївна _____

Засвідчую, що у цій магістерській дисертації
немає запозичень з праць інших авторів
без відповідних посилань
Студент (*підпис*): _____

Київ – 2024

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

Рівень вищої освіти — другий (магістерський)
Спеціальність — 124 «Системний аналіз»
Освітньо-професійною програмою «Системний аналіз фінансового ринку»

ЗАТВЕРДЖУЮ
Завідувач кафедри ММСА
_____Оксана ТИМОЩУК
«___» _____ 2024 р.

ЗАВДАННЯ

на магістерську дисертацію студенту

Коваленку Олександру Максимовичу

1. Тема дисертації: «Моделі інтелектуального аналізу даних для оцінювання фінансових даних», науковий керівник дисертації Гуськова Віра Геннадіївна, старший викладач кафедри ШІ, д.ф., затверджені наказом по університету від «08» листопада 2023 р. № 5200-с

2. Строк подання студентом дисертації _____

3. Об’єкт дослідження: фінансові дані

4. Предмет дослідження: математична статистика, нейронні мережі, моделі і методи інтелектуального аналізу даних.

5. Перелік завдань, які потрібно розробити, наведений нижче:

1) проаналізувати існуючі методи, що використовуються для аналізу даних, обробки відсутності даних;
2) обрати методи для аналізу, фільтрації, обробки, прогнозу;
3) побудувати моделей навчання та їх аналіз;
4) проаналізувати результати;
5) зробити висновки.

6. Перелік графічного (ілюстративного) матеріалу:

1) презентація.

7. Орієнтовний перелік публікацій:

1) Коваленко О.М., Гуськова В.Г. Моделі інтелектуального аналізу даних для оцінювання фінансових даних. ІІ Всеукраїнська науково-практична конференція «Системні науки та інформатика» з нагоди 125-річчя КПІ ім. Ігоря Сікорського, м. Київ, 04-08 грудня 2023 року. С. 139-146.

2) Коваленко О.М., Муравльов А.Д., Петровський В.Є., Гуськова В. Г. Прогнозування фінансових показників шляхом удосконалення аналітичних методів та моделей на основі передпроцесингу даних. XXII-а Міжнародна науково-практична конференція «Інформаційно-комунікаційні технології та сталий розвиток». Інститут телекомунікацій і глобального інформаційного простору НАН України. С. 70-73.

8. Консультанти розділів дисертації*

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

9. Дата видачі завдання 01.09.2023 р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Строк виконання етапів магістерської дисертації	Примітка
1.	Інструктаж з техніки безпеки	01.09.2023 – 06.09.2023	Виконано
2.	Затвердження теми магістерської дисертації (МД) Ознайомлення зі структурою МД згідно з Положення про випускну атестацію студентів КПІ ім. Ігоря Сікорського [Електронний ресурс]	07.09.2023 – 13.09.2023	Виконано
3.	Ознайомлення з ДСТУ 3008:2015 та стандарти Єдиної системи програмної документації (ЄСПД)	21.09.2023 – 29.09.2023	Виконано
4.	Ознайомлення з державним стандартом України ДСТУ ГОСТ 7.1:2006 “Система стандартів з інформації, бібліотечної та видавничої справи. Бібліографічний запис. Бібліографічний опис. Загальні вимоги та правила складання”	28.09.2023 – 04.10.2023	Виконано

5.	Проведення дослідження за темою МД під керівництвом наукового керівника	05.10.2023 – 11.10.2023	Виконано
6.	Завершення роботи над першим варіантом основної частини МД	12.10.2023 – 18.10.2023	Виконано
7.	Продовження роботи над експериментальною частиною МД та програмним забезпеченням, оформлення розділу із стартап проекту.	19.10.2023 – 04.12.2023	Виконано
8.	Оформлення магістерської дисертації.	05.12.2023 – 31.12.2023	Виконано

Студент _____ Олександр КОВАЛЕНКО

Науковий керівник дисертації _____ Віра ГУСЬКОВА

РЕФЕРАТ

Магістерська дисертація: 82 с., 22 табл., 10 рис., 1 дод., 19 джерел.

Об’єкт дослідження: фінансові дані, а саме записи продажів акцій фондового ринку.

Предмет дослідження: прогнозування зростання/спадання курсу продажів.

Метою роботи є дослідження і створення моделей та розробка ефективних методів і алгоритмів інтелектуального аналізу фінансових даних з задачею прогнозування.

Постановка задачі: основною поставленою задачею було створення і порівняння моделей інтелектуального аналізу фінансових даних для дослідження їх ефективності.

Результатом роботи є програмний продукт розроблений з використанням мови програмування Python, що приймає індекс компанії і прогнозує подальшу ціну акції.

Ключові слова: LSTM, ARIMA, НЕЙРОННА МЕРЕЖА, ФОНДОВИЙ РИНОК, ІАД, ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, МОДЕЛЬ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ, ПРОГНОЗУВАННЯ.

ABSTRACT

Master's thesis: 82 p., 22 tables, 10 fig., 1 appendix, 19 sources.

Object of research: financial data, specifically records of stock market sales.

Subject of research: forecasting the growth/decline of sales rates.

The purpose of this work is to investigate and create models and develop effective methods and algorithms for data mining of financial data for the purpose of forecasting.

Problem statement: The main task was to create and compare models of data mining of financial data to investigate their effectiveness.

The result of this work is a software product in the Python programming language that takes a company index and forecasts the future stock price.

Keywords: LSTM, ARIMA, NEURAL NETWORK, STOCK MARKET, DATA MINING, MODEL OF DATA MINING, FORECASTING.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ІАД – Інтелектуальний Аналіз Даних (Data Mining)

RDBMS – Relational Database Management System – реляційна система управління базами даних

OLAP – Online Analytical Processing – аналітична обробка в реальному часі

ML – машинне навчання

SQL – Structured Query Language – Мова структурованих запитів

ЧР – часовий ряд

ВСТУП

Сучасний світ щоденно виробляє 2,5 мільярди гігабайт інформації на день, що накопичуються, зберігаються і потребують аналізу. Те ж стосується і економічних даних. Однак, зростання обсягів даних та ускладнення їх структури створюють виклики для традиційних методів аналізу.

Відповідно, Інтелектуальний Аналіз Даних (ІАД) стає все більш актуальним і потужним інструментом для виявлення складних зв'язків, прогнозування тенденцій та виявлення аномалій у фінансових даних. Застосування новітніх методів дозволяє отримати нові інсайди і зробити обґрунтовані рішення, що сприяє підвищенню ефективності фінансового управління.

У цій магістерській роботі міститься дослідження та розробка моделей інтелектуального аналізу фінансових даних. Конкретно, робота спрямована на створення ефективних алгоритмів та моделей для оцінювання фінансових даних, прогнозування трендів та виявлення аномалій.

У першому розділі йдеться про актуальність теми, постановка задачі, огляд Інтелектуального Аналізу Даних та його моделей.

У другому розділі надана інформація про існуючі математичні моделі, що використовуються для вирішення задач ІАД та нормування даних.

У третьому розділі більш глибоко описано процес ІАД, маніпуляцій з даними, оцінка моделей та результати дослідження.

Четвертий розділ присвячений дослідженню доцільності створення стартап-проекту, що мав би за ціль створення програмного продукту для аналізу та прогнозування фінансових даних.

Додаток А містить лістинг коду програми мовою програмування Python, за допомогою якої були отримані результати даного дослідження.

РОЗДІЛ 1 АКТУАЛЬНІСТЬ ТЕМИ ДОСЛІДЖЕННЯ, ОГЛЯД ІСНУЮЧИХ РІШЕНЬ. ОГЛЯД МОДЕЛЕЙ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАННИХ, ПОСТАНОВКА ЗАДАЧІ

1.1 Актуальність теми, методи та моделі для прогнозування, постановка задачі

Актуальність теми базується на кількох ключових аспектах, які визначають потребу в такому дослідженні.

По-перше, зростання обсягів та складності фінансових даних створює потребу у вдосконаленні методів їх аналізу. Традиційні підходи до статистичного аналізу можуть бути недостатніми для отримання достатньо точних та надійних результатів. Тому, зростає необхідність використання інтелектуальних методів для здійснення більш точних прогнозів, виявлення складних зв’язків та ідентифікації аномалій у фінансових даних.

По-друге, за умов сучасного глобалізованого ринку, фінансовий аналіз відіграє критичну роль для різних зацікавлених сторін, таких як інвестори, фінансові установи та регулятори. Точні прогнози, ефективна ідентифікація ризиків та аномалій є ключовими для прийняття раціональних рішень та запобігання фінансовим кризам.

Методи інтелектуального аналізу даних використовуються для виявлення закономірностей і залежностей у великих обсягах неструктурованих даних.

Загальний результат ІАД можна охарактеризувати як знання щодо закономірностей і тенденцій, що повинно мати такі властивості:

- він відображає результати дослідження системи, відображаючи об'єктивну реальність;
- він представлений у зрозумілій людині формі, використовуючи загальноприйняті символи, поняття та природну мову;

– він компактний у своєму описі, що дозволяє його легко розуміти, інтерпретувати та використовувати.

Data Mining надає можливість знаходити нові гіпотези про поведінку невідомих, але реально існуючих залежностей в даних, створювати моделі, які можуть оцінити ступінь впливу факторів, що досліджуються [1].

Існує два загальних типи ІАД – на основі верифікації (verification-driven data mining) та на основі виявлення (discovery-driven data mining).

Аналіз даних на основі верифікації використовує структуровані запити (SQL) та багатовимірний, статистичний аналіз для отримання результатів. Це включає прогностичне моделювання, виявлення аномалій, аналіз зв'язків та сегментацію баз даних.

Прогностичний та описовий Інтелектуальний Аналіз Даних - це дві основні категорії інтелектуального аналізу даних, що засновані на виявленні даних. Прогностичний аналіз використовується для прогнозування майбутніх подій або тенденцій, тоді як описовий аналіз аналізує кореляцію, перехресну табуляцію та частоту для виявлення схожостей в даних і виявлених закономірностей [2].

Прогностичний аналіз є однією з основних категорій методів інтелектуального аналізу даних. Його основна мета полягає в прогнозуванні майбутніх подій, трендів або значень на основі наявних даних і встановлення взаємозв'язків між змінними.

Інтелектуальний Аналіз Даних найчастіше розв'язує чотири основних завдання: асоціацію, кластеризацію, класифікацію і регресію. Нижче наведений їх короткий опис.

1. Асоціація: завдання асоціації полягає в виявленні кореляцій та залежностей між різними елементами набору даних. Він дозволяє виявити часті комбінації елементів і встановити, які елементи часто зустрічаються разом.

2. Кластеризація: кластеризація використовується для групування схожих об'єктів разом на основі їхніх спільних характеристик. Метою кластеризації є

створення груп або кластерів, де об'єкти всередині кожного кластера подібні один до одного, а об'єкти з різних кластерів відрізняються.

3. Класифікація: класифікація відносить об'єкти до певних попередньо визначених категорій або класів на основі їхніх властивостей. Це завдання полягає в розробці моделі, яка може прогнозувати, до якого класу належить новий об'єкт на основі знань, набраних з попереднього набору даних.

4. Регресія: регресія використовується для прогнозування числових значень на основі залежностей між змінними. Вона дозволяє побудувати математичну модель, яка може передбачати значення однієї змінної на основі інших змінних.

Навчання у методах ІАД може бути як з вчителем, так і без вчителя. Це означає, що при навчанні з вчителем є дані, які містять маркер для прогнозування, тоді як без вчителя заздалегідь правильних результатів для навчання немає [3].

Нижче наведена порівняльна характеристика обох методів (таблиця 1.1).

Таблиця 1.1. Різниця навчання з вчителем і без вчителя

Навчання з вчителем	Навчання без вчителя
Вхідні дані марковані	Вхідні дані не марковані
Має механізм зворотного зв’язку	Не має механізму зворотного зв’язку
Дані класифікуються на основі навчального набору даних	Надаються властивості заданим даним для їх класифікації
Задачі регресії та класифікації	Задачі кластеризації, зменшення розмірності та асоціації
Основна задача - прогнозування	Основна задача – аналіз
Відома кількість класів	Невідома кількість класів

Також існують ситуації нерівномірного розподілення даних у датасеті, коли з’являється необхідність застосовувати гібридний тип навчання, як з вчителем, так і без вчителя. Наприклад, кількість маркованих даних більша. В такому разі відбувається навчання з вчителем, будується аналітична модель, після цього відбувається навчання без вчителя з підкріпленням і побудовою аналітичної моделі та просто використовуючи навчання з підкріпленням. В таблиці 1.2 наведені приклади алгоритмів, які використовуються в цих методах.

Таблиця 1.2. Приклади алгоритмів навчання різних методів ІАД

№	Навчання з вчителем	Навчання без вчителя	Часткове навчання з вчителем	Навчання з підкріпленням
1	Лінійна регресія (Linear regression)	Кластеризація методом к-середніх (K-means clustering)	Самонавчений наївний байєсів класифікатор (Selftrained Naïve Bayes classifier)	Q-навчання (Qlearning)
2	Метод к-найближчих сусідів (Knearest neighbor algorithm)	Багатовимірне шкалювання (Multidimensionals caling - MDS)	Трансдуктивний метод опорних векторів (Transductive support vector machines)	Розширення значення на основі моделі (Model-based value expansion – MVE)
3	Метод опорних Векторів (Support vector machines - SVM)	Ієрархічна кластеризація (Hierarchical clustering)	Генеративна змагальна мережа (Generative adversarial networks)	
4	Дерево прийняття рішень (Decision tree)	Метод головних компонент (Principal component analysis)	Поширення міток на основі графу (Graphbased label propagation)	

Кінець таблиці 1.2

5	Наївний байєсів Класифікатор (Naïve Bayes)	Алгоритм Apriori	Довга короткочасна пам’ять (Long shortterm memory networks - LSTM)	
6	Нейронні мережі (Neural networks)	Метод зворотнього поширення помилки (Backpropagation)		
7	Логістична регресія (Logistic regression)	Глибинна мережа переконань (Deep belief network)		

Виходячи з конкретних умов, даних і чинників кожен з підходів може бути відповідним і ефективним для вирішення задачі.

1.2 Огляд сучасних методів при роботі з даними

SAS, R, Python та SQL є найрозповсюдженішими і найуживанішими мовами програмування, що застосовуються для інтелектуального аналізу даних.

SAS (Statistical Analysis System) - це інтегроване програмне забезпечення для аналізу даних і статистичних обчислень. SAS має багатий набір функцій і можливостей, що дозволяє проводити широкий спектр статистичних аналізів, включаючи дисперсійний аналіз, регресійний аналіз, кластерний аналіз, аналіз виживаності та багато інших. Крім того, SAS надає інструменти для візуалізації даних, створення звітів та графіків. Одна з переваг SAS полягає в його високій надійності та стабільності. Він має довгу історію використання в індустрії і вважається одним з провідних інструментів для аналізу даних.

В той самий час про R говорять так: «R - це мова програмування для статистичного аналізу, обчислення та графіки. Однією з сильних сторін R є легкість створення графіків, включаючи математичні позначення та формули.» [4]. Завдяки багатому набору пакетів і бібліотек, R дозволяє виконувати різноманітні завдання, такі як експлораторний аналіз даних, моделювання, кластеризація, аналіз тексту, машинне навчання та інші. Багато вчених, статистиків і аналітиків використовують R як основний інструмент для вирішення завдань аналізу даних та статистики.

Python є потужним інструментом для інтелектуального аналізу даних (ІАД). Він надає широкі можливості для обробки, аналізу та візуалізації даних. У Python існує велика кількість бібліотек, таких як NumPy, Pandas, SciPy та scikit-learn, які дозволяють виконувати різноманітні завдання ІАД, включаючи класифікацію, регресію, кластеризацію та інше. Багато інструментів для візуалізації, таких як Matplotlib та Seaborn, також доступні в Python.

SQL (Structured Query Language) - це стандартизована мова програмування, що використовується для управління та роботи з реляційними базами даних. Вона є основним інструментом для збереження, вибору, оновлення та видалення даних в базах даних. SQL дозволяє виконувати різноманітні операції з базою даних, такі як створення таблиць, вставку даних, вибірку, зміну та видалення записів, а також з'єднання декількох таблиць для складніших запитів. Одна з переваг SQL полягає в його простоті та зрозумілості. Він має декларативний підхід, що означає, що користувач описує бажаний результат, а не конкретний спосіб його отримання. Це робить SQL дружнім для користувачів, навіть для тих, хто не має глибоких знань в програмуванні. Варто відзначити, що багато RDBMS використовують мову SQL і мають свої відмінності в роботі, побудові процесів, інколи і в фукнціоналі. Найвідоміші серед яких, це MSSQL Server, PostgreSQL, Oracle, MySql.

Також, деякі компанії постачають інструменти та сервіси для аналізу даних. До таких компаній можна віднести Amazon з AWS, Google з BigQuery, DataBricks, H2O.ai, Microsoft, Tibco Software [5].

Останнім інструментом розглянемо RapidMiner. RapidMiner - це комплексна платформа для науки про дані з візуальним проектуванням робочих процесів та повною автоматизацією. Це означає, що розробнику не потрібно писати код для завдань з data mining. Інтерфейс можна переглянути на рисунку 1.1

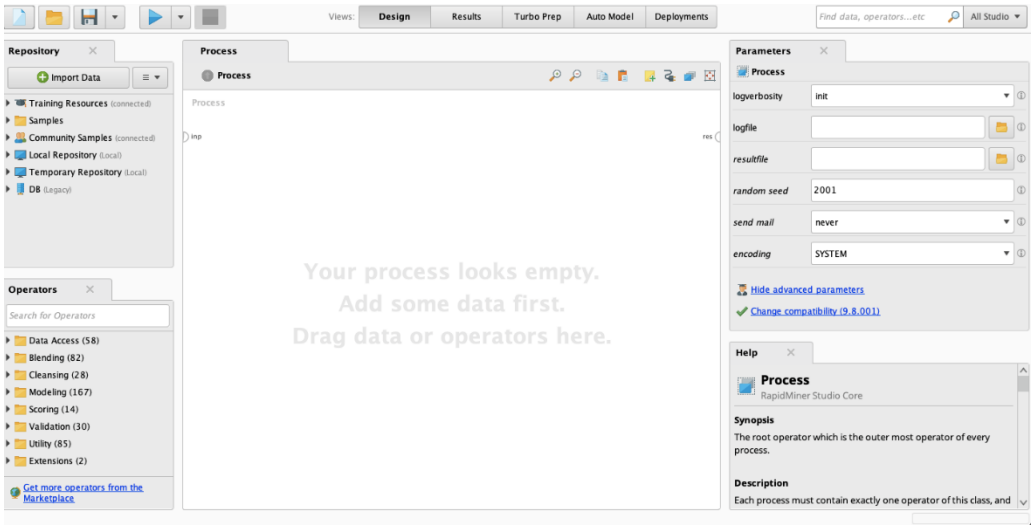

Рисунок 1.1 Графічний інтерфейс RapidMiner

Це графічний інтерфейс порожнього процесу в RapidMiner. Він має репозиторій, який містить наш набір даних. Існує можливість імпортувати свої власні набори даних. Також доступно багато користувацьких наборів даних, які розробник може завантажити та обробляти. Також є функція відтворення з'єднання з базою даних.

Понизу від репозиторію є вікно з оператором. Оператор включає в собі все, що необхідно для побудови процесу інтелектуального аналізу даних. Наприклад, очищення даних, моделювання, валідація, оцінювання.

Права частина графічного інтерфейсу містить вікно з параметрами. Параметри можуть змінювати поведінку і налаштовувати оператори [6].

1.3 Обґрунтування вибору мови програмування та середовищ розробки

1.3.1 Мова програмування Python

Python - це потужна мова програмування, яка знайшла широке застосування у сфері аналізу даних. Вона має велику кількість бібліотек і інструментів, що роблять її популярним вибором для роботи з даними.

Вкажемо декілька причин чому варто обирати Python [7].

1. Якість фінального програмного продукту. Код на Python проектується таким чином, аби бути читабельним, послідовним та повторно використовуваним. Це робить його легким для розуміння, навіть якщо він був написаний іншим розробником.

2. Продуктивність розробки. Порівнюючи зі статично типізованими мовами програмування, як от C, C++, Java, код на Python писати легше і швидше, його сумарний лістинг може бути вполовину меншим.

3. Портативність програм. Перенесення коду між Windows та Linux зазвичай зводиться до копіювання коду між машинами. До того ж, пропонується кілька варіантів написання портативних графічних інтерфейсів користувача, програм доступу до баз даних, систем на основі вебу та інше.

4. Бібліотеки. По-перше, Python з нуля має велику кількість вбудованих функцій. По-друге, велика кількість додаткового матеріалу вже пропонується і постійно підтримується у вільному доступі. По-третє, написання власних бібліотек не є складною задачею і часто використовується програмістами для досягнення власних цілей.

5. Інтеграція компонентів. Скрипти Python легко обмінюються з іншими частинами додатку за допомогою різних механізмів інтеграції. Сьогодні код Python може використовувати бібліотеки C і C++, може бути

викликаний з програм на C і C++, може інтегруватися з компонентами Java і .NET, може взаємодіяти з API через інтерфейси SOAP, XML-RPC і CORBA.

1.3.2 Мова програмування R

Згідно з офіційною документацією R, R – це мова програмування та середовище для статистичних обчислень і графіків. Ця мова з відкритим кодом зазвичай використовується для аналізу даних.

Середовище R складається з інтегрованого набору програмних засобів, створених для обробки даних, обчислень і графічного відображення. Середовище має наступні особливості:

- високопродуктивний механізм зберігання та обробки даних;
- набір операторів для роботи з масивами, зокрема матрицями;
- великий, легко зрозумілий та інтегрований асортимент засобів, призначених для аналізу даних;
- графічні можливості для аналізу та відображення даних, які працюють як на екрані, так і для друку;
- добре розроблена, проста та ефективна мова програмування, яка має можливість визначення рекурсивних функцій користувача, циклів, умовних операторів та засобів введення-виведення.

Синтаксис R складається з трьох елементів:

1) змінних, які зберігають дані;
2) коментарів, які використовуються для покращення зрозумілості коду;
3) ключових слів, зарезервованих слів, які мають особливий зміст для компілятора.

Пакети R визначаються як збірки функцій R, вибіркових даних, документації та компільованого коду. Ці елементи зберігаються в каталозі,

який називається "library" всередині середовища R і встановлюються за замовчуванням під час установки.

Пакети R підвищують потужність R, поліпшуючи наявні функціональні можливості і збираючи набори функцій R в один блок. Крім того, пакет R є пере використовуваним ресурсом, що робить життя програміста набагато простішим.

Загалом доступно більше 7 тисяч різних пакетів, які розширяють можливості написання коду. Слід відзначити такі пакети.

1. dplyr: пакет dplyr надає потужні функції для маніпулювання даними. Він пропонує простий та консистентний синтаксис для фільтрації, сортування, групування, вибору та з'єднання даних. Використання dplyr допомагає зменшити код та зробити маніпуляцію даними більш ефективною.

2. ggplot2: цей пакет надає потужні можливості для створення графіків та візуалізації даних. Він базується на "граматиці графіків", що дозволяє легко створювати високоякісні графіки з гнучкістю в налаштуванні. ggplot2 допомагає відтворити дані у зрозумілій та привабливій формі, сприяючи розумінню залежностей та паттернів.

3. Caret: пакет caret (Classification And REgression Training) надає зручний фреймворк для тренування та оцінки моделей машинного навчання. Він має широкий вибір алгоритмів для класифікації, регресії, кластеризації та інших завдань. caret спрощує процес вибору та оптимізації моделей, надаючи зручні інструменти для налаштування параметрів та оцінки результатів.

4. Tidyr: пакет Tidyr допомагає перетворювати та організовувати дані у вигляді "довгих" або "широких" таблиць. Він має функції для розпакування, групування та форматування даних. Tidyr дозволяє легко маніпулювати структурою даних, щоб підготувати їх для аналізу та візуалізації.

1.4 Висновки до розділу

У даному розділі було розглянуто декілька аспектів, що визначають потребу в такому дослідженні, а саме:

- потреба в удосконаленні методів аналізу зростаючих об’ємів фінансових даних;
- потреба у точніших прогнозах, ефективнішій ідентифікації ризиків та аномалій для зацікавлених сторін.

Було розглянуто поняття ІАД, його типи, задачі. В розділі наведені таблиці порівняння навчання з вчителем і без вчителя, а також приклади алгоритмів різних методів ІАД.

Також, було наведено декілька найуживаніших мов програмування задля здійснення інтелектуального аналізу даних. Серед них SAS, R, Python та SQL. Було висвітлено їх основні переваги та недоліки. Для R та Python слабкі та сильні сторони були описані більш детально.

РОЗДІЛ 2 МАТЕМАТИЧНІ МОДЕЛІ ДЛЯ АНАЛІЗУ ДАНИХ ТА ПРОГНОЗУВАННЯ

2.1 Основні поняття

Для створення моделі, яка б описувала та пояснювала закономірності даних, потрібно розуміти, що дані можуть використовуватися для побудови просторових моделей і для побудови моделей часових рядів. Перший вид описує певну кількість процесів у конкретний момент часу t, а другий описує тільки один процес за інтервал часу t.

Часовий ряд – це послідовність числових показників, що упорядковані у часі і описують рівень стану і зміни досліджуваного об’єкту. Він характеризується своєю сезонністю, трендом, циклом та похибкою. Поділяються на стаціонарні та нестаціонарні.

Стаціонарними називаються такі часові ряди, характер яких не змінюється з часом. Вимога до стаціонарного часового ряду полягає у присутності сталого середнього значення і в той самий час інші значення коливаються навколо цього середнього зі сталою дисперсією. Така особливість невластива нестаціонарним часовим рядам, при цьому зберігаючи властивості сезонності та тренду. Приклади таких рядів наведені нижче (рисунок 2.1 та рисунок 2.2).

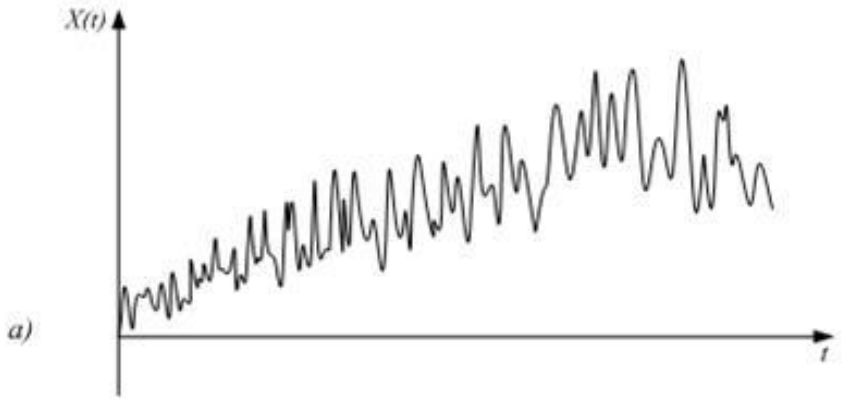

Рисунок 2.1. Вигляд нестаціонарного часового ряду [18]

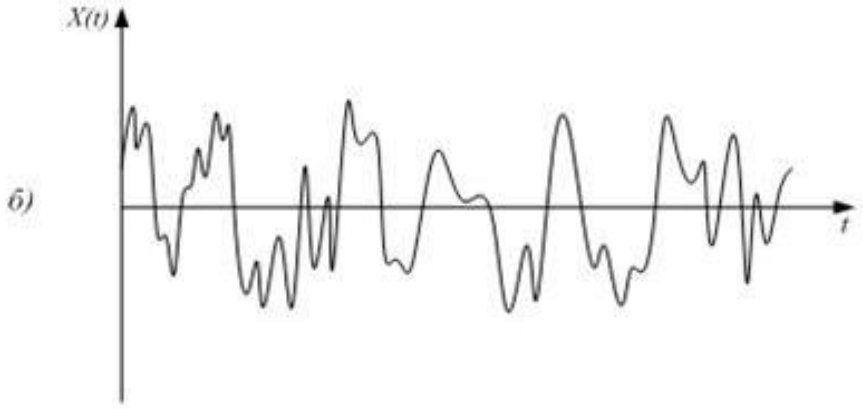

Рисунок 2.2 Вигляд стаціонарного часового ряду [18]

Процес породження наявних даних є лінійним для стаціонарних часових рядів і зазвичай не мають тренду, або періодичної зміни середнього та дисперсії. Перевірити гіпотезу стосовно сталості середнього значення та дисперсії часового ряду можна виконати кількома способами. Вкажемо один з найпростіших способів.

1. Перевірити значущість різниці двох середніх значень підмножин вибірки за критерієм перевірки гіпотези про рівність середніх двох нормально розподілених вибірок (z-критерій).

2. Перевірити сталість дисперсії. Наприклад, використовуючи F-критерій (критерій Фішера про відношення вибіркових дисперсій).

Розглянемо метод перевірки різниць середніх рівнів. Початковим етапом є розділення набору даних на дві приблизно однакові частини. Для кожної частини обраховуються середні значення й дисперсії.

Далі відбувається перевірка однорідності дисперсій за допомогою критерію Фішера за формулою:

$$F = \frac{\sigma_1^2}{\sigma_2^2}$$

і відбувається порівняння із табличним значенням критерію Фішера із заданим рівнем значущості. Якщо розрахункове значення менше за табличне, то гіпотеза про рівність приймається. В інакшому випадку це означає, що запропонований метод не відповідає на питання наявності тренду.

Наступним кроком є перевірка гіпотези про відсутність тренду по критерію Стьюдента (t-критерій). Відбувається розрахунок за формулою:

$$t = \frac{|M_1 - M_2|}{\sqrt{\frac{(N_1 - 1)\sigma_1^2 + (N_2 - 1)\sigma_2^2}{N_1 + N_2 - 2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}},$$

де M – середнє арифметичне, σ – стандартне відхилення, а N – розміри виборок. Якщо розраховане значення t менше за табличне, тоді тренд відсутній, інакше – тренд є. Цей метод застосовується суто для рядів із монотонною тенденцією, що і є його недоліком, бо він не може правильно визначити існування тренду в разі, коли ЧР містить точку зміни тенденції у середині ряду.

2.2 Робота з пропущеними даними

У роботі з даними завжди присутні фактори, що зможуть погіршити умови подальшої їх обробки. Це може бути системний збій, проблеми з технічної сторони або людський фактор. Все це може призвести до часткової або повної втрати певних даних. Тому постає питання про вирішення питання прогалин і неточностей в наборах даних.

Найпростішим рішенням обробки даних, що мають прогалини, є виключення некомпетентних спостережень, що містять пропуски, і подальша робота з такими «повними» даними. За такого підходу можемо спостерігати сильну відмінність між результатами, що були отримані з початкових даних і вже оброблених. В такому разі маємо звернутися до інших методів – методів заповнення пропусків перед аналізом масиву даних.

Для того, щоб належно елімінувати пропуска, необхідно зрозуміти механізми їх формування. Відповідно до загальноприйнятих визначень, описують три види формування пропусків.

1. MCAR (Missing Completely At Random) – вид наявності пропусків, в якому кожне поле має однакову імовірність пропуску. За такого механізму, ігнорування чи вилучення записів, зазвичай, не веде до сильного погіршення результатів, а часто і зовсім не впливає.
2. MAR (Missing At Random) – вид наявності пропусків, в якому ймовірність пропуску може бути визначена на основі іншої наявної інформації без пропусків. Таким чином це не випадково пропущені характеристики і причиною їм є певні закономірності, тому як і в MCAR, це не веде до суттєвого спотворення результатів.
3. MNAR (Missing Not At Random) – вид наявності пропусків, коли невідомі чинники визначають існування чи не існування прогалин в наборі даних. В наслідок, на основі даних неможливо якось оцінити ймовірність прогалин.

Метод Hot Deck. Цей метод використовує заміну пропущеного значення найближчим доступним інформаційним об'єктом. Пропущені дані можна відновлювати з усієї групи повних спостережень або з певної підгрупи, такої як кластер, до якого належить цільовий об'єкт. Для заповнення пропуску використовується значення цієї характеристики з найближчого об'єкта до цільового. Вибір типу функції відстані для визначення найближчого спостереження залежить від типу досліджуваних даних та характеру зв'язку між змінними, а також від постановки конкретного дослідження.

Метод Барлета. Він включає два етапи: спочатку на першому етапі використовується заміщення пропущених значень початковими згенерованими значеннями; на другому етапі проводиться коваріаційний аналіз цільової змінної та побудова дихотомічного індикатора повноти спостережень за цією змінною. Індикатор повноти спостережень завжди

дорівнює 0, крім одного випадку: якщо і-те значення є пропущеним для цільової змінної, то індикатор набуває значення 1.

Алгоритм ZET. Алгоритм ZET використовує частину сукупності спостережень, відому як компонентна матриця, для заповнення пропущених значень. Компонентна матриця складається з компонентних рядків і стовпців, і вона обмежена лише на деяку частину даних, а не на всю сукупність спостережень. Кожен компонентний рядок має величину, яка залежить від декартової відстані до цільового рядка, який має пропущене значення. За допомогою компонентної матриці будується функціональна залежність між прогнозним значенням і відповідним значенням в компонентній матриці, що дозволяє прогнозувати пропущені значення.

Регресійне моделювання. Регресійне моделювання є одним з методів заповнення пропущених даних. Цей метод використовує статистичну модель, яка описує залежність між змінними, щоб прогнозувати пропущені значення на підставі наявних даних.

Основна ідея регресійного моделювання полягає в тому, що вибірка даних використовується для побудови математичної моделі, яка визначає залежність між змінними. Ця модель може бути лінійною, поліноміальною або використовувати інші функції для опису залежності. За допомогою цієї моделі можна прогнозувати значення пропущених даних на основі значень інших змінних, що вже відомі.

Процес регресійного моделювання включає такі етапи.

1. Вибір моделі: вибір підходящої математичної моделі, яка найкраще підходить для опису залежності між змінними.

2. Підготовка даних: підготовка даних шляхом видалення пропущених значень або їх заповнення за допомогою інших методів.

3. Будування моделі: побудова математичної моделі, використовуючи наявні дані.

4. Оцінка моделі: оцінка точності та якості моделі за допомогою метрик, таких як середня квадратична помилка (MSE) або коефіцієнт детермінації (R^2).

5. Прогнозування: застосування побудованої моделі для прогнозування пропущених значень на основі вхідних даних.

Регресійне моделювання дозволяє заповнити пропущені значення, використовуючи інформацію, що міститься в інших змінних. Воно може бути корисним методом, коли дослідник має достатньо даних для побудови адекватної моделі. Однак, варто враховувати, що результати заповнення пропусків залежать від вибору моделі та вірогідності правильної оцінки залежності між змінними. Тому важливо аналізувати результати та проводити валідацію моделі, щоб переконатися в її адекватності і надійності.

2.3 Масштабування. Нормування

Масштабуванням ознак у розрізі інтелектуального аналізу даних називається метод нормалізації незалежних ознак у фіксованому діапазоні. Цей процес є необхідним етапом попередньої обробки набору даних, бо забезпечує ефективне та швидке виконання використовуваного алгоритму, оскільки дозволяє алгоритму навчання не зважувати більші значення, що є затратним в плані обчислювальних можливостей.

Загалом, можна виділити такі техніки масштабування ознак.

- Нормалізація мінмаксна (min-max scaling) – метод масштабування даних з розподільчими значенням між 0 та 1:

$$X_{new} = \frac{X_i - min\ (X)}{max(x) - min\ (x)}.$$

- Стандартизація / стандартна нормалізація (standardization / Z-score normalization) – метод масштабування ознаки, щоб вона мала розподіл з середнім значенням в 0, а величина стандартного відхилення дорівнює 1 (μ - середнє арифметичне значення і σ – стандартне відхилення):

$$X_{new} = \frac{x_i - \mu}{\sigma}.$$

- Робастне масштабування (robust scaling) – метод масштабування згідно з міжквартильним розмахом (IQR – міжквантильний розмах):

$$X_{new} = \frac{x_i - median}{IQR}.$$

- Гаусівські перетворення.

- Нормалізація одиночного вектора (unit vector normalization): $X_{new} = \frac{x_j}{||x_j||}$За такого варіанту, кожен запис у наборі даних розглядається як p-вимірний вектор і масштабування стискає/розтягує вектор до одиночної сфери. Дані візуалізуються як низка різнонаправлених векторів на p-вимірній одиничній сфері.

Нормалізація та гаусівські перетворення застосовуються в різних випадках залежно від розподілу ознак. Нормалізацію застосовують, коли розподіл ознаки не є гаусівським, наприклад, для алгоритмів KNN та нейронних мереж. Вона дозволяє привести значення ознак до одного діапазону, забезпечуючи рівномірну шкалу значень. Стандартизацію використовують, коли розподіл ознаки є гаусівським. Цей вид масштабування зберігає стандартне відхилення та середнє значення, що корисно для алгоритмів, які залежать від цих параметрів. Робастне масштабування використовують у випадках, коли є викиди даних в наборі. Воно використовує медіану та міжквартильний розмах, що робить його менш чутливим до викидів.

Алгоритми, які використовують градієнтний спуск (наприклад, лінійна та логістична регресія, нейронні мережі) або містять розрахунок відстані між точками даних (наприклад, KNN, K-means, SVM), вимагають масштабування даних. Для алгоритмів, які використовують дерева прийняття рішень або ансамблеві методи, такі як Gradient Boosting або Random Forest, нормалізація даних не є обов'язковою, оскільки масштабування не впливає на їх ефективність, враховуючи їх специфіку.

2.4 Математичні моделі та методи прогнозування фінансових даних

Фінансові ринки є складними системами, де кожна зміна може мати значний вплив на ціни активів та рівень ризику. Врахування цих факторів та прогнозування їх впливу стає важливим завданням для інвесторів, трейдерів та фінансових установ. У цьому розділі буде розглянуто різноманітні моделі та методи, які допомагають приймати обґрунтовані рішення на фінансових ринках.

Передбачення майбутньої поведінки фінансових ринків є складною задачею, оскільки вони піддаються впливу багатьох непередбачуваних факторів, таких як економічні події, політична нестабільність та інші зовнішні чинники. Використання математичних моделей та аналітичних методів дозволяє підвищити точність прогнозів та зробити більш обґрунтовані рішення на фінансових ринках. Ці моделі базуються на статистичних методах, математичних алгоритмах та комп'ютерному моделюванні, що дозволяє аналізувати великі обсяги даних та виявляти складні залежності між різними факторами. У цьому розділі буде розглянуто основні математичні моделі та методи, їх переваги та обмеження, а також викладемо приклади їх застосування для прогнозування фінансових ринків

2.4.1 Регресійний аналіз та регресійні моделі

Регресійний аналіз є одним з ключових методів прогнозування у фінансових ринках. Цей аналітичний підхід базується на встановленні залежностей між залежною змінною (цільовою змінною, яку необхідно прогнозувати) та набором незалежних змінних (факторів), що впливають на

цю залежну змінну. Регресійний аналіз дозволяє побудувати математичну модель, яка описує статистичну залежність між цими змінними і використовується для прогнозування майбутніх значень цільової змінної на основі відомих факторів. Цей підхід дозволяє інвесторам, трейдерам та фінансовим аналітикам отримувати числові прогнози щодо цін активів, доходності портфелів чи інших фінансових показників. Регресійний аналіз заснований на статистичних методах, таких як метод найменших квадратів, і використовується в різних варіаціях, включаючи лінійну регресію, логістичну регресію та нелінійні моделі. Цей метод є потужним інструментом для прогнозування на фінансових ринках, але його ефективність залежить від якості даних, вибору правильних факторів і правильного підбору моделі.

Розглянемо лінійну модель. У такій моделі шукана змінна має дорівнювати 1, або 0. Наприклад такою змінною може бути відповідь на питання «буде зріст економіки, чи не буде?», або «видавати даному клієнтові кредит, чи ні». Математично лінійна модель представляється у вигляді правила прийняття рішень:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon,$$

де у – залежна змінна; x_k – незалежні змінні, β_k – вагові коефіцієнти, присвоєні незалежним змінним, ε – довільна похибка, розподіл якої залежить від значень екзогенних змінних, але математичне сподівання має бути нульовим.

2.4.2 Логіт та пробіт моделі

Логіт-моделі є формою статистичної моделі, яка використовується для прогнозування ймовірності виникнення події. Логіт-моделі також називають моделями логістичної регресії. Логіт-модель базується на логістичній функції

(також відомій як сигмоїдна функція), яка використовується для моделювання ситуацій, де є два / бінарні можливі результати або категоріальні результати. Логістична функція може бути використана для моделювання різноманітних ситуацій, включаючи бінарні залежні змінні, дихотомічні залежні змінні та категоріальні дані. Логіт-функція використовується для моделювання зв'язку між прогностичними змінними та ймовірністю виникнення події і надає вихідний результат на неперервній шкалі, яка варіюється від 0 до 1. На рисунку 2.3 зображено криву, зробрену логіт-функцією.

Логіт-функція записується таким чином:

$$logit(I) = log[P/(1-P)] = Z = b_0 + b_1X_1 + \cdots .. + b_kX_k$$

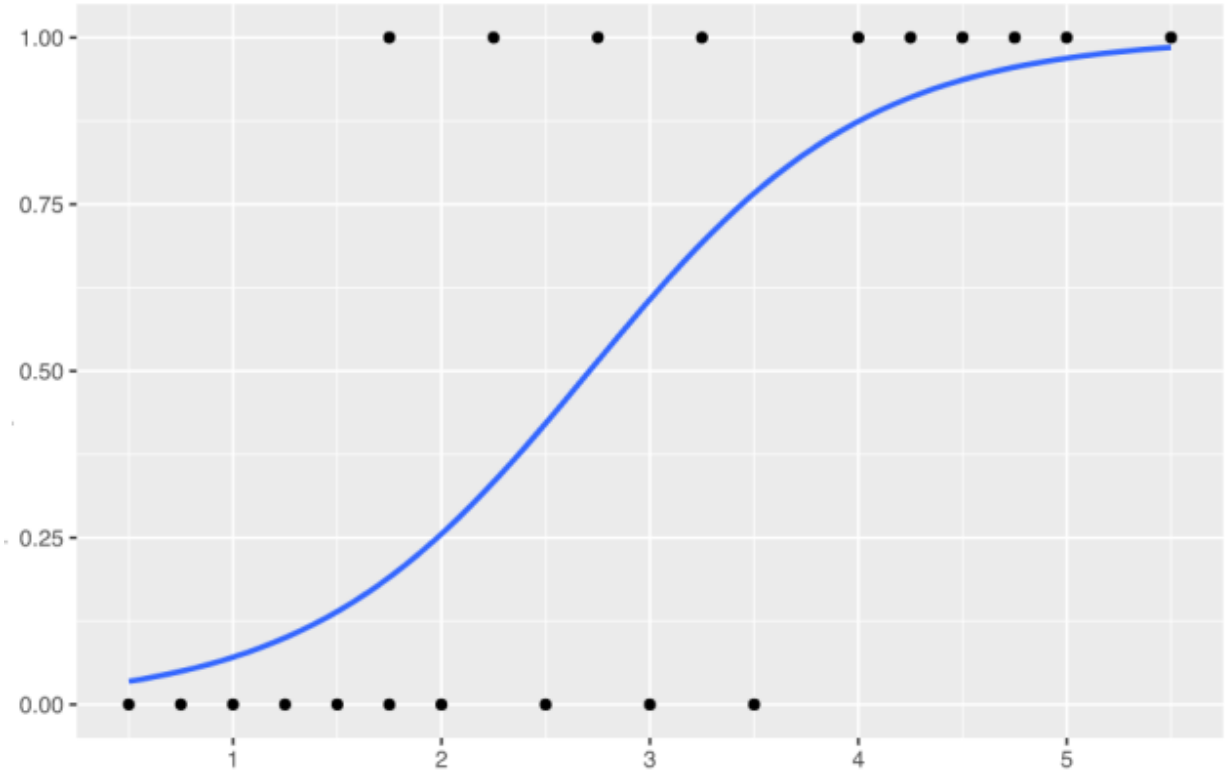

Рисунок 2.3 Вигляд кривої, зробленої логіт-функцією

Пробіт-моделі є формою статистичної моделі, яка використовується для прогнозування ймовірності виникнення події. Пробіт-моделі схожі на логіт-моделі, але вони базуються на пробіт-функції замість логістичної функції. Пробіт-функція також називається функцією пробіт-зв'язку. У пробіт-моделі використовується кумулятивна стандартна нормальна функція розподілу Φ(·)

для моделювання зв'язку між прогностичними змінними та ймовірністю виникнення події. Вихідні дані пробіт-моделі також варіюються від 0 до 1, так само як і в логіт-моделі.

Пробіт функція може бути репрезентована в наступному вигляді:

$$Pr(Y = 1|X) = \Phi(Z),$$

де $Z = b_0 + b_1X_1 + \cdots .. + b_kX_k$.

Тут Y є залежною змінною і представляє ймовірність виникнення події (тобто Y = 1) при заданих змінних X. Z є лінійною комбінацією незалежних змінних (X) з коефіцієнтами (b0, b1, b2... bn). У випадку логіт-моделі використовується логістична або сигмоїдна функція замість Ф, яка є кумулятивною функцією розподілу стандартного нормального розподілу. Параметри (такі як b0, b1 і т. д.) оцінюються за допомогою методу максимальної правдоподібності [10].

2.4.3 Метод на основі нечіткої логіки

Метод на основі нечіткої логіки є одним з підходів до інтелектуального аналізу даних, що дозволяє моделювати та аналізувати нечіткі, неоднозначні або неясні дані. Він використовує нечіткі множини та правила нечіткої логіки для представлення та обробки цих даних.

Нечітка логіка дозволяє враховувати ступінь належності об'єктів до певних категорій чи описувати їх відносною нечіткістю. Вона використовує поняття нечітких множин, де елементи можуть належати множині з різною ступенем належності. Таким чином, метод на основі нечіткої логіки дозволяє враховувати неоднозначність та невизначеність даних, що часто зустрічаються в реальних задачах.

При застосуванні методу на основі нечіткої логіки для інтелектуального аналізу даних, спочатку використовуються експертні знання та досвід для

формулювання нечітких правил та визначення нечітких множин. Потім ці правила та множини використовуються для класифікації, прогнозування або прийняття рішень на основі нових даних.

2.4.4 Модель випадкового лісу

Велика частина машинного навчання зосереджена на класифікації - визначення до якого класу (групи) належить спостереження. Здатність точно класифікувати спостереження має велику цінність для різних бізнес-застосувань, таких як прогнозування того, чи купить певний користувач продукт, або передбачення того, чи буде заданий кредит невиконаним чи ні.

Випадковий ліс, як і його назва вказує, складається з великої кількості окремих дерев рішень, які працюють як ансамбль. Кожне окреме дерево випадкового лісу дає прогноз класу, і клас з найбільшою кількістю голосів стає прогнозом нашої моделі.

Основна ідея за випадковим лісом - проста, але потужна: оцінка від натовпу. З точки зору науки про дані, причина, чому модель випадкового лісу працює настільки добре, полягає в наступному:

Велика кількість майже некорельованих моделей (дерев) як комітет буде працювати краще, ніж будь-яка з окремих моделей-компонентів [11]. Приклад наведено на наступному рисунку 2.4.

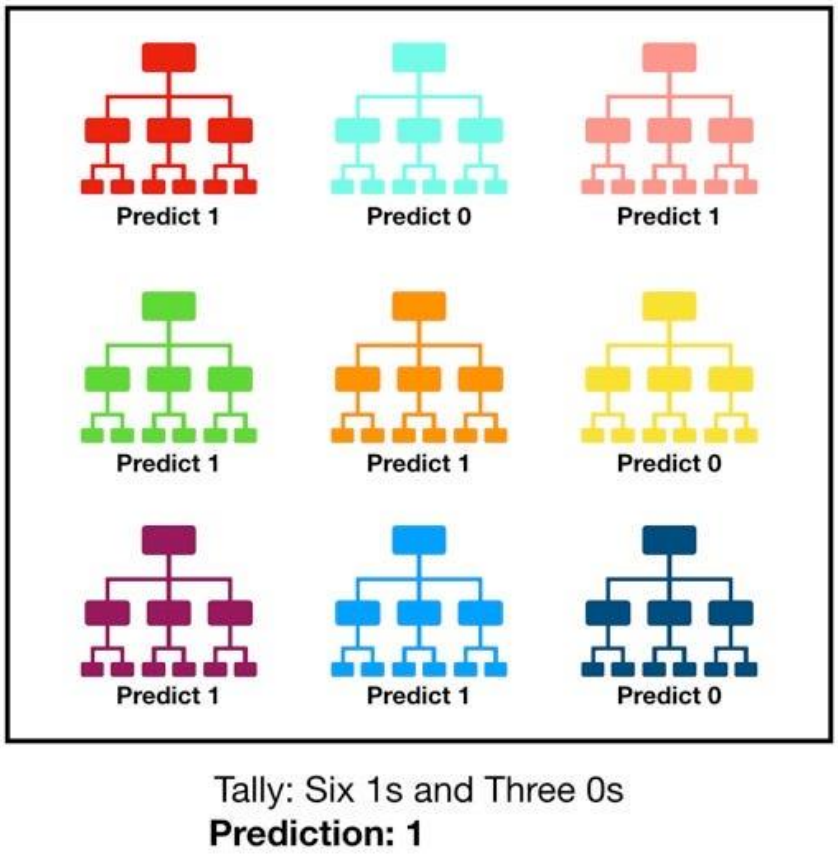

Рисунок 2.4. Візуалізація методу випадкового лісу, що робить передбачення [19]

Низька кореляція між моделями є ключовою. Так само, як і випадкові інвестиції з низькою кореляцією (такі як акції та облігації) поєднуються, утворюючи портфель, який перевищує суму його складових, некорельовані моделі можуть створювати прогнози ансамблю, які є більш точними, ніж будь-які окремі прогнози. Причина цього чудового ефекту полягає в тому, що дерева захищають одне одного від своїх індивідуальних помилок (якщо вони не постійно роблять помилки в одному напрямку). Хоча деякі дерева можуть бути неправильними, багато інших дерев будуть правильними, тому як група дерев вони здатні рухатися в правильному напрямку. Отже, передумови для успішної роботи випадкового лісу такі:

У ознак повинен бути певний справжній сигнал, щоб моделі, побудовані з використанням цих ознак, виконувалися краще, ніж простий випадковий вибір.

Прогнози (і, отже, помилки) індивідуальних дерев повинні мати низьку кореляцію одне з одним.

2.5 Висновки до розділу

У даному розділі було наведено визначення ЧР, його стаціонарності, причини та способи усунення пропущених даних. Було висвітлено тему нормалізації незалежних ознак, а саме необхідність цього етапу і техніки для виконання.

Також, у розділі наявний опис таких математичних моделей та методів для прогнозування фінансових даних:

- регресійний аналіз та регресійні моделі;
- логіт та пробіт моделі;
- методи на основі нечіткої логіки;
- модель випадкового лісу.

РОЗДІЛ 3 ДАНІ ДЛЯ АНАЛІЗУ, ПРОГНОЗУВАННЯ, АНАЛІЗ РЕЗУЛЬТАТІВ

3.1 Вимоги до даних та основи ІАД

В час стрімкого розповсюдження і впровадження найрізноманітніших технологій та систем, світ обростає мільйонами серверів з екзабайтами даних на них. Майже кожна організація, комерційна чи державна, наукова чи розважальна, має підв’язку під систему з різних технологій чи сервісів для ефективного діловодства. Починаючи з обліку працівників і аж до розрахунків орбіт планет, усе потребує зберігання даних. А нашою основною метою в цей момент є пошук будь-якої корисної інформації, що можна з цих даних здобути.

Довгий час, поки кількість даних була невеликою, а потужності обчислювальних приладів ще не вимірювалися терафлопсами(одиниця вимірювання швидкодії обчислювальних приладів), більшість аналітиків користувалися більш традиційними підходами до аналізу даних, такими як алгоритми традиційної математичної статистики, що зазвичай базувалися на операціях з фіктивними величинами(наприклад, середніх значень). Таке дослідження давало лише загальну картину, дуже часто зрізаючи частину зображення, і зараз вважається достатньо «грубим» підходом, що становить основу OLAP [12].

Ми ж маємо справу з ІАД, в основі якого лежить пошук закономірностей, концепція патернів, що наслідують багатоаспектні характеристики досліджуваного об’єкта. Основна ідея пошуку закономірностей це не підтвердження власних гіпотез про дані, а їх спростування, пошук несподіваних результатів, прихованого знання, що може призвести до оптимізації певних процесів, або інших цілей ІАД. Саме це і є причиною необхідності великого обсягу даних, бо стало зрозуміло, що

необроблені дані можуть розповісти багато цікавого і корисного, при грамотному дослідженні.

Як правило, основою Інтелектуального Аналізу Даних є поглиблена обробка статистичних даних, які побудовані для дослідження певних процесів. Можемо виділити такі характеристики подібних досліджень.

1. Набори даних, що представляють певну кількість статистичних показників, зібрані з первинних та вторинних джерел.

Первинними джерелами називають соціальні опитування, збір свідчень, обстеження, які надають дані такого змісту і форми, що потребуються для запланованих прогнозами кількостей. Отримання таких даних є результатом спеціально підготовленої і ретельно спланованої роботи (часто, з виділеними під це коштами та спеціальними засобами). Розроблюється склад показників, метод організації вибірки, її розподіл, а інколи і заздалегідь прописана взаємодія реєстрації одних параметрів від інших [13].

Вихідні дані, що були опубліковані в тому чи іншому виді, а також зібрані кимось, хто не перетинається з конкретною задачею аналітика, але надають інформацію, яка певним чином корисна для розв’язання цієї задачі, називається вторинним джерелом.

2. Вихідні статистичні дані мають слідувати вимогам. При формуванні вибірки даних з первинних чи вторинних джерел, потрібно притримуватися вимог до їх якості.

Відповідність. Дані, що беруться для дослідження певного процесу чи явища мають бути реальним відображенням цих подій і повинні відповідати дійсним подіям, потрібних об’єктів.

Репрезентативність. Щоб забезпечити цю характеристику, вибірка повинна бути структурована таким чином, щоб відображати всі важливі властивості досліджуваної групи чи сукупності. Іншими словами, вибірка

повинна адекватно відображати характеристики цієї групи, від якої вона була взята.

Порівнянність. Інформація, зібрана в дослідженні, повинна супроводжуватися поясненнями, які стосуються змісту показників і методології їх вимірювання. Це дозволить здійснювати порівняння даних в різних часових і просторових контекстах і враховувати зміни в методології та коригування змінних[12].

Надійність і точність. Ця характеристика вимагає забезпечити високу достовірність вихідних даних шляхом використання різних методів для перевірки надійності джерел і дотримання встановленої методології вимірювань. Також важливо контролювати правильність відповідей респондентів та виявляти можливі помилки в записах.

Розрахунок прогнозу та проведення відповідних аналітичних та експериментальних перевірок (верифікація) ймовірнісно-статистичної моделі зазвичай базуються на одночасному використанні інформації двох видів:

- апріорна інформація про характер та сутність аналізованого явища, яка, як правило, представлена теоретичними законами, обмеженнями, або гіпотезами;
- початкові статистичні дані, що відображають процес та результати функціонування вивченого явища чи системи.

Цей підхід передбачає спільне використання інформації обох видів для розробки та перевірки ймовірнісно-статистичної моделі.

Загалом, виділяють такі основні етапи прогнозування:

1. Постановка задачі. Планування процесу аналізу сильно спрощує подальший хід виконання, тому завчасне визначення фінальних цілей, їх пріоритетність, набір змінних, опис існуючих взаємозв’язків, ролі цих змінних та зв’язків.

2. Попередній аналіз даних. Він проводиться перед будуванням моделей і необхідний для побудови апріорних гіпотез та припущень на основі формалізованих знань.
3. Реєстрація значень, що приймають участь в аналізі показників і параметрів на різних просторових або часових тактах роботи модельованої системи.
4. На даному етапі розробки моделі, який базується на прийнятих гіпотезах та початкових припущеннях, виконується формування загальної структури модельних взаємозв'язків між вхідними та вихідними змінними. Мова йде про визначення загального опису моделі, де визначається таємничий сенс величин, які поки що не мають числових значень (їх називають параметрами моделі), і вони вимагають статистичного оцінювання. Тобто на цьому етапі визначається сама структура моделі та формальне вираження взаємозв'язків між змінними.
5. Процес ідентифікації моделі полягає у статистичному аналізі наявних даних з метою налаштування невідомих параметрів на наявні значення. Цей етап передбачає визначення, чи можна взагалі відновити невідомі параметри моделі на основі доступних вихідних статистичних даних для структури моделі, що була визначена на попередньому етапі. Ця задача відома як проблема ідентифікації моделі. Якщо відповідь на це запитання позитивна, наступним кроком є вирішення проблеми параметричної ідентифікації моделі. Іншими словами, потрібно розробити та застосувати математично обґрунтований метод для оцінки невідомих параметрів моделі на основі наявних статистичних даних. Якщо ідентифікація моделі виявляється неможливою, то необхідно повернутися до попереднього етапу і внести необхідні зміни у структуру моделі для вирішення завдання її специфікації.

6. Верифікація моделі полягає в проведенні різних процедур для порівняння результатів, отриманих за допомогою цієї моделі, з реальною дійсністю. Цей етап також називається статистичним аналізом точності та відповідності моделі. У випадку, якщо результати цього етапу є незадовільними, необхідно повернутися до четвертого етапу, а в окремих випадках навіть до першого. Однак, якщо результати цієї перевірки позитивні, то модель може бути використана для оцінювання прогнозу, згідно з загальною схемою, описаною вище.

В першому етапі було згадано про важливість визначення фінальної цілі прогнозування. Мається на увазі ще й визначення необхідного виду прогнозу. Він може бути визначений двома факторами:

- прогнозувальним горизонтом;
- рівнем ієрархії прогнозованого показника.

Горизонт прогнозування може бути:

- короткостроковим (1-2 такти часу вперед)
- середньостроковим (3-5 тактів)
- довгостроковим (понад 5 тактів часу вперед)

Такт часу визначається, зазвичай, періодичністю даних(година, день, тиждень, місяць, рік).

3.2 Характеризація математичної моделі

Математична модель (ММ). Математична модель - це абстрактна математична умова, яка відтворює взаємозв'язки між реальними об'єктами, використовуючи відповідні математичні концепції. Зазвичай ці взаємозв'язки виражаються у вигляді рівнянь і нерівностей між змінними, які описують

функціонування моделі. Створення математичної моделі включає в себе завдання балансу між математичною простотою і достатньою точністю для відтворення важливих аспектів реальної системи, які цікавлять дослідника.

Для опису об’єкта при створенні ММ можна використати безліч математичних інструментів. Серед них: математична логіка, математичне програмування, теорія множин, теорія графів, теорія ймовірностей, диференціальні, комплексні чи інтегральні рівняння тощо. Модель має відповідати наступним умовам.

1. Модель має бути адекватною процесу чи об’єкту.

Це означає, що модель повинна:

- показувати найхарактерніші зв’язки та взаємодію між параметрами процесу;
- брати до уваги імовірні керуючі сигнали;
- враховувати позапроцесні шуми та зовнішні збурення;
- враховувати початкові значення змінних та їх обмеження.

Загалом, за допомогою рядку статистичних величин визначити адекватність[14].

Так як використання лиш однієї величини для визначення адекватності не є переконливим і достатнім підходом, оскільки оцінка параметрів – це випадкова величина, зазвичай використовується декілька критеріїв, що сприяє підвищенню ймовірності вибору адекватнішої моделі.

Наприклад, використовується коефіцієнт кореляції, середньоквадратична помилка, або коефіцієнт детермінації, який обчислюється за формулою:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}},$$

де SS_{res} – сума квадратів лишків регресії, а SS_{tot} – загальна сума квадратів.

Суми квадратів розраховуються за наступними формулами:

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2,$$

де $y_i, \hat{y}_i$– фактичне і розрахункове значення досліджуваної змінної, $\bar{y}_i$ – середнє арифметичне значення досліджуваної функції.

2. Кожному рівнянню моделі має бути хоча б один відповідний аналітичний розв'язок, або, коли це не є можливим, то чисельний розв'язок. Він суттєво необхідний для проведення обчислень оцінок прогнозів та аналізу поведінки процесу (збіжність).

Перш за все, при побудові моделі варто зважати на такий принцип: модель не має містити щось зайве окрім необхідного. Логічно, що дотримуватися такого принципу досить нелегко і буває таке, шо модель має бути надзвичайно розвиненою і мати складну архітектуру для досягнення високого ступеня її адекватності процесу. Це частий випадок для нелінійних процесів. Але у випадку лінійних моделей, такі як авторегресії (АР) чи авторегресії з ковзнім середнім (АРКС), достатньо створити модель, що мала б статистичні характеристики ідентичні до характеристик часового ряду, на основі якого вона оцінюється. Інколи, такі прості моделі можуть повністю задовольнити заздалегідь поставлені цілі з прогнозування та прийняття рішень. Можна сказати, що кожен індивідуальний випадок обрання складності моделі варто вирішувати окремо.

3. Варто попіклуватися про універсальність моделі, аби її застосування не обмежувалося одним процесом, або роботою лише за певних умов.

Наприклад, при описі моторної функції людини (реакція на зовнішні сигнали) використовують звичайне диференціальне рівняння другого порядку, що відображене у формі функції передачі того ж порядку:

$$W(s) = \frac{Ke^{-\tau s}}{(1 - T_1 s)(1 - T_2 s)},$$

де К – сталий коефіцієнт передачі об'єкта, τ – час затримки входу (у випадку з людиною, це зазвичай 300-350 мс), T_1, T_2 - постійні часу [5,6].

Функція такого виду може бути використана, наприклад, для описання реакції людини на світлові чи звукові сигнали, що проходять через систему візуального сприйняття чи аудіосистему (як приклад може слугувати водіння транспортного засобу, де потрібно знати час реакції). Значення параметрів можуть відрізнятися від людини до людини, проте структура моделі лишатиметься незмінною. Тут можна сказати, що оглянута модель відповідає умові універсальності і описує обширний клас біологічних систем.

Моделювання технічних систем характеризується частим використанням ланки першого і другого порядку, що відповідають звичним диф. рівнянням тих же порядків. Це дає змогу на основі базових ланок побудувати вже більш складні моделі. Наприклад, дифузія домішок в атмосфері та водному середовищі, механічні коливання запчастин літака, вітряка чи автомобіля з причепом і багато подібних. Динаміку і рух таких об'єктів та систем переважно описують за допомогою диференціальних рівнянь з частинними похідними.

4. Вимога моделі до міцності, робастності. Це означає, що модель має надавати достовірні результати не тільки на тому часовому відрізку, на якому вона була побудована, а й на будь-якому іншому, що має схожий режим роботи. Іншими словами, це можна назвати стійкістю моделі по відношенню до помилок, збурень та пропущених значень. Цей критерій особливо важливий для таких систем, що працюють в реальному часі оскільки може причинити аварійну ситуацію при виході з ладу.

5. Адаптивність. Ця вимога означає, що щонайменше один параметр мусить мати можливість уточнювати по мірі надходження нових даних від досліджуваного об’єкта. Цей критерій є обов’язковим для побудови моделей, параметри яких є функціями часу, тобто, нестаціонарних систем. Системи керування таких нестаціонарних процесів називаються адаптивними. Вони є досить непростими для аналізу збіжності оцінок параметрів та помилок керування і через це, під час проведення проектних робіт адаптивних систем потрібно приділяти особливу увагу питанням достатнього збудження процесу, вибору методів оцінювання параметрів, вагів, оцінці прогнозів та значень керуючих впливів [15].

3.3 Аналіз якості моделей

Для оцінки адекватності регресійних моделей використовується аналіз послідовності залишків, тобто помилок моделі. Цей аналіз включає обчислення залишків шляхом підстановки фактичних значень усіх факторів, що враховані в моделі. Залишкова послідовність перевіряється на відповідність характеристикам випадкової складової в економічному часовому ряді, таким як:

1) близькість до нуля математичного очікування, що означає, що середнє значення залишків дорівнює нулю;
2) випадковий характер відхилень, що показує, що залишки не мають системного або систематичного відхилення від нуля;
3) відсутність автокореляції, що означає, що залишки не залежать від самих себе в попередніх часових точках;
4) нормальний закон розподілу, що вказує на те, що залишки розподілені за нормальним законом розподілу.

Параметри адекватності, наведені нижче допоможуть провести попередній, а інколи і ґрунтовний, аналіз адекватності моделі.

1. Сума квадратів залишків – це сума квадратів величин розбіжності між змодельованими і фактичними значеннями пояснюючої змінної на період ідентифікації, формула була наведена вище.

2. Коефіцієнт детермінації – статистичний параметр, який використовують в статистичних моделях, як міра інформативності моделі стосовно даних. Це можна описати як те, на скільки існуючі спостереження відповідають побудованій моделі. Формула наведена вище.

3. Критерій Дарбіна-Уотсона – це статистичний критерій, що необхідний для перевірки автокореляції першого порядку елементів досліджуваної послідовності. Найчастіше використовується під час аналізу часових рядів і залишків моделей регресії:

   a. $d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} {e_t}^2} \approx 2(1 - p_1)$,

де p_1 – коефіцієнт автокореляії першого порядку. d = 2 при відсутності автокореляції, при негативній дорівнює 4 і в інших випадках прямує до нуля:

$$\begin{cases} p_1 = 0 \rightarrow d = 2 \\ p_1 = 1 \rightarrow d = 0 \\ p_1 = -1 \rightarrow d = 4 \end{cases}$$

4. Інформаційний критерій Акайке (ІКА, Akaike Information Criterion, AIC) – це критерій відносної якості статистичних даних для наданій вибірці даних. ІКА проводить оцінку кожної з моделей відносно кожної з інших моделей з сукупності моделей для наявних даних. ІКА виокремлюється своєю ґрунтованістю в теорії інформації. Він служить для визначення, наскільки добре обрана модель адекватно відтворює процес, який створив набір даних. З іншими словами, ІКА надає важливий показник,

який розкриває баланс між точністю моделі та її складністю. Припустимо, що існує деяка статистична модель для деякого набору даних. Максимальне значення функції правдоподібності для цієї моделі нехай буде L, і нехай k буде кількістю оцінюваних параметрів. Тоді значення ІКА обчислюватиметься так:

a. $AIC = 2k - 2ln\,(L)$

5. Також, для оцінки адекватності моделі використовують критерій Байєса-Шварца.

3.4 Аналіз якості прогнозу

Процес прогнозування включає в себе важливий аспект - об'єктивне визначення, наскільки надійний отриманий прогноз. Оскільки передбачувані значення представляють собою випадкові величини, для оцінки їх точності необхідно використовувати різні статистичні критерії. Графік на рисунку 3.1 ілюструє часовий аспект та інтервали часу, на яких проводиться оцінка моделі та перевірка точності прогнозу.

Розумна практика полягає в тому, щоб розділити наявну вибірку даних на дві частини: навчальну та перевірочну. У навчальній частині проводиться аналіз і оцінка параметрів моделі процесу, і також здійснюється "історичний" прогноз, що допомагає оцінити якість однокрокового прогнозу для цього сегмента даних. Прогноз на перевірочній частині вибірки даних іноді називається "прогнозом ex post". В різних наукових дослідженнях рекомендується виділити для перевірки від 5% до 40% значень у вибірці даних. Однак, при аналізі коротких часових рядів, доцільно використовувати значно більшу частину ряду для оцінювання параметрів моделі.

Прогнозування значень поза межами вибірки даних відоме як "прогноз ex ante" (рисунок 3.1).

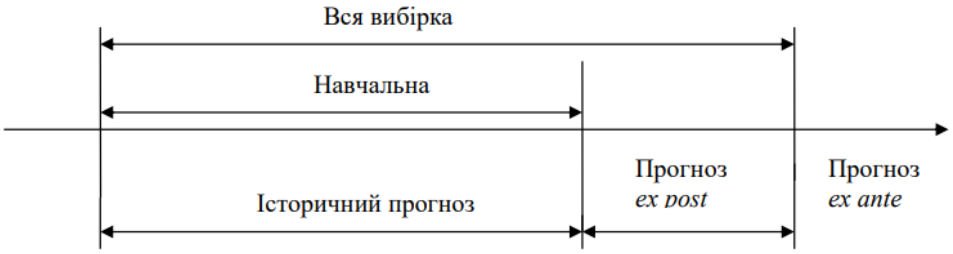

Рисунок 3.1 Види прогнозування відносно часового ряду

Зазвичай, для оцінювання якості прогнозів використовують різноманітні статистичні критерії. Наприклад, середньоквадратична похибка, яка зазвичай використовується, має свої обмеження, оскільки залежить від масштабу даних. Тому необхідно використовувати різні статистичні показники для аналізу якості прогнозу.

На останньому етапі, важливо забезпечити точність і обґрунтованість прогнозів, використовуючи різноманітні критерії, підходи і процедури для оцінки їхньої якості.

Є такі показники якості прогнозу:

- MSE (Mean Squared Error). Середньоквадратичне відхилення - це середньоквадратична похибка, статистичний показник, який вимірює середньоквадратичну відстань між прогнозованими значеннями і спостережуваними даними

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n},$$

- RMSE (Root Mean Squared Error) - це квадратний корінь середньоквадратичної похибки, використовується для вимірювання середньоквадратичної відстані між прогнозованими і фактичними значеннями, і він виражається в одиницях вихідних даних.

- Абсолютна помилка прогнозу теж може бути корисною і визначається як різниця фактичного(y) і прогнозованого значення(ŷ):
  $\Delta = y_t - \hat{y}$Середнє абсолютне значення помилки дорівнюватиме:
  $$\Delta_{\text{сер}} = \frac{\sum_{t=1}^{n} y_t - \hat{y}_t \vee}{n},$$

- Коефіцієнт Тейла - це статистичний показник, який вказує на ступінь зв'язку між змінними. Він вимірює кореляцію між двома змінними і приймає значення в межах від -1 до 1. Значення близьке до -1 вказує на негативний зв'язок, близьке до 1 - на позитивний, а значення близьке до 0 - на відсутність зв'язку.
  $$U = \frac{\sqrt{\frac{1}{N}\sum_{k=1}^{N}[y(k) - \hat{y}(k)]^2}}{\sqrt{\frac{1}{N}\sum_{k=1}^{N} y^2(k)} + \sqrt{\frac{1}{N}\sum_{k=1}^{N} \hat{y}^2(k)}}$$

- SAP (Standardized Abnormal Performance, стандартна абсолютна похибка в процентах, САПП) - це стандартизований показник, який використовується для оцінки відхилення результатів від очікуваних значень у фінансовій аналітиці. SAP вимірює, наскільки фінансові результати перевищують або залишаються поза очікуваннями у стандартизованих одиницях.

Оскільки дана міра відображає відносну якість прогнозу, то її використовують головним чином для порівняння точності прогнозів різних об'єктів або процесів. Проте вона завжди корисна при проведенні порівняльного аналізу якості прогнозування одного і того ж процесу за допомогою різних методів. Це так, оскільки відносна міра є чіткою і

зрозумілою як для дослідника, так і для практичних користувачів. У таблиці 3.1 наведені стандартні значення САПП та їх можлива інтерпретація.

Таблиця 3.1 Інтерпретація типових значень критерію САПП

САПП, %	Інтерпретація
<10	Висока точність
10-20	Хороша точність
20-50	Задовільна точність
>50	Погана точність (незадовільна)

- Середня похибка – не є відносним показником, бо вона описує характер зміщення прогнозованих значень від реальних і розраховується за формулою:

$$\text{СП} = \frac{1}{N}\sum_{k=1}^{N}[y(k) - \hat{y}(k)] = \frac{1}{N}\sum_{k=1}^{N} e(k)$$

Або

$$\text{СП} = \frac{1}{S}\sum_{i=1}^{S}[y(k+s) - \hat{y}(k+s,k)]$$

СП буде зменшуватися, якщо похибки мають різні знаки, що очевидно. Середня похибка в процентах обчислюється за виразом:

$$\text{СПП} = \frac{1}{N}\sum_{k=1}^{N}\frac{[y(k) - \hat{y}(k)]}{y(k)} \times 100\%$$

Середньоквадратична помилка (СКП) також вказує на те, яким чином прогноз відхиляється від реальної ситуації, зокрема, чи тенденція завищення прогнозу балансується з тенденцією його заниження. Ідеальний прогноз вважається незміщеним, коли втрати, пов'язані з переоцінкою та недооцінкою фактичних значень у майбутньому, компенсують одна одну, і в цьому випадку

значення СП та СКП прагнуть до нуля. Зрозуміло, що досягнення абсолютної нульової помилки є ідеалом, але на практиці це надзвичайно складно. Дослідження показують, що прийнятні значення для СКП (так само, як і для САПП) розташовуються в певному діапазоні.

- Максимальна Абсолютна Похибка (МАП), її можна визначити таким чином:

$$\text{МАП} = max_k\{|y(k) - \hat{y}(k)|\}, 1 \leq k \leq N,$$

- Мінімальна абсолютна похибка (МіАП) визначається як:

$$\text{МіАП} = min_k\{|y(k) - \hat{y}(k)|\}, 1 \leq k \leq N,$$

Критерії МАП та МіАП можуть також бути корисні при порівняльному аналізі кількох методів прогнозування, особливо тоді, коли нас цікавить максимальне або мінімальне можливе відхилення прогнозів від фактичних значень на конкретному інтервалі часу.

3.5 Аналіз результатів

Поняття «фінансові дані» є дуже різноплановим, бо може включати безліч інформації про фінансовий стан, діяльність і результати підприємств, організацій, чи індивідуальних осіб. Вони описують найрізноманітніші аспекти фінансів і господарської діяльності. Серед них можна виділити такі:

– податкова інформація: дані про податки, сплачені фізичними та юридичними особами;
– кредитування: кредитна історія, історія боргів та інших зобов’язань;
– бюджети, тобто заплановані доходи та витрати на певний майбутній період часу;
– фінансова звітність, що представляють фінансову інформацію про баланс, звіти про прибутки, звіти про готові продукти та інші;

- фінансові показники, які обчислюються на основі фінансових звітів і слугують для подальшого аналізу, наприклад, рентабельність, ліквідність, маржу, ефективність управління запасами;
- фінансові ринки: предметом їх опису є інформація про стан акцій, облігацій, валют та інших фінансових інструментів на ринках.

Саме дані з фінансових ринків будуть використовуватися для розробки і дослідження моделей Інтелектуального Аналізу Даних.

Так як передбачення фінансових ринків завжди була актуальною та привабливою ідеєю для багатьох, зараз можемо спостерігати безліч моделей та методів для такого моделювання. Серед найбільш популярних варто відзначити статистичні моделі аналізу часових рядів, рекурентні та нейронні мережі, текстовий аналіз (для відстеження трендів в новинах, соціальних мережах), генетичні алгоритми, дерева рішень, тощо.

Процес аналізу фінансових даних є складною операцією, що включає в себе кілька послідовних етапів для оцінки та розуміння ринкової ситуації. Кожен окремий випадок має розглядатися окремо. Узагальнюючи, можна навести такий приблизний план.

1. Перш за все потрібно здійснити підготовку даних.
    1) пошук та збір фінансових даних;
    2) обробка та підготовка для подальшого аналізу.
2. Візуалізація як перший крок в усвідомленні базових трендів сильно допоможе в наступних етапах.
    1) дослідження графіків, діаграм, що побудовані на виокремлених параметрах;
    2) визначення трендів, волатильності, сезонності та інших явно виражених характеристик.

3. Технічний аналіз. Використання технічних індикаторів, таких як рухоме середнє, стандартне відхилення, осцилятори та інші нестандартні допоміжні інструменти.
4. Прогнозування та моделювання. Використання вже ІАД для створення прогнозів щодо майбутньої ринкової поведінки
5. Моніторинг та аналіз результатів. Відстеження власних прогнозів та перевірка аналізу з отриманням вже реальних результатів

Варто відзначити, що існують і інші етапи для аналізу фінансових ринків, які базуються на внутрішніх та зовнішніх факторах впливу. Наприклад, перед купівлею чи продажем суттєвої кількості акцій потрібно розраховувати не тільки на статистичні оцінки, а й на поведінкові, потрібно розрахувати реакцію ринку на такі події і діяти відповідно. До того ж, не варто забувати, що в разі суттєвих політичних змін котирування тієї чи іншої компанії може змінитися кардинально.

Для аналізу було обрано вибірку щоденного курсу з 2010-01-01 до 2023-10-24 акцій успішної компанії, що займається виробленням електронних пристроїв та програмного забезпечення. Маємо 3475 записів про ці дні з такими ознаками:

- Open – ціна на початку торгового періоду;
- Close – ціна в кінці торгового періоду;
- High – найвища ціна, що була досягнута за даний період;
- Low – найнижча ціна, що була досягнута за даний період;
- Volume – кількість угод, укладених протягом торгового періоду.

Нижче наведена таблиця 3.2 з прикладом даних, використано перші та останні 5 записів.

Таблиця 3.2. Початкові дані. Перші та останні 5 записів

	Open	**High**	**Low**	**Volume**	**Close**
0	7.622500	7.660714	7.585000	4,94E+08	6.487534
1	7.664286	7.699643	7.616071	6,02E+08	6.498749
2	7.656429	7.686786	7.526786	5,52E+08	6.395379
3	7.562500	7.571429	7.466071	4,77E+08	6.383556
4	7.510714	7.571429	7.466429	4,48E+08	6.425996
...	...	...	...	...	...
3470	176.649994	178.419998	174.800003	57549400	177.149994
3471	175.580002	177.580002	175.110001	54764400	175.839996
3472	176.039993	177.839996	175.190002	59302900	175.460007
3473	175.309998	175.419998	172.639999	64189300	172.880005
3474	170.910004	174.009995	169.929993	55934300	173.000000

В даному випадку, дані було зібрано без жодних пропусків та в повному обсязі, тому можна переходити до кроку візуалізацій.

Для кращої наочності візьмемо останні 1500 записів з нашої вибірки.

На Рисунку 3.2 зображено вартість акції наприкінці торгового періоду. Такий графік дає попередні уявлення про успіхи компанії, є можливість відстежити вплив пандемій, війн, економічних криз і те, як швидко після цього оговтувалася компанія.

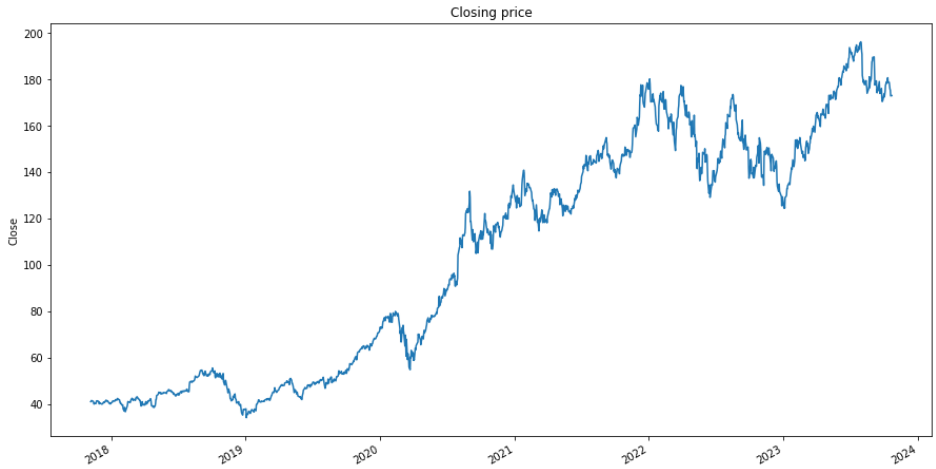

Рисунок 3.2 Close параметр відносно часу

На Рисунку 3.3 зображено ковзне середнє за 10, 20 та 50 днів і порівнюється з реальними даними протягом 500 днів. Даний метод дає змогу згладити піки даних шляхом усереднення значень. Неважко побачити, що ковзне середнє на 10 та 20 днів краще справилися за ковзне середнє на 50 днів, бо досі мають змогу точно відображати тренди, тому добре репрезентують загальне положення справ.

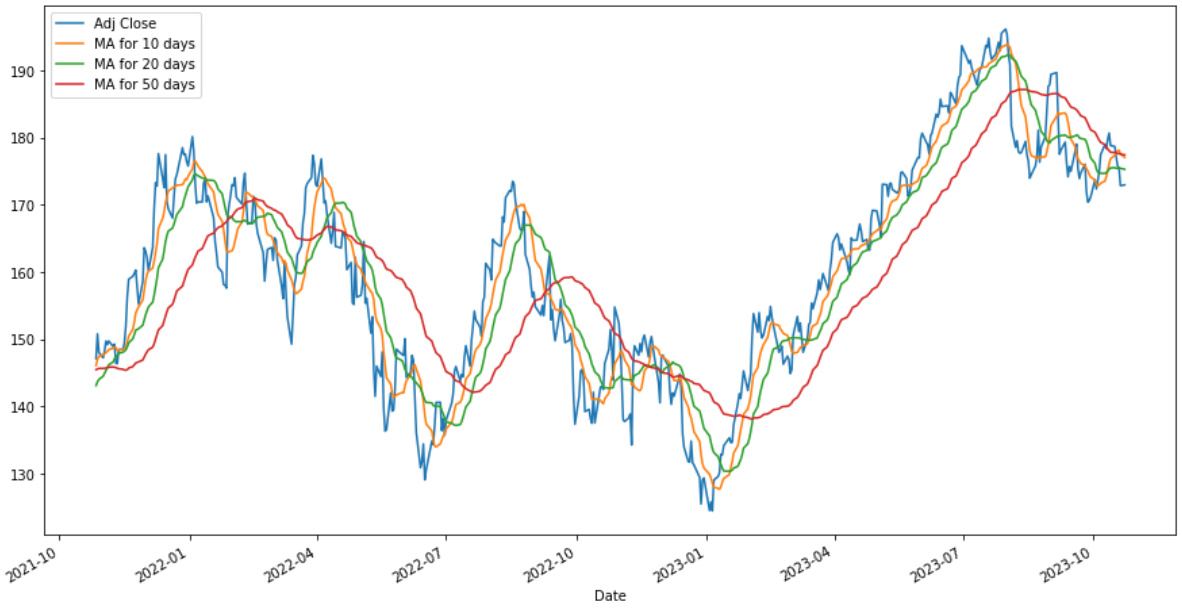

Рисунок 3.3 Ковзне середнє на 10, 20, 50 днів

Нижче можна побачити кореляційну матрицю (рисунок 3.4) для звичайних даних з фондових ринків і її використання в простому вигляді не є доречним. Проте, під час маніпулювання даними і створення нових параметрів та індикаторів на основі вже наявних можна визначити ступінь їх кореляції до інших параметрів, що часто може бути корисним при роботі з іншими типами фінансових даних.

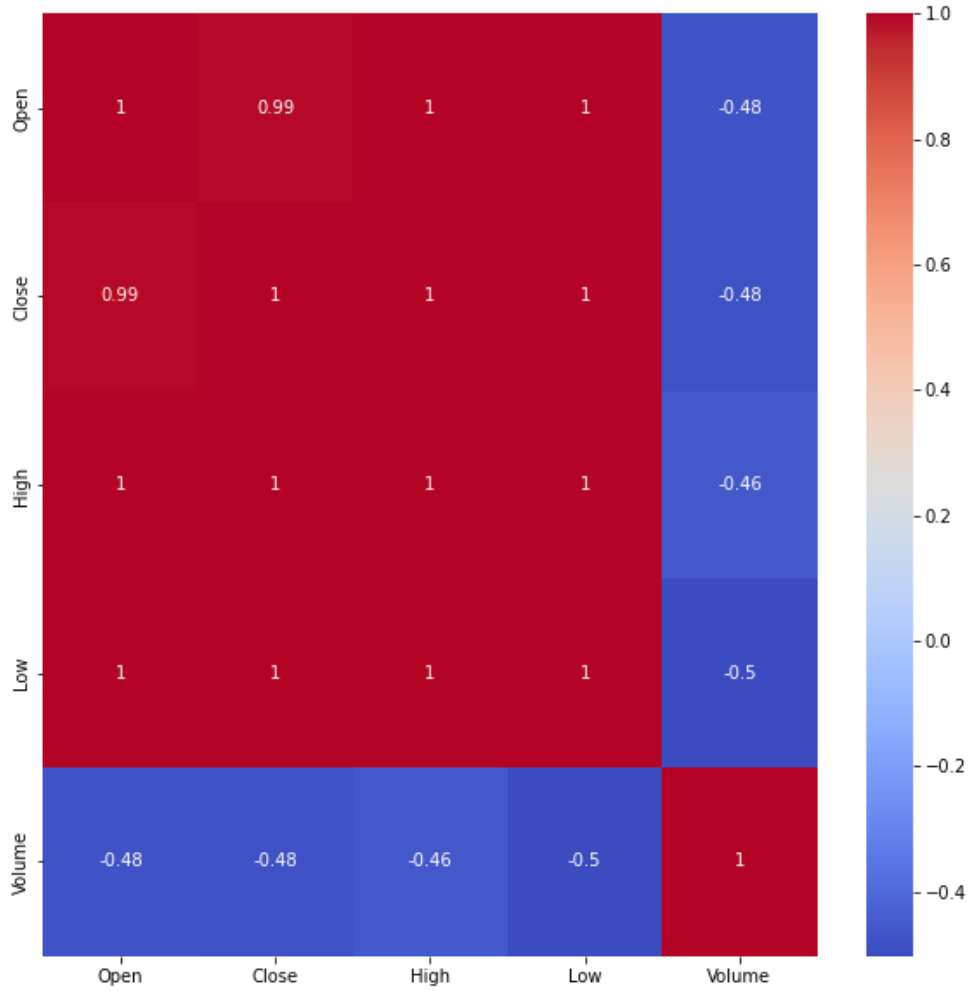

Рисунок 3.4 Кореляційна матриця

Для прогнозування даного набору даних було створено 7 моделей.

1. Vanilla LSTM. Це стандартна реалізація мережі Long Short-Term Memory, що є варіацією рекурентної нейронної мережі і призначена для роботи послідовностями даних, такими як звук, текст, часові ряди. Основна ідея – відтворення LSTM без модифікацій або розширень. Має два шари, центральний та прихований, які оновлюються і передають інформацію в кожному кроці.
2. Stacked LSTM. Є модифікацією стандартної LSTM моделі, де декілька прихованих шарів розташовані один за одним (або один на одному – звідси і назва Stacked). Вони краще розуміють інгібіторні та зворотні залежності, мають більшу глибину, краще

адаптовуються до складних завдань та підходять для задач, де часові залежності є важливими.

3. Bi-directional LSTM. Чергова варіація LSTM, що дозволяє інформації переміщуватися в обох напрямках вздовж послідовності даних, на відміну від стандартного відтворення. Головна перевага – здатність враховувати контекст як попередньої так і наступної частини даних. Хоч дана модель і потребує більших обчислювальних потужностей, вона є досить універсальною та може використовуватися для різних задач.
4. ARIMA – модель часового ряду, яка використовується для аналізу та прогнозування часових рядів. Має такі характеристики:
   (a) AR – авторегресія, залежність поточного значення часового ряду від попередніх.
   (b) I – інтегрування, процедура диференціації, що призначена для перетворення ряду до стаціонарного, тобто позбавленого тренду та сезонності.
   (c) MA – ковзне середнє, середнє значення змінних за певний період часу
5. SARIMA – модифікація ARIMA з додаванням компоненту сезонності. Така модель призначена для моделювання з вираженими сезонними паттернами.
6. Decision tree, дерево рішень, один з найбільш популярних алгоритмів Інтелектуального Аналізу Даних для вирішення проблем регресії та класифікації. Має структуру дерева, де внутрішні вузли представлені прийняттям рішень на основі конкретної ознаки, а кожен листок представляє прогнозоване значення. Ця модель легко інтерпритується та візуалізується для розуміння, як модель приймає рішення, корисна в визначенні ознак, проте попри все, може бути легко перенавченою

7. Random Forest – це ансамбльний алгоритм машинного навчання, що базується на деревах рішень. Така модель робить багато прогнозів для кожного прикладу даних і, в результаті, дані об’єднуються для отримання точного, зазвичай, результату. Однією з переваг можна назвати можливість працювати з великою кількістю ознак і даних. Random Forest є популярним через високу точність, стійкість до перевантаження та можливість працювати з різними типами даних.

Для оцінки моделей було обрано коефіцієнт детермінації (R^2), середньо квадратичне відхилення (MSE), квадратний корінь середньо квадратичного відхилення (RMSE), середнє абсолютне значення похибки (MAE) та середнє абсолютне значення похибки у відсотках (MAPE) (таблиця 3.3).

Таблиця 3.3 Результати якостей моделей

	R2	MSE	RMSE	MAE	MAPE
Vanilla LSTM	0.828	60.889	7.803	6.041	0.037
Stacked LSTM	0.889	39.221	6.263	5.454	0.035
Bi-directional LSTM	0.968	11.324	3.365	2.627	0.017
ARIMA	-2.507	0.053	0.229	0.197	0.039
SARIMA	-1.434	0.037	0.191	0.152	0.03
Decision Tree	-1.158	763.652	27.634	21.764	0.131
Random Forest	-1.131	754.069	27.46	21.572	0.13

Найкращі результати показали найбільш складні та продумані моделі для аналізу часових рядів, а саме найбільш досконала версія LSTM – Bi-directional LSTM та покращена модель ARIMA – SARIMA.

Для порівняння візьмемо дані іншої компанії, що має суттєвий обвал акцій на початку 2022 року після різкого зростання. Надалі вартість то

зростала, то спадала. Нижче на малюнку 3.5 наведений графік вартості акції наприкінці торгівельного періоду.

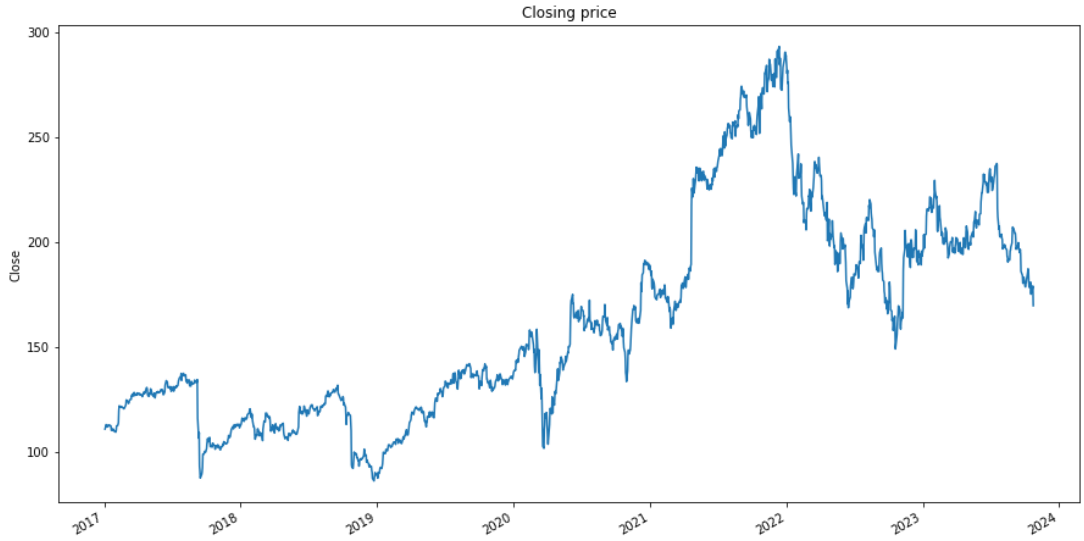

Рисунок 3.5 Вартість акцій компанії_2

Тепер побудуємо аналогічні моделі та поспостерігаємо за їх метриками (таблиця 3.4).

Таблиця 3.4 Результат якості моделей компанії_2

	R2	MSE	RMSE	MAE	MAPE
Vanilla LSTM	0.926	22.065	4.697	3.689	0.019
Stacked LSTM	0.929	21.225	4.607	3.572	0.018
Bi-directional LSTM	0.939	18.154	4.261	3.167	0.016
ARIMA	-4.131	0.041	0.203	0.18	0.034
SARIMA	-2.704	0.03	0.172	0.154	0.029
Decision Tree	0.962	11.646	3.413	2.634	0.014
Random Forest	0.972	8.477	2.912	2.311	0.012

Для моделей LSTM ситуація змінилась і середня похибка зросла значуще. Натомість, моделі, засновані на деревах, проявили себе достатньо добре, особливо Random Forest.

3.6 Висновки до розділу

Фінансові дані, попри загальну тенденцію до подібних поведінок, багато чим відрізняються поміж собою. В розглянутому випадку був приклад світової корпорації яка добре регулює стан своїх акцій і уміє справлятися з непередбачуваними ситуаціями, форс-мажорами державного рівня. Враховуючи постійний зріст та зрозумілі тренди, аналіз та дослідження цих фінансових даних було абсолютно безперешкодним та прогнозованим. В таких умовах досить легко обрати стратегію для подальших капіталовкладень, базуючись на простих і логічних методиках, які були наведені вище.

В іншому розглянутому випадкові є компанія з незрозумілими, на перший погляд, коливаннями вартості акцій. Більш того, модель Bi-directional LSTM, що чудово проявила себе в першому випадку, тут показала достатньо посередній результат.

Таким чином, можна дійти до деяких висновків. При розробці моделей для Інтелектуального Аналізу Даних потрібно завжди брати до уваги декілька варіантів для його здійснення, так як зосередження на якомусь одному може мати серйозні негативні наслідки у вигляді поганого прогнозу чи оцінці ситуації. Компетентний та структурований підхід має неабияку роль для досягнення високого результату. Хотілося б додати, що лише в стабільних умовах за відсутності зовнішніх чинників моделювання дає точні результати завжди. Для реального світу з сучасними проблемами завжди бажано надавати перевагу комплексу рішень.

Результати даного розділу були апробовані на конференції [16,17].

РОЗДІЛ 4 РОЗРОБКА ВЛАСНОГО СТАРТАП ПРОЕКТУ

4.1 Опис ідеї

Зміст ідеї полягає в створенні певного механізму, моделі чи то системного продукту, що за допомоги сучасних методів аналізу та прогнозування, як то авторегресійні методи, нейронні чи рекурентні мережі, буде прогнозувати різноманітні фінансові процеси.

Така структура буде корисною першочергово для фінансових інституцій, як то банки, фонди, що беруть за основу постійну роботу з фінансовими даними. Також можливе застосування для державного апарату при побудові довгострокових перспектив покращення життя співгромадян та розвитку економіки.

Користувачі даного продукту матимуть доволі складний інструмент, що міг би виконувати значну частину роботи з аналізу та внутрішніх процесів автоматизовано, без залучення людського ресурсу, таким чином виключаючи так званий «людський фактор».

Варто провести аналіз потенційних техніко-економічних переваг подібного продукту, а саме:

- визначити техніко-економічні властивості та характеристики даної ідеї (таблиця 4.1);
- визначити список конкуруючих продуктів, провести аналіз їх можливостей та діяльності, порівняти з власними ідеями.

Таблиця 4.1 Аналіз сильних, слабких та нейтральних характеристик ідеї

Техніко-економічні хар-ки ідеї	Продукти конкурентів				W	N	S
	Проект	К.1	К.2	К.3	Слабка сторона	Нейтральна сторона	Сильна сторона
Форма продукту	Програма	Програма	Програма	Програма		+	
Собівартість	Висока	Висока	Висока	Висока		+	
Наявність адміністратора	Не треба	Треба	Треба	Не треба			+
Наявність інтернету	Треба	Треба	Треба	Треба		+	
Кросплатформеність	Так	Ні	Ні	Ні			+

В межах даного розподілу варто провести аудит технології, яка допоможе реалізувати задуманий проект. Серед потрібних характеристик:

- будування моделей для аналізу, технології існують;
- будування моделей для прогнозування, технології існують.

Враховуючи дані характеристики для реалізації задуманої логіки проекту, можна скористуватися мовою програмування Python чи R.

4.2 Дослідження ринкових можливостей запуску стартап-проекту

Для ефективного планування впровадження проекту на ринку необхідно здійснити аналіз ринкових можливостей та загроз. Цей аналіз допоможе визначити потенційні шляхи розвитку проекту та врахувати фактори, які можуть вплинути на його успішність.

Почнемо з аналізу попиту на продукт чи послугу, яку надає стартап-проект. Оцінимо наявність попиту на ринку, обсяг цього попиту, а також динаміку його розвитку (таблиця 4.2). Важливо визначити, чи існує попит на конкретну продукцію чи послугу, яку плануємо надавати.

Таблиця 4.2. Попередня характеристику потенційного ринку

№	Параметр стану ринку	
1	Кількість головних гравців, од	3
2	Загальний обсяг продаж, грн/ум.од	4000
3	Динаміка ринку (якісна оцінка)	Зростає
4	Наявність обмежень для входу (вказати характер обмежень)	Відсутні
5	Специфічні вимоги до стандартизації та сертифікації	Відсутні
6	Середня норма рентабельності в галузі (або по ринку), %	30%

Далі проведемо характеристику потенційних клієнтів, що можуть бути зацікавлені в проекті (таблиця 4.3).

Таблиця 4.3. Характеристика потенційних клієнтів стартап-проєкту

№ п/п	Потреби, що формує ринок	Цільова аудиторія (цільові сегменти ринку)	Відмінності у поведінці різних цільових груп клієнтів	Вимоги споживачів до товару
1	Потреба в програмному забезпеченні для прогнозування фінансових даних підприємства чи державної установи	Державні підприємства, фінансові установи, венчурні фонди, середній та великий бізнес, аудиторські компанії	Різні потреби, обмеження в бюджеті, сфера діяльності	Точність, швидкість, ефективність

Визначимо фактори загроз та можливостей (таблиця 4.4 і таблиця 4.5).

Таблиця 4.4. Фактори загроз

№ п/п	Фактор	Зміст загрози	Можлива реакція компанії
1	Конкуренція	Поява великої корпорації з більшими можливостями за меншу ціну	А) Продаж компанії конкуренту Б) Вихід з ринку В) Перехід на інший ринок
2	Ціна збуту	Висока ціна збуту товару буде перешкодою для багатьох потенційних клієнтів	Передбачення різних тарифів для покриття більшої кількості клієнтів

Таблиця 4.5. Фактори можливостей

№ п/п	Фактор	Зміст можливості	Можлива реакція компанії
1	Технічна підтримка	Постійна робота над вдосконаленням продукту шляхом реакції на відгуки споживачів	Створення тісного каналу комунікації для постійного вдосконалення
2	Забезпечення різних потреб	Більш широкий спектр послуг	Розширення варіативності клієнтських запитів

Беручи до уваги модель 5 сил М. Портера варто розробити перелік факторів конкурентоспроможності для певного ринку (таблиця 4.6).

Таблиця 4.6. Аналіз конкуренції на ринку за М. Портером

Складові аналізу	Прямі конкуренти в галузі	Постачальники	Клієнти	Потенційні конкуренти	Товари-замінники
Висновки	Займають велику частину ринку, мають напрацьовані рішення	Відсутні	Клієнтам потрібен стабільно точний продукт	Можливості для входу на ринок існують через простоту входу, хоч це і займе певний час	Замінники можуть використовувати більш складну схему, що буде готовою до надання кращих результатів

Наступним етапом варто навести перелік факторів конкурентоспроможності (таблиця 4.7).

Таблиця 4.7. Обґрунтування факторів конкурентоспроможності

№ п/п	Фактор конкурентоспроможності	Обґрунтування (наведення чинників, що роблять фактор для порівняння конкурентних проектів значущим)
1	Інновації	Інноваційні рішення можуть стати ключовим фактором, що відрізняє від конкурентів
2	Функціонал	Функціонал має бути достатнім для різних вимог клієнтів
3	Підтримка	Продукт мусить мати зв’язок з усіма клієнтами задля утримання і вдосконалення продукту

Для фінального етапу ринкового аналізу можливостей впровадження проекту і складання SWOT-аналізу (матриці аналізу сильних (Strength) та слабких (Weak) сторін, загроз (Troubles) та можливостей (Opportunities)) потрібно виділити сильні та слабкі сторони за RFS (таблиця 4.8).

Таблиця 4.8. Порівняльний аналіз сильних та слабких сторін RFS

№ п/п	Фактор конкурентоспроможності	Бали 1-20	Рейтинг товарів-конкурентів у порівнянні з нашим підприємством					
			-3	-2	-1	0	+1	+2
1	Інновації	20		+				
2	Функціонал	15					+	
3	Підтримка	10	+					

Аналіз ринкових загроз і ринкових можливостей базується на дослідженні впливу факторів, які можуть впливати на діяльність підприємства в маркетинговому середовищі. Ринкові загрози і можливості виникають як наслідок цього впливу, і, незважаючи на те, що їх реалізація на ринку ще не відбулася, вони мають певний ступінь ймовірності.

Підприємство повинно дбати про аналіз ринкових загроз і можливостей (таблиця 4.9), щоб приймати інформовані рішення та розробляти стратегії, які дозволять використовувати можливості та запобігати загрозам на ринку. Це допомагає зберегти конкурентну перевагу та забезпечити успішну діяльність на ринку.

Таблиця 4.9. SWOT-аналіз стартап-проекту

Сильні сторони:	Слабкі сторони:
Постійна підтримка клієнтів, вдосконалення продукту, інновації	Висока ціна
Можливості:	Загрози:
Розробка ефективного інструменту, оптимізація	Конкуренція, недостатній функціонал

Альтернативи ринкового впровадження проекту розглянуто в наступній таблиці (таблиця 4.10).

Таблиця 4.10. Альтернативи ринкового впровадження стартап-проекту

№ п/п	Альтернатива (орієнтовний комплекс заходів) ринкової поведінки	Ймовірність отримання ресурсів	Строки реалізації
1	Вихід на ринок з якістю, нижче	60%	4 місяці

	очікуваної		

Кінець таблиці 4.10

2	Створення програмного забезпечення на хмарному сервісі	30%	6 місяців
3	Відсутність повного функціоналу	80%	3 місяці

У даному пункті був проведений детальний аналіз ринку та продукту в рамках конкурентоспроможності. Можемо зробити висновок на основі описаних стратегій та ідей, що існують достатньо сприятливі умови для виходу продукту на ринок.

4.3 Розробка ринкової стратегії стартап-проекту

Першочергово необхідно проаналізувати цільову аудиторію проекту(таблиця 4.11).

Таблиця 4.11. Вибір цільових груп потенційних споживачів

№ п/п	Опис профілю цільової групи потенційних клієнтів	Готовність споживачів сприйняти продукт	Орієнтовний попит у межах цільової групи (сегменту)	Інтенсивність конкуренції в сегменті	Простота входу у сегмент
1	Персональні користувачі	Низька	5%	Низька	Середня

2	Великі бізнеси	Висока	30%	Середня	Висока

Кінець таблиці 4.11

3	Малі та середні бізнеси	Середня	15%	Середня	Середня
4	Держава	Середня	10%	Висока	Низька
5	Фінансові установи різного масштабу	Висока	70%	Висока	Середня
Які цільові групи обрано: 3, 5					

Маючи аналіз цільових груп, далі визначимо базову стратегію розвитку продукту (таблиця 4.12).

Таблиця 4.12. Визначення базової стратегії розвитку

№ п/п	Обрана альтернатива розвитку проекту	Стратегія охоплення ринку	Ключові конкурентоспроможні позиції відповідно до обраної альтернативи	Базова стратегія розвитку*
1	3, 5	Ринкове позиціювання	Відсутність настільки просунутих можливостей в точності та ефективності	Вдосконалення вже існуючого, поповнення функціоналу

Для роботи в обраних сегментах ринку сформовано базову стратегію розвитку (таблиці 4.13, 4.14).

Таблиця 4.13. Визначення базової стратегії конкурентної поведінки

Чи є проект «першопрохідцем» на ринку?	Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів?	Чи буде компанія копіювати основні характеристики товару конкурента, і які?	Стратегія конкурентної поведінки*
Так	Так	Можливе запозичення ідей для кращого користувацького досвіду	Інноваційна неперевершеність

Таблиця 4.14. Визначення стратегії позиціонування

Вимоги до товару цільової аудиторії	Базова стратегія розвитку	Ключові конкурентоспро можні позиції власного стартап-проекту	Вибір асоціацій, які мають сформувати комплексну позицію власного проекту (три ключових)
Висока якість аналізу фінансових даних,	Зв’язок з клієнтами, розповсюдження через «сарафанне радіо»,	Зв’язок з клієнтами, вдосконалення функціоналу,	Аналіз, фінанси, прогнозування, точність, ефективність

ефективний розподіл потужностей	вдосконалення функціоналу	ефективність роботи програми	

4.4 Розробка маркетингової програми стартап-проекту

Після проведеного детального аналізу можемо повноцінно описати ключові переваги концепції потенційного продукту та побудувати концепцію маркетингових комунікацій (таблиця 4.15 і таблиця 4.16).

Таблиця 4.15. Ключові переваги концепції потенційного товару

№ п/п	Потреба	Вигода, яку пропонує товар	Ключові переваги перед конкурентами (існуючі або такі, що потрібно створити
1	Точність результатів та ефективність	Продукт надає великі можливості для того, хто його використовує	Перевага в точності та швидкості
2	Функціональність	Великий перелік можливих задач	Наявність унікальних можливостей
3	Проста комунікація з розробниками	Швидке вирішення проблем та усунення несправностей	Необхідність розвивати власну тісно кооперацію з клієнтами

Таблиця 4.16. Концепція маркетингових комунікацій

№ п/п	Специфіка поведінки цільових клієнтів	Канали комунікацій, якими користуються цільові клієнти	Ключові позиції, обрані для позиціонування	Завдання рекламного повідомлення	Концепція рекламного звернення
1	Пошук спеціалізованих систем	Реклама у соціальних мережах; Публікація в спеціалізованих виданнях, журналах; Пошук через знайомих	Точність Якість Орієнтованість на клієнтів Прогнозування Аналіз	Показати унікальність інновацій і спектр можливостей	Презентування можливостей системи і швидкодія якісного продукту

4.5 Висновки до розділу

Даний розділ присвячений дослідженню доцільності створення стартап-проекту, що мав би за ціль створення програмного продукту для аналізу та прогнозування фінансових даних. Було розглянуто стратегії виходу на ринок, маркетингову складову. Оскільки перспективи вбачаються реалістичними та прогнозованими, проект має значну перевагу над конкурентами.

Можна дійти до висновку, що подальша імплементація є доцільною.

ВИСНОВКИ

У магістерській дисертації було проведено теоретичне дослідження методів Інтелектуального Аналізу Даних для фінансових даних. Визначено основні потреби у використанні ІАД та ретельно проаналізовані результати таких досліджень.

Зроблено висновок, що різноманітні методи для здійснення Інтелектуального Аналізу фінансових даних мають свої переваги та недоліки, на які необхідно зважати при виборі коректного інструменту для виконання поставлених цілей, оскільки не існує універсального рішення для всіх можливих задач. Також, причинами невідповідності результатів роботи певного методу до реальних показників та значень можуть бути більш глибинні, приховані аспекти внаслідок недостатньої кількості даних про досліджуваний об’єкт. Наприклад, світова криза чи війна неодмінно повпливають на курс акцій і саме тому прогнозування є такою обширною і складною темою.

Розглянуто сучасні методи для аналізу нестаціонарних фінансових явищ. Також, було запропоновано для побудованих моделей адекватні критерії оцінки. Було розроблено стратегію стартап-проекту та його комерційного використання, досліджено ринкові можливості запуску, визначено фактори загроз та можливостей, проведений аналіз конкурентів, описана ринкова стратегія, сформовані слабкі та сильні сторони та виготовлена концепція маркетингових комунікацій.

Встановлено, що методи, використані в рамках даного дослідження, дають запланований результат з певним рівнем похибки.

В умовах невизначеності та нестабільності економічного та глобального ринкового середовища, оцінка розглянутих методів для прогнозування курсу акцій підтверджує можливість ефективного застосування таких прогнозів.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Шарапов О.Д., Дербенцев В.Д., Семьонов Д.Є. Економічна кібернетика: навч. посіб. / К.: КНЕУ, 2004. 231 с. URL: https://buklib.net/books/24506/ (дата звернення 15.12.2023)
2. Д. В. Ланде, І. Ю. Субач, Ю. Є. Бояринова, Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки : навч. посіб. / КПІ ім. Ігоря Сікорського, Київ, 2018. 297 с.
3. Гороховатський В. О. Методи інтелектуального аналізу та оброблення даних : навч. посіб. / М-во освіти і науки України, Харків. нац. ун-т радіоелектроніки. Харків : ХНУРЕ, 2021. 92 с.
4. Python Vs R: Know The Difference URL: https://www.interviewbit.com/blog/python-vs-r/ (дата звернення 15.12.2023)
5. Що таке Data Mining (Аналіз Даних)? URL: https://futurenow.com.ua/shho-take-data-mining-analiz-danyh/ (дата звернення 15.12.2023)
6. Intro to Rapidminer: A No-Code Development Platform for Data Mining URL: https://www.analyticsvidhya.com/blog/2021/10/intro-to-rapidminer-a-no-code-development-platform-for-data-mining-with-case-study/ (дата звернення 15.12.2023)
7. Lutz M. Learning Python. 3rd ed. Sebastopol, CA : O'Reilly, 2008. 700 p.
8. Mishchuk O. S., Tkachenko R. O. Методи оброблення та заповнення пропущених параметрів у даних екологічного моніторингу. Scientific Bulletin of UNFU. 2019. Vol. 29, No.6. P. 119–122. URL: https://doi.org/10.15421/40290623 (дата звернення: 08.06.2023).
9. Kuznietsova N. V. Identification and dealing with uncertainties in the form of incomplete data by data mining methods. System research and information technologies. 2016. No. 2. P. 104-115. URL: https://doi.org/10.20535/srit.2308-8893.2016.2.10 (дата звернення 15.12.2023).

10. Logit vs Probit Models: Differences, Examples URL: https://vitalflux.com/logit-vs-probit-models-differences-examples/#What_are_Logit_Models (дата звернення 15.12.2023)
11. Understanding Random Forest URL: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (дата звернення 15.12.2023)
12. Колодчак О. М. Інтелектуальний аналіз даних. Вісник Національного університету "Львівська політехніка". Комп'ютерні системи та мережі. 2013. № 773. С. 49-58.
13. Medova E., Smith R. A framework to measure integrated risk. Quantitative Finance, Taylor & Francis Journals, 2005, Vol.5(1), P. 105-121.
14. Тен В. В. Проблеми аналізу кредитоспроможності позичальника. Банківська справа. 2006. № 3.
15. Crouhy M., Mark R., A Comparative Analysis of Current Credit Risk Models. Manuscript, Conference on Credit Risk Modelling and Regulatory Implications, 1998. 117 p.
16. Коваленко О.М., Гуськова В.Г. Моделі інтелектуального аналізу даних для оцінювання фінансових даних. II Всеукраїнська науково-практична конференція «Системні науки та інформатика» з нагоди 125-річчя КПІ ім. Ігоря Сікорського, м. Київ, 04-08 грудня 2023 року. С. 139-146.
17. Коваленко О.М., Муравльов А.Д., Петровський В.Є., Гуськова В. Г. Прогнозування фінансових показників шляхом удосконалення аналітичних методів та моделей на основі передпроцесингу даних. XXII-а Міжнародна науково-практична конференція «Інформаційно-комунікаційні технології та сталий розвиток». Інститут телекомунікацій і глобального інформаційного простору НАН України. С. 70-73.
18. Стаціонарні та нестаціонарні процеси URL: https://studopedia.com.ua/1_280009_statsionarni-y-nestatsionarni-protsesi.html
19. Як працює випадковий ліс URL: https://shorturl.at/rDLTW

ДОДАТОК А

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import SimpleRNN
from keras.layers import Dropout
#
import plotly.graph_objs as go
import plotly.offline as offline
#
import warnings
warnings.filterwarnings('ignore')
import yfinance as yf

from datetime import datetime
start_date = '2017-01-01'
end_date = datetime.today().strftime('%Y-%m-%d')

# Fetch historical stock data from Yahoo Finance
df = yf.download(ticker_symbol, start=start_date, end=end_date)
df = df.reset_index()[['Open','High','Low', 'Volume', 'Adj Close']]

import numpy
import math
from keras.layers import Bidirectional
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import LSTM
from sklearn.tree import DecisionTreeRegressor
```

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from keras.layers import Flatten
from keras.layers import TimeDistributed
from keras.layers import Conv1D
from keras.layers import MaxPooling1D

from pmdarima.arima import auto_arima
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.tsa.statespace.sarimax import SARIMAX

scaler = MinMaxScaler(feature_range=(0, 1))
def reshape_split(df, divider=0.8, time_stemp=10, scaling=True):
    dataset = df.iloc[:,-1].values
    dataset = dataset.reshape(-1,1)
    dataset = dataset.astype("float32")
    dataset.shape

    if scaling:
        dataset = scaler.fit_transform(dataset)

    train_size = int(len(dataset) * divider)
    test_size = len(dataset) - train_size
    train = dataset[0:train_size,:]
    test = dataset[train_size:len(dataset),:]
    print("train size: {},\ntest size: {} ".format(len(train),
len(test)))

    x_train, y_train, x_test, y_test = [], [], [], []

    for i in range(len(train)-time_stemp-1):
        a = train[i:(i+time_stemp), 0]
```

```
        x_train.append(a)
        y_train.append(train[i + time_stemp, 0])
    x_train = numpy.array(x_train)
    y_train = numpy.array(y_train)

    for i in range(len(test)-time_stemp-1):
        a = test[i:(i+time_stemp), 0]
        x_test.append(a)
        y_test.append(test[i + time_stemp, 0])
    x_test = numpy.array(x_test)
    y_test = numpy.array(y_test)


    x_train = numpy.reshape(x_train, (x_train.shape[0], 1,
x_train.shape[1]))
    x_test = numpy.reshape(x_test, (x_test.shape[0], 1,
x_test.shape[1]))
    print(f"x_train: {len(x_train)},\t y_train: {len(y_train)},\t
x_test: {len(x_test)},\t y_test: {len(y_test)} ")
    return x_train, y_train, x_test, y_test

def close_data(df, divider=0.8):
    df_close = df['Adj Close']
    df_log = np.log(df_close)
    train_data, test_data = df_log[:int(len(df_log)*0.8)],
df_log[int(len(df_log)*0.8):]
    return train_data, test_data


def basic_split(df, divider=0.8):
    train = df.iloc[:int(len(df)*0.8)]
    test = df.iloc[int(len(df)*0.8):]
    x_train = train.drop(columns=['Adj Close'])
    y_train = train['Adj Close']
```

```
    x_test = test.drop(columns=['Adj Close'])
    y_test = test['Adj Close']
    return x_train, y_train, x_test, y_test


def dec_tree(x_train, y_train):
    model = DecisionTreeRegressor()
    model.fit(x_train,y_train)
    return model

def rf_tree(x_train, y_train):
    model = RandomForestRegressor(n_estimators = 10, random_state=42)
    model.fit(x_train,y_train)
    return model

def arima(df, order=(2,0,1), divider=0.8):
    train_ar = close_data(df, divider)[0]
    model = ARIMA(train_ar, order=order)
    model_fit = model.fit()
    return model_fit


def sarima(df, order=(2, 0, 1), seasonal_order=(1, 1, 1, 7),
divider=0.8):
    train_ar = close_data(df, divider)[0]

    model = SARIMAX(train_ar, order=order,
seasonal_order=seasonal_order, enforce_stationarity=False,
enforce_invertibility=False)
    model_fit = model.fit()

    return model_fit

def vanilla_lstm(x_train, y_train, time_stemp=10):
```

```
    model = Sequential()
    model.add(LSTM(10, input_shape=(1, time_stemp)))
    model.add(Dense(1))
    model.compile(loss='mean_squared_error', optimizer='adam')
    model.fit(x_train, y_train, epochs=50, batch_size=1)
    return model

def stacked_lstm(x_train, y_train, time_stemp=10):
    model = Sequential()
    model.add(LSTM(50,    activation='relu',    return_sequences=True,
input_shape=(1,time_stemp)))
    model.add(LSTM(50, activation='relu'))
    model.add(Dense(1))
    model.compile(optimizer='adam', loss='mse')
    model.fit(x_train, y_train, epochs=50, batch_size=1)
    return model

def bidirectional_lstm(x_train, y_train, time_stemp=10):
    model = Sequential()
    model.add(Bidirectional(LSTM(50,               activation='relu'),
input_shape=(1,time_stemp)))
    model.add(Dense(1))
    model.compile(optimizer='adam', loss='mse')
    model.fit(x_train, y_train, epochs=50, batch_size=1)
    return model


from    sklearn.metrics    import    r2_score,    mean_absolute_error,
mean_absolute_percentage_error
def build_metrics(models, xtest, ytest):
    result  =  pd.DataFrame(index=models.keys(),  columns=['R2',  'MSE',
'RMSE', 'MAE', 'MAPE'])

    for model in models.keys():
```

```
        print(model)
        if 'lstm' in model:
            y_pred = models[model].predict(xtest)
            y_pred = scaler.inverse_transform(y_pred).reshape(-1,1)
            y_test = scaler.inverse_transform([ytest]).reshape(-1,1)
        elif 'arima' in model:
            y_pred = models[model].forecast(steps=xtest.shape[0])
            y_test = close_data(df)[1][:len(ytest)]
        elif 'tree' in model or 'forest' in model:
            y_pred = models[model].predict(basic_split(df)[2])
            y_test = basic_split(df)[3]


        result.loc[result.index==model]  =      round(r2_score(y_test,
y_pred),3),\

round(mean_squared_error(y_test, y_pred),3),\

round(math.sqrt(mean_squared_error(y_test, y_pred)),3),\

round(mean_absolute_error(y_test, y_pred),3),\

round(mean_absolute_percentage_error(y_test, y_pred),3)
    return result

x_train, y_train, x_test, y_test = reshape_split(df, divider=0.8)

models = {
    'vanilla_lstm': vanilla_lstm(x_train, y_train),
    'stacked_lstm': stacked_lstm(x_train, y_train),
    'bidirectional_lstm': bidirectional_lstm(x_train, y_train),
    'arima': arima(df, order=(2,0,1)),
    'sarimax': sarima(df),
    'decision tree': dec_tree(basic_split(df)[0], basic_split(df)[1]),
```

```
    'random forest': rf_tree(basic_split(df)[0], basic_split(df)[1])
}

build_metrics(models, x_test, y_test)

start_date = '2017-01-01'
end_date = datetime.today().strftime('%Y-%m-%d')

# Fetch historical stock data from Yahoo Finance
dfg = yf.download(ticker_symbol, start=start_date, end=end_date)
company = dfg

company['Adj Close'].plot(figsize=(15,8))
plt.ylabel('Close')
plt.xlabel(None)
plt.title(f"Closing price")

ma_day = [7, 14]

for ma in ma_day:
    column_name = f"MA for {ma} days"
    company[column_name] = company['Adj Close'].rolling(ma).mean()

company[['Adj Close', 'MA for 7 days', 'MA for 14 days']].iloc[-
500:].plot(figsize=(15,8))



dfg['Returns'] = dfg['Close'].pct_change()
data = dfg.iloc[-500:].copy()
data['Returns'] = data['Returns'].fillna(data['Returns'].mean())

from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
```

```
data['Returns'] = scaler.fit_transform(data['Returns'].values.reshape(-
1,1))
data.head()
from sklearn.ensemble import IsolationForest
model = IsolationForest(contamination=0.06)
model.fit(data[['Returns']])

# Predicting anomalies
data['Anomaly'] = model.predict(data[['Returns']])
data['Anomaly'] = data['Anomaly'].map({1: 0, -1: 1})

# Ploting the results
plt.figure(figsize=(15,7))
plt.plot(data['Close'], label='Close')
plt.scatter(data[data['Anomaly'] == 1].index, data[data['Anomaly'] ==
1]['Close'], color='red')
plt.legend(['Close', 'Anomaly'])

plt.xticks(pd.to_datetime(data.loc[data.Anomaly==1].index).strftime('%
Y-%m-%d'),

pd.to_datetime(data.loc[data.Anomaly==1].index).strftime('%Y-%m-%d'),
rotation=90)
plt.show()

import seaborn as sns
corr = data[['Open', 'Close', 'High', 'Low', 'Volume']].corr()
plt.figure(figsize=(10,10))
sns.heatmap(corr, annot=True, cmap='coolwarm')
```