КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Економічний факультет
Кафедра економічної кібернетики

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА «ПРОГНОЗУВАННЯ ЦІН НА АКЦІЇ З ВИКОРИСТАННЯМ СЕНТИМЕНТ-АНАЛІЗУ»

студента 4 курсу
спеціальності 051 «Економіка»
ОПП «Економічна кібернетика»
денної форми навчання
Плющова Владислава Юрійовича

Науковий керівник:
к. ф.-м. н., доцент Кравець Тетяна Вікторівна

Засвідчую, що у цій дипломній роботі немає запозичень із праць інших авторів без відповідних посилань

Студент_____(підпис)

Роботу допущено до захисту перед ЕК
рішенням кафедри економічної кібернетики
від 12.06.2025р., протокол №15
Завідувач кафедри:
доктор економічних наук, професор
Ляшенко Олена Ігорівна _____
(підпис)

КИЇВ – 2025

РЕФЕРАТ

Кваліфікаційна робота бакалавра містить: 74 ст., 24 рис., 20 табл., 62 джерела, 1 додаток.

Ключові слова: часові ряди, акції, моделювання, прогнозування, сентимент аналіз, економетрика, машинне навчання, глибоке навчання

Об'єкт дослідження: ринок акцій.

Предмет дослідження: прогнозування цін на акції з використанням сентимент-аналізу, економетричних, статистичних методів і моделей машинного навчання та нейронних мереж.

Мета дослідження: побудувати різноманітні моделі для прогнозування цін на акції, порівняти їх між собою, визначити ступінь впливу фактору настроїв акціонерів на якість прогнозування.

Методи дослідження: системний підхід, аналіз та синтез, індукція і дедукція, абстрагування, порівняння та узагальнення, статистичний аналіз, економіко-математичне моделювання.

Наукова новизна, теоретична значимість дослідження: у роботі проведене детальне дослідження динаміки котирування акцій компаній Tesla та Apple, чинників впливу як на ринок акцій загалом, так і на акції обраних компаній, використані сучасні підходи для обробки даних та тренування моделей, створені нові фактори (в тому числі на основі технічного аналізу динаміки цін на акції), побудовані економетричні моделі, моделі машинного та глибокого навчання для прогнозування цін на акції, оцінений вплив фактору настроїв акціонерів на формування цін на акції на основі моделювання та візуального аналізу.

Практична цінність: можливість використання результатів кваліфікаційної роботи для прогнозування цін на акції та прийняття зважених рішень стосовно інвестування в обрані цінні папери, розуміння впливу настроїв акціонерів на динаміку котирування акцій.

RESUME

Taras Shevchenko National University of Kyiv,
Faculty of Economics, Department of Economic Cybernetics.

Key words: time series, stocks, modeling, sentiment analysis, econometrics, machine learning, deep learning.

The graduation thesis: Stock prices forecasting using sentiment analysis of stock tweets.

The practical value: Possibility to use the results to forecast stock prices and make informed decisions about investing in stocks, understand the impact of shareholder sentiment on stock price dynamics.

Pages – 74, figures – 24, tables – 20, references – 62, append. – 1.

Зміст

ВСТУП

Важливим аспектом функціонування фінансового ринку є прогнозування цін на акції. Якісне передбачення майбутніх цін відіграє ключову роль для інвесторів, фінансових аналітиків і менеджерів, які ухвалюють рішення щодо розміщення інвестицій, адже від цих рішень залежить їхній прибуток.

Прогнозування цін на акції та інші фінансові інструменти є комплексною задачею, оскільки ціни залежать від багатьох факторів, таких як економічні умови, рішення, прийняті тією чи іншою компанією, настрої акціонерів та інших, які є дуже складними для передбачення. Саме тому ця тема є цікавою не тільки для гравців на ринку акцій, а і для науковців, які постійно знаходять нові моделі та підходи для все більш точних прогнозів.

З появою та подальшим розвитком соціальних мереж набув популярність сентимент-аналіз, що являє собою розділ глибинного аналізу даних (data mining) і область комп'ютерної лінгвістики, що займається вилученням думок та емоцій з текстів. При правильному використанні, висновки, зроблені внаслідок аналізу почуттів акціонерів, можуть стати вагомим та цінним фактором у передбаченні майбутніх цін на акції.

Останні дослідження.

Фондовий ринок - це динамічне та мінливе середовище, де ціни на акції компаній постійно коливаються під впливом різноманітних випадкових та невипадкових факторів, які можуть мати значний вплив як на ціни окремих акцій, так і на ринок в цілому. Огляду фондового ринку присвячені роботи К. Ліма та Р. Брукса [1], а також М. Кендалла [2]. Інвестори та трейдери використовують різні методи аналізу та прогнозування для прийняття обґрунтованих інвестиційних рішень. Зокрема, прогнозуванню цін на акції присвячені роботи Л. Рілла та С. Сейденса [3], Л. Ді Персіо та О. Гончара [4], Дж. Патела та ін. [5]. До найбільш популярних і поширених підходів належить технічний аналіз, який базується на вивченні історичних даних про ціни,

обсяги торгів та інші показники. До робіт з технічного аналізу належать Р. Назаріо та ін. [6], Г. Атсалакіс і К. Валаваніс [7], А. Акедоя та ін. [8].

Однією з перших робіт, яка продемонструвала зв'язок між загальними настроями в Твіттері та коливаннями на фондовому ринку, була робота Дж. Боллена та ін. [9]. У роботі В. Паголу та ін. [10] розглядається використання аналізу настроїв у Твіттері для прогнозування руху цін на акції. У статті Т. Нгуєн та ін. [11] описано підхід до аналізу настроїв у твітах та його інтеграцію з технічними індикаторами для прогнозування цін на акції. С. Дінг та ін. [12] зосереджуються на виявленні новинних подій та їхньому впливі на ціни акцій за допомогою глибокого навчання.

А. Міттал та А. Гоел [13] запропонували підхід, який передбачає збір твітів про певні акції, визначення їх тональності (позитивної, негативної чи нейтральної) за допомогою аналізу настроїв і використання цих даних як вхідних змінних для моделі машинного навчання для прогнозування майбутніх цін. Г. Юхименко та І. Лазаренко [14] запропонували поєднати аналіз настроїв у соціальних мережах та генеративно-конкурентні мережі для підвищення точності прогнозування цін на акції. Однак важливо враховувати різні фактори, що впливають на фондовий ринок. Зокрема, М. Іззельдін та ін. [15] дослідили вплив російсько-української війни на світові фінансові ринки. Їхні результати підтверджують значний негативний вплив військової агресії.

Також все більше уваги приділяється балансуванню даних та застосуванню машинного навчання для прогнозування корпоративного банкрутства, що є важливим елементом фундаментального аналізу [16].

Об’єкт дослідження: ринок акцій.

Предмет дослідження: прогнозування цін на акції з використанням сентимент-аналізу, економетричних, статистичних методів і моделей машинного навчання та нейронних мереж.

Мета дослідження: побудувати різноманітні моделі для прогнозування цін на акції, порівняти їх між собою, визначити ступінь впливу фактору настроїв акціонерів на якість прогнозування.

Завдання дослідження:

- Провести очистку даних («data cleaning»).
- Обробити викиди.
- Провести сентимент-аналіз текстових даних, що стосуються акцій компанії Tesla, виділити фактор почуттів акціонерів. Додати інші фактори, сконструйовані з вже наявних.
- Побудувати моделі прогнозування, вибравши оптимальні гіперпараметри моделей.
- Визначити та порівняти похибки моделей при наявності і відсутності даних про настрої акціонерів.

Робота складається з 3-х розділів та висновку:

- В І розділі описується об’єкт дослідження (ринок акцій), його характеристики, особливості, чинники впливу. Надається опис компанії Tesla, її історії та поточному стану.
- В ІІ розділі представлений предмет дослідження - методи та моделі, за допомогою яких здійснюється прогнозування.
- В ІІІ розділі представлена практична частина: обробка даних, прогнозування та порівняння похибок.
- У висновку зазначений аналіз проведених дій та отриманих результатів.

Програмна реалізація поставлених завдань здійснювалась на мові програмування Python. Сентимент-аналіз був реалізований за допомогою бібліотеки VADER (Valence Aware Dictionary and sEntiment Reasoner) [17], моделі для прогнозування будувались за допомогою трьох бібліотек: statsmodels [18], TensorFlow [19] та scikit-learn [20].

Апробація результатів дослідження. Результати дослідження були апробовані на міжнародній науковій конференції «14th International Conference on Advanced Computer Information Technologies (ACIT)» (Чеські Будейовиці, Чехія, 2024). Вони були включені у бібліотеку досліджень IEEE Xplore [21], яка індексується Web of Science, Scopus та іншими базами наукових робіт. Робота також була апробована на Першій Всеукраїнській науковій конференції «Когнітивні дослідження: результати, виклики та перспективи» [22].

РОЗДІЛ 1. ХАРАКТЕРИСТИКА РИНКУ АКЦІЙ

1.1. Визначення та особливості ринку акцій

Різні вчені та експерти пропонують різні визначення ринку акцій залежно від своїх досліджень та підходів. Так, [23] визначає ринок акцій наступним чином: «Ринок акцій - це ринок, на якому інвестори купують і продають цінні папери, такі як акції та облігації, що представляють право власності на компанії, акції яких котируються на біржі. Він надає компаніям платформу для залучення капіталу, а інвесторам - для отримання прибутку від своїх інвестицій». За визначенням [24] «Ринок акцій - складна система взаємодій між покупцями та продавцями, де ціни на цінні папери визначаються на основі динаміки попиту та пропозиції. Він слугує механізмом розподілу капіталу, ціноутворення ризиків та сприяння інвестиціям». А для [25] «Ринок акцій - це фінансовий інститут, який сприяє обміну правами власності на підприємства шляхом купівлі-продажу акцій. Він забезпечує ліквідність для інвесторів і дозволяє оцінювати компанії на основі ринкового сприйняття та очікувань».

Ці визначення відображають різні погляди на природу і суть функціонування ринку акцій. Важливо зазначити, що фондовий ринок - це складна і динамічна система, яка підлягає постійному дослідженню та аналізу. Розглянемо особливості ринку акцій.

Ринок акцій має кілька ключових особливостей, які визначають його функціонування та впливають на поведінку учасників.

Однією з основних характеристик є обмін цінними паперами. Фондовий ринок виконує роль платформи, де компанії можуть залучати капітал, випускаючи та продаючи акції, а інвестори отримують можливість стати частковими власниками цих компаній. Випуск акцій дає підприємствам доступ до додаткових фінансових ресурсів для розширення бізнесу, розробки нових продуктів чи інвестування в інновації [26]. Водночас інвестори купують акції не лише для отримання права власності, а й з метою отримання прибутку через зростання їхньої вартості або виплату дивідендів. Завдяки

цьому процесу фондовий ринок відіграє важливу роль у розподілі фінансових ресурсів в економіці.

Другим важливим аспектом є учасники ринку, серед яких виділяють індивідуальних і інституційних інвесторів, трейдерів, брокерів, маркет-мейкерів та регуляторні органи. Індивідуальні інвестори – це приватні особи, які купують і продають акції для власного прибутку. Інституційні інвестори, такі як банки, пенсійні фонди та страхові компанії, управляють великими обсягами капіталу і часто впливають на ринкові тренди [27]. Трейдери здійснюють короткострокові операції, використовуючи стратегії спекуляції, тоді як брокери виконують роль посередників між покупцями та продавцями, забезпечуючи виконання угод. Маркет-мейкери підтримують ліквідність, пропонуючи постійні котирування для купівлі та продажу акцій. Регуляторні органи контролюють діяльність учасників ринку, встановлюючи правила для захисту інвесторів і забезпечення стабільності фінансової системи.

Ще однією важливою особливістю є механізм ціноутворення. Вартість акцій на ринку визначається взаємодією попиту та пропозиції [28]. Якщо попит на певні акції зростає, їхня ціна підвищується, тоді як зниження інтересу інвесторів призводить до падіння котирувань. На попит і пропозицію впливає багато факторів, включаючи фінансові показники компаній, макроекономічні умови, ринкові очікування та навіть психологію інвесторів. Важливу роль у ціноутворенні відіграють новини та події: позитивні фінансові звіти, успішний запуск нових продуктів або зростання галузі можуть стимулювати підвищення цін на акції, тоді як економічна нестабільність, політичні кризи або проблеми компанії можуть спричинити падіння котирувань.

Фондовий ринок також відзначається високою ліквідністю, що означає можливість швидко купувати або продавати активи без значного впливу на їхню вартість. Ліквідність визначається обсягом торгів і кількістю учасників ринку [29]. Чим більше акцій перебуває в обігу та чим більше трейдерів зацікавлені в їхній купівлі-продажу, тим вища ліквідність. Ліквідні акції, такі як Apple, Microsoft або Tesla, користуються

стабільним попитом і можуть бути продані в будь-який момент за ринковою ціною. Менш ліквідні активи, зокрема акції маловідомих компаній, можуть мати значні цінові коливання через недостатню кількість покупців і продавців.

Ще одна особливість фондового ринку – це інвестиційні можливості, які він надає учасникам. Ринок дозволяє інвесторам розподіляти капітал між різними компаніями, галузями та навіть країнами, що сприяє диверсифікації портфеля. Диверсифікація допомагає знизити ризики, адже падіння вартості одних активів може компенсуватися зростанням інших. Крім того, фондовий ринок дає змогу заробляти не тільки на купівлі-продажу акцій, але й на отриманні дивідендів, що робить його привабливим як для довгострокових, так і для короткострокових інвесторів.

Разом із можливостями інвестування ринок акцій супроводжується волатильністю та ризиками. Волатильність – це рівень змін у ціні активу протягом певного періоду. Чим вищі коливання, тим вищий ризик, але водночас і більший потенціал для прибутку. Фондові ринки схильні до періодів нестабільності, викликаних глобальними економічними кризами, політичними подіями чи спекулятивними факторами. Поведінка інвесторів, панічні розпродажі або ажіотаж навколо певних акцій можуть значно змінювати ринкові тренди. Саме тому успішне інвестування вимагає аналізу ризиків і розуміння потенційних втрат.

Фондовий ринок функціонує в межах регуляторних норм, які встановлюють спеціальні органи. Регулятори, такі як Комісія з цінних паперів і бірж США (SEC), Європейське управління з цінних паперів і ринків (ESMA) або Китайська комісія з регулювання цінних паперів (CSRC), контролюють діяльність учасників, запобігають шахрайським схемам та інсайдерській торгівлі. В Україні фондовий ринок регулює Національна комісія з цінних паперів та фондового ринку (НКЦПФР) [30]. Встановлення прозорих правил сприяє довірі до ринку, залученню нових інвесторів і підтримці стабільності фінансової системи.

Важливу роль у функціонуванні фондового ринку відіграють ринкові індекси, які використовуються для аналізу динаміки ринку. За [31] «Індекси фондових ринків є індикаторами глобальної економіки, груп країн або національної економіки, інвестиційного клімату в країні, ситуаційного аналізу ринку цінних паперів та прогнозування їх тренду». Індекси, такі як S&P 500, Dow Jones Industrial Average або Nasdaq Composite, представляють середньозважену вартість групи акцій та слугують орієнтирами для оцінки загального стану економіки. Вони відображають зміну ринкових тенденцій, допомагають інвесторам порівнювати ефективність своїх вкладень і прогнозувати майбутні коливання цін [32]. Наприклад, зростання індексу S&P 500 може свідчити про позитивні очікування на ринку, тоді як падіння Dow Jones може бути ознакою економічного спаду.

Таким чином, ринок акцій є складним і динамічним середовищем, яке надає можливості для зростання капіталу, але також містить ризики, що потребують ретельного аналізу. Завдяки механізмам ціноутворення, ліквідності, регулювання та ринкових індексів він залишається ключовим елементом світової фінансової системи.

1.2. Ключові фактори, що впливають на динаміку ринку акцій

Фондовий ринок зазнає впливу широкого спектра факторів, які можуть як підвищувати, так і знижувати ціни окремих акцій та ринку загалом. Вони поділяються на випадкові та невипадкові, тобто такі, що мають передбачуваний або системний характер. Серед основних факторів, що визначають динаміку ринку, виділяють економічні показники, результати діяльності компаній, галузеві тенденції, глобальні події та настрої інвесторів.

Економічні фактори мають визначальне значення для фондового ринку. Валовий внутрішній продукт (ВВП), рівень інфляції, процентні ставки та безробіття впливають на інвестиційну активність [33]. Зростання ВВП свідчить про економічну стабільність, що підвищує довіру інвесторів. Висока інфляція, навпаки, знижує купівельну спроможність і може викликати посилення монетарної політики, що робить кредити

дорожчими та обмежує розвиток бізнесу. Водночас зміни процентних ставок прямо впливають на ринкову динаміку: їхнє зниження стимулює інвесторів вкладати кошти в акції, а підвищення — змушує переорієнтовуватись на менш ризикові активи.

Результати діяльності компаній безпосередньо впливають на їхню ринкову вартість [34]. Фінансові звіти, темпи зростання доходів і рівень боргового навантаження формують уявлення про перспективи компаній. Позитивні звіти можуть стимулювати купівлю акцій, тоді як негативні новини призводять до зниження їхньої вартості. Інвестори уважно аналізують фундаментальні показники, щоб оцінити потенційну прибутковість вкладень.

Галузеві та секторальні тенденції визначають динаміку акцій окремих секторів [35]. Наприклад, технологічний прогрес стимулює зростання вартості компаній у сфері штучного інтелекту та електромобілів, тоді як традиційні галузі можуть зазнавати тиску через регуляторні обмеження або зміну споживчих уподобань. Крім того, законодавчі зміни, податкові реформи та державна політика впливають на розвиток окремих секторів економіки, визначаючи їхню інвестиційну привабливість.

Глобальні та геополітичні події також відіграють важливу роль. Політична нестабільність, війни [15], торговельні конфлікти, санкції та стихійні лиха можуть викликати різкі коливання фондових ринків. Наприклад, міжнародні санкції або нові торговельні обмеження впливають на прибутковість компаній, що працюють на глобальних ринках, а військові конфлікти створюють невизначеність і спричиняють відтік капіталу до більш захищених активів, таких як золото та державні облігації.

Настрої інвесторів та ринкова психологія можуть призводити до значних коливань цін. Оптимізм і довіра до ринку стимулюють зростання котирувань, тоді як панічні настрої можуть викликати масові розпродажі [36]. Ефект "стада", коли інвестори повторюють дії більшості, часто підсилює волатильність. Крім того, засоби масової інформації та соціальні мережі можуть впливати на ринок, поширюючи як аналітичні прогнози, так і спекулятивні заяви, що змінюють інвестиційні настрої.

В практичній частині роботи прогнозувались ціни на акції компанії Tesla, для валідації результатів прогнозування відбувалось також для акцій компанії Apple. Рух цін на акції даних компаній суттєво залежить від наведених вище особливостей ринків акцій та факторів, проте їм також властиві більш специфічні фактори впливу, враховуючи технологічну направленість цих компаній.

Таким чином для Tesla ключову роль відіграють технологічні інновації та розвиток електромобілів [37]. Постійне вдосконалення акумуляторних батарей, покращення програмного забезпечення для автономного водіння та впровадження енергозберігаючих рішень значно впливають на інтерес інвесторів і вартість акцій компанії.

Окремим аспектом, що є специфічним для Tesla та впливає на котирування її акцій, є дії її генерального директора Ілона Маска. Його висловлювання в соціальних мережах або стратегічні ініціативи, такі як зміна фокусу на штучний інтелект або робототехніку, часто спричиняють суттєві коливання ціни акцій компанії. Наприклад, у 2022 році після придбання Маском компанії Twitter, акції Tesla знизилися з майже $400 на початку року до $123 наприкінці, що викликало занепокоєння серед інвесторів та працівників компанії. [38]

У випадку Apple одним із визначальних факторів є запуск нових продуктів. Вихід оновлених моделей iPhone, Mac та інших пристроїв викликає підвищений інтерес ринку, адже успішні продажі напряму впливають на фінансові показники компанії [39]. Паралельно Apple активно розвиває екосистему послуг, зокрема платформи підписок Apple Music, iCloud та Apple TV+. Стабільне зростання цього напряму забезпечує компанії прогнозовані доходи, що позитивно позначається на привабливості її акцій.

Ще одним важливим аспектом є стратегія Apple щодо мікропроцесорів. Перехід на власні чіпи серії M, замість використання процесорів Intel, дозволив компанії покращити продуктивність своїх пристроїв, оптимізувати витрати та зміцнити незалежність від зовнішніх постачальників. Водночас Apple залишається вразливою до перебоїв у

постачанні, оскільки значна частина її виробничих потужностей розташована в Китаї [40]. Будь-які торговельні обмеження або геополітична напруженість можуть негативно позначитися на ланцюгах постачання та темпах випуску продукції.

Крім цього, на ціну акцій Apple впливають фінансові стратегії компанії, зокрема програми викупу власних акцій та дивідендна політика. Регулярні викупи підтримують вартість акцій на високому рівні, а виплата дивідендів робить Apple привабливою для довгострокових інвесторів.

Таким чином, хоча Tesla та Apple загалом підпадають під вплив загальних ринкових чинників, їхні акції демонструють залежність від унікальних особливостей бізнес-моделі, технологічного розвитку та стратегічних рішень керівництва.

1.3. Особливості компаній Tesla та Apple як представників технологічного сектору фондового ринку

Tesla Inc. - американська автомобільна компанія-стартап із Кремнієвої долини. Орієнтована на виготовлення та збут електромобілів і компонентів до них. Названа на честь всесвітньо відомого американського фізика сербського походження Ніколи Тесли.

Tesla - одна з найдорожчих компаній світу за ринковою капіталізацією. З липня 2020 року вона утримує титул найдорожчого автовиробника у світі. У жовтні 2021 року компанія досягла капіталізації у трильйон доларів США, ставши сьомою американською компанією, якій це вдалося.

Особливістю та драйвером успіху Tesla є те, що вона продовжує зосереджуватися на створенні електромобілів та виробництві систем і компонентів для силових агрегатів електромобілів (EV). Станом на 2022 рік компанія має мережу з 438 магазинів і галерей, а також близько 100 сервісних центрів. [41]

Акції Tesla постійно зростали з першого дня торгів у 2010-2019 роках, але головним чином їх стримувала нездатність компанії стати прибутковою. У 2020 році акції компанії зросли майже в 10 разів після отримання першого річного прибутку. Причиною цьому

став бум електромобілів, під час якого Tesla займала найсильнішу позицію на американському ринку. [42]

Окрім випуску електромобілів, Tesla займається розробкою технологій автономного керування та програмного забезпечення, що дозволяє автомобілям функціонувати у режимі часткового автопілота. Це ще більше підвищило привабливість її продукції для споживачів і сприяло зростанню капіталізації. Компанія також активно працює над розвитком енергетичних рішень, зокрема виробництвом акумуляторних систем зберігання енергії, таких як Powerwall та Powerpack, а також сонячних панелей. Таким чином, Tesla не просто є виробником електромобілів, а й ключовим гравцем у сфері відновлюваної енергетики.

Apple Inc. - американська технологічна компанія, що спеціалізується на розробці, виробництві та збуті електронних девайсів, програмного забезпечення та онлайн-сервісів. Вона є однією з найвпливовіших компаній у світі, відомою завдяки своїм інноваційним рішенням і стратегічним підходам до розвитку бренду та своєї продукції. Заснована в 1976 році Стівом Джобсом, Стівом Возняком і Рональдом Вейном, Apple пройшла довгий шлях від виробника персональних комп'ютерів до глобального технологічного лідера.

Головними напрямками діяльності Apple є виробництво смартфонів iPhone, планшетів iPad, комп'ютерів Mac, носимих пристроїв Apple Watch та AirPods, а також програмних продуктів, таких як операційні системи iOS і macOS. Компанія також активно розвиває свої сервіси, серед яких App Store, iCloud, Apple Music, Apple TV+ та Apple Pay. Саме сегмент сервісів став ключовим джерелом прибутку для Apple у 2020-х роках, поступово знижуючи залежність від продажів апаратного забезпечення.

Акції Apple традиційно демонструють високу стабільність і привабливість для інвесторів завдяки постійному зростанню виручки та значній маржинальності бізнесу. У серпні 2018 року Apple стала першою компанією в історії США, ринкова капіталізація якої перевищила трильйон доларів. У серпні 2020 року вона першою подолала позначку

у два трильйони, а у 2022 році – три трильйони доларів США. Це робить її найдорожчою компанією у світі.

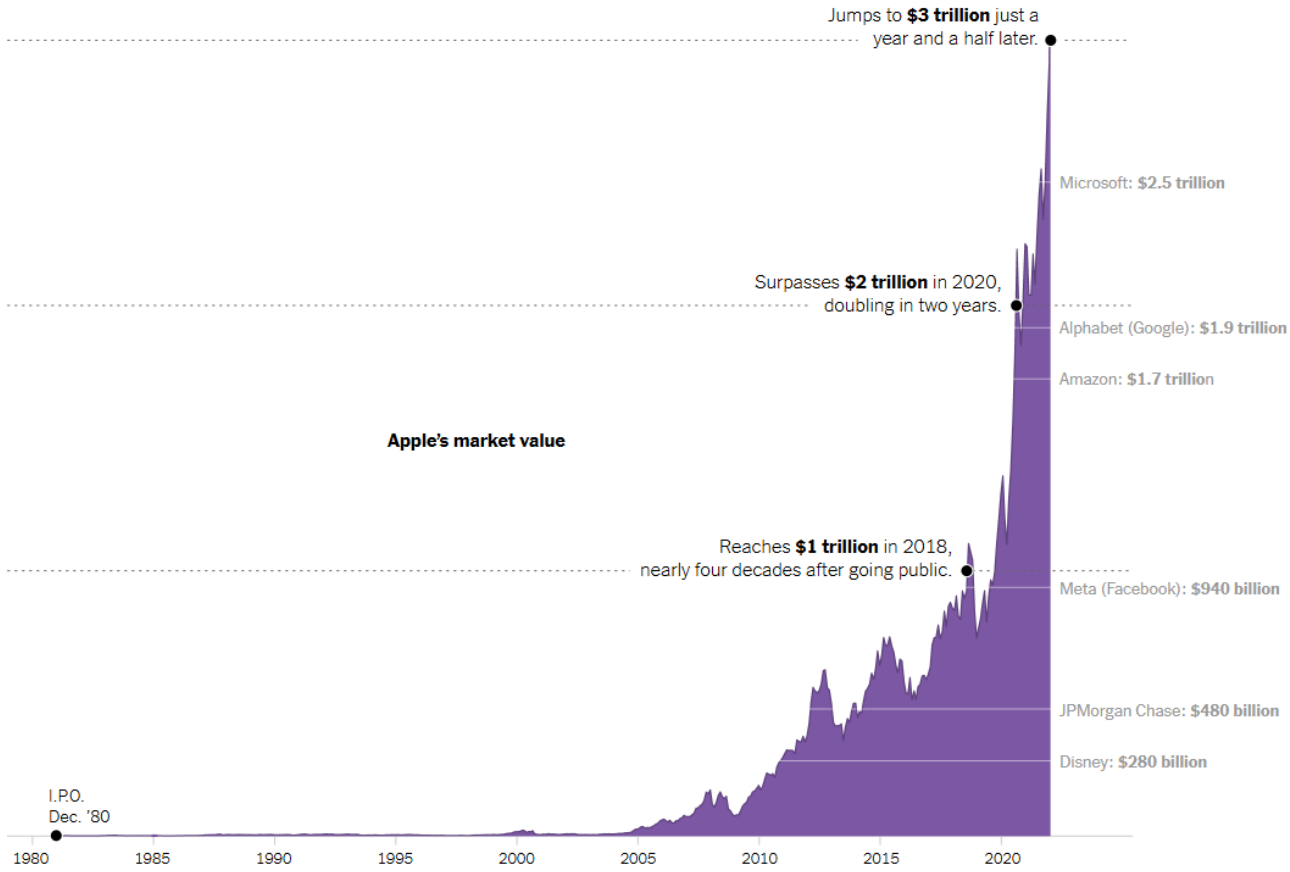

Рис. 1.1. Динаміка цін на акції Apple до моменту досягнення позначки капіталізації в $3 трлн.

Джерело: [43]

Особливістю Apple є її замкнута екосистема, що дозволяє глибоко інтегрувати апаратне та програмне забезпечення. Завдяки цьому компанія створює продукти, які забезпечують бездоганний користувацький досвід і високу лояльність клієнтів. Бренд Apple асоціюється з інноваціями, дизайном і преміальним сегментом ринку, що дає змогу зберігати високу маржинальність продукції. Крім того, значна увага приділяється питанням конфіденційності та безпеки даних, що вигідно вирізняє компанію серед конкурентів.

Таким чином, Tesla та Apple є провідними представниками технологічного сектору фондового ринку, кожна з яких має унікальні особливості ведення бізнесу. Tesla змінює автомобільну індустрію завдяки електромобілям і відновлюваній енергетиці, тоді як Apple задає тренди у сфері споживчої електроніки та цифрових сервісів. Обидві компанії демонструють значний вплив на глобальні фінансові ринки, що підтверджується їхньою ринковою капіталізацією та інвестиційною привабливістю.

Наведені в цьому розділі визначення, особливості та чинники впливу дають підґрунтя для подальшого поглиблення в предметну область - прогнозування цін на акції.

РОЗДІЛ 2. ТЕОРЕТИЧНІ ЗАСАДИ ВИКОРИСТАНИХ МОДЕЛЕЙ ТА ПІДХОДІВ

2.1. Огляд моделей та методів прогнозування цін на акції

В сучасному світі існує велика кількість моделей, які здатні описувати взаємозалежності між факторами та за допомогою цього прогнозувати майбутні ціни на акції на основі минулих даних. В цій роботі були використані 6 різних за своєю суттю та структурою моделей та ансамбль моделей. Детальнішу інформацію про них наведено нижче.

Bidirectional LSTM (BiLSTM) - це тип рекурентної нейронної мережі (RNN), яка стала популярною завдяки своїй здатності моделювати та прогнозувати дані часових рядів, включаючи ціни на акції. Дана модель є розширенням традиційної LSTM-мережі, яка здатна обробляти дані не лише в прямому, а й у зворотному напрямку. Це дозволяє моделі враховувати як попередній контекст, так і майбутній, що є важливим у задачах часових рядів, зокрема для прогнозування цін на акції. У фінансовому аналізі ця здатність дозволяє краще враховувати складні взаємозв'язки та тренди, які можуть впливати на майбутні зміни ціни [44].

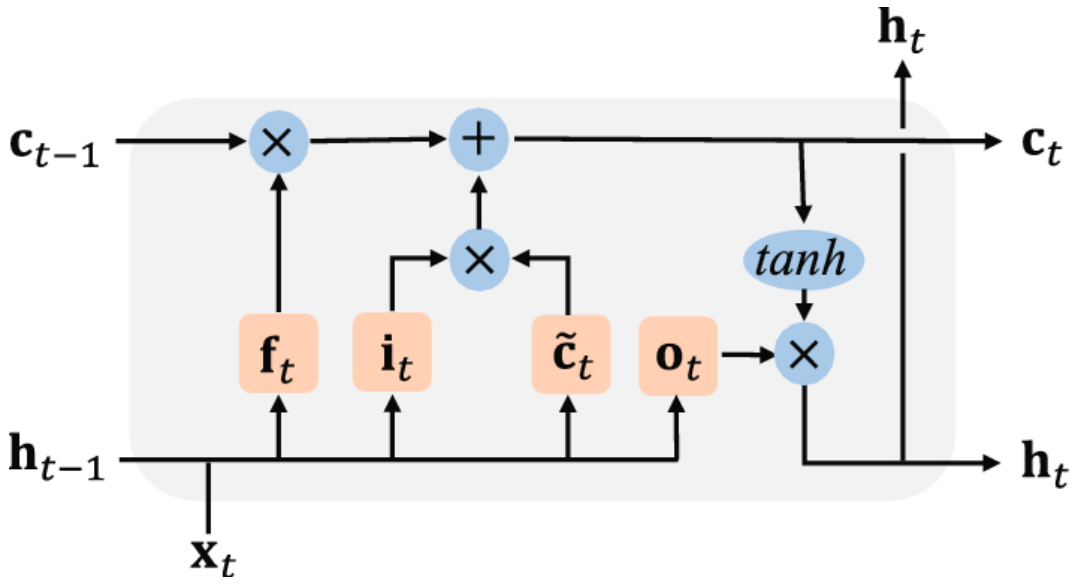

Рис. 2.1. Архітектура комірки моделі BiLSTM

Джерело: [45]

Історичні дані про ціни на акції подаються у BiLSTM як послідовність вхідних даних. Завдяки двонаправленій архітектурі модель обчислює представлення даних двома способами: з початку до кінця та у зворотному порядку. Це дає змогу отримати багатше представлення інформації та зменшити ризик втрати важливих залежностей у часових рядах.

Як і в традиційній LSTM, під час навчання BiLSTM-мережа налаштовує ваги своїх зв'язків, щоб мінімізувати функцію втрат. У даній роботі для цієї мети використовувалося середньоквадратичне відхилення, яке забезпечує вимірювання різниці між прогнозованими та фактичними значеннями [46]

k-Nearest Neighbors (k-NN) - це тип алгоритму машинного навчання, який часто використовується для прогнозування цін на акції. k-NN - це непараметричний метод, що означає, що він не робить жодних припущень про основний розподіл даних [47]. Замість цього він використовує самі дані для прогнозування.

Загальна ідея k-NN полягає в порівнянні точок даних у навчальному наборі з точкою, що прогнозується, та визначає k найближчих сусідів на основі обраної метрики відстані. Потім алгоритм робить прогноз на основі значень найближчих сусідів, усереднюючи їх.

Однією з переваг k-NN для прогнозування акцій є те, що він простий і легкий в реалізації. Його можна використовувати як для регресії (прогнозування безперервних змінних, таких як ціни на акції), так і для класифікації (прогнозування класів, наприклад, чи піде ціна вгору або вниз).

Приклад використання k-NN зображений на рис. 2.2. Синє коло уособлює новий екземпляр, значення якого потрібно визначити. Жовті кола – тренувальні дані. При k = 3 червоний екземпляр прийме значення $\frac{29+30+31}{3} = \frac{90}{3} = 30$. При k = 6 - $\frac{29+30+31+30+32+34}{6} = \frac{186}{6} = 31$.

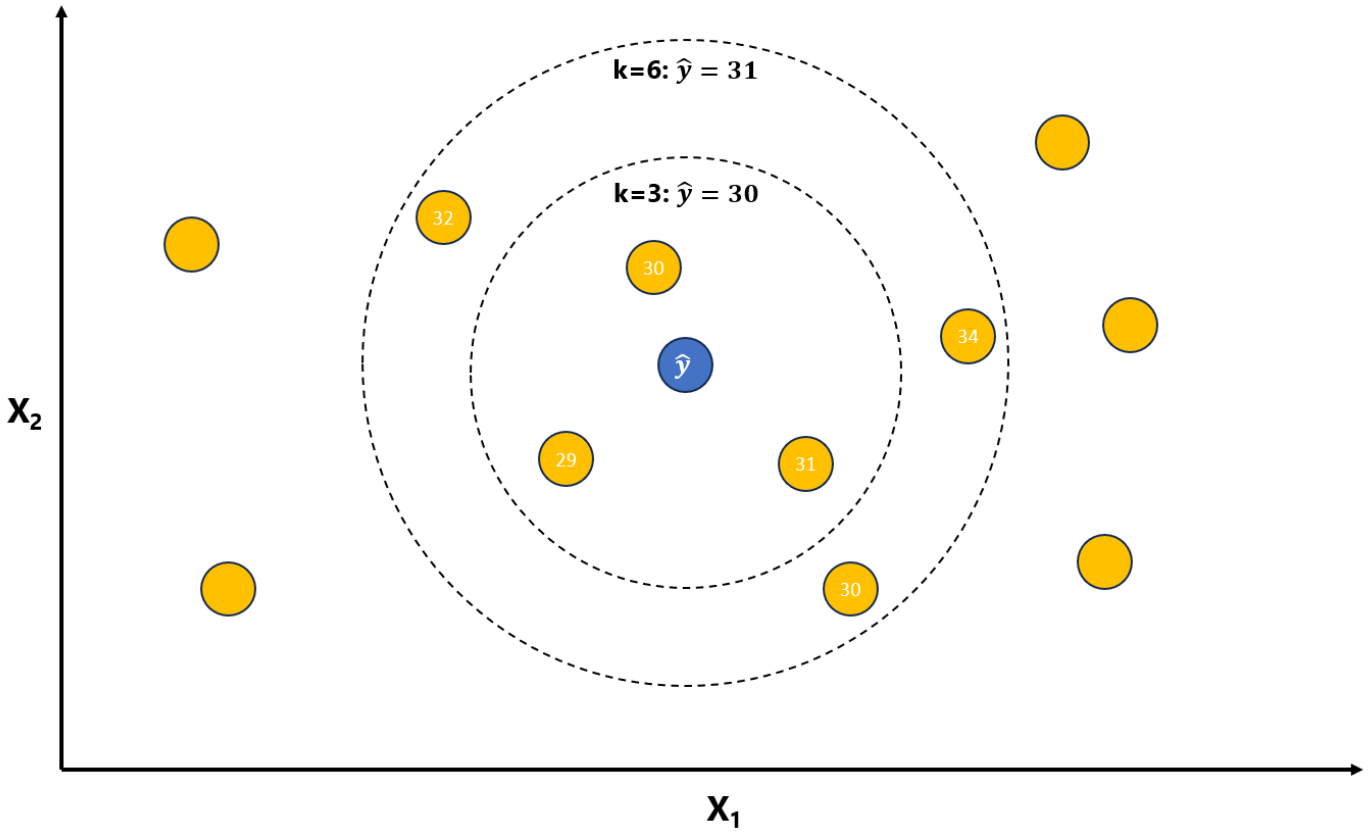

Рис. 2.2 Приклад використання k-NN для задачі регресії

Джерело: розрахунки автора

ARIMA - це популярна і широко використовувана техніка аналізу часових рядів, у тому числі руху цін на акції.

ARIMA моделі складаються з трьох основних компонентів [48]:

1. авторегресійного (AR) компонента. Запис AR(p) позначає, що авторегресія змінної y була побудована на p лагів її значень.

2. інтегрального (I) компонента. I(d) позначає, що беруться d послідовних різниць часового ряду.

3. компонента ковзного середнього (MA). Запис MA(q) означає, що регресія змінної y побудована на q лагів її збурень.

Авторегресійний компонент базується на ідеї, що майбутні значення часового ряду можна передбачити на основі його минулих значень. Компонент ковзного середнього ґрунтується на тому, що майбутні значення можна передбачити на основі помилок

минулих прогнозів. Інтегральний компонент використовується для приведення ряду до стаціонарного шляхом взяття послідовних різниць компонентів.

В практичній імплементації цієї моделі в даній роботі була використана AutoARIMA з бібліотеки statsmodels, що автоматично підбирає оптимальні параметри моделі. Підбір параметру d відбувається за статистичним тестом ADF на стаціонарність ряду. Якщо за ним ряд не стаціонарний, то береться різниця елементів. Процес продовжується допоки ряд не стане стаціонарним. Параметри p та q підбираються за максимізацією інформаційного критерію AIC.

Лінійна регресія – це популярна і широко використовувана модель прогнозування цін на акції. Це простий та інтуїтивно зрозумілий метод, який можна використовувати для моделювання взаємозв'язку між ціною акції та різними факторами. Функціональна залежність між незалежними змінними і залежною є лінійною [49]. Коефіцієнти при змінних знаходять за допомогою мінімізації квадратичної функції втрат.

Однією з переваг лінійної регресії є її простота та зрозумілість. Модель знаходить коефіцієнти для кожної змінної, вказуючи на силу і напрямок її зв'язку з ціною акції. Це може допомогти аналітикам та інвесторам визначити, які змінні є найбільш важливими для прогнозування майбутньої ціни акції.

Однак є й обмеження у використанні лінійної регресії для прогнозування акцій. Наприклад, лінійна регресія передбачає, що зв'язок між незалежними змінними та залежною змінною є лінійним, що не завжди відповідає дійсності. Крім того, лінійні регресійні моделі можуть бути чутливими до викидів або інших аномалій у даних.

У цій роботі замість традиційної лінійної регресії була використана **Lasso-регресія** (Least Absolute Shrinkage and Selection Operator), яка є вдосконаленим методом лінійного моделювання. Lasso-регресія застосовує L1-регуляризацію, додаючи до функції втрат суму абсолютних значень коефіцієнтів [50]. Це дозволяє моделі виконувати одночасно відбір ознак і зменшення ваг змінних, що робить її особливо ефективною в задачах з великою кількістю факторів або мультиколінеарністю.

Однією з ключових переваг Lasso-регресії є її здатність автоматично виключати незначущі змінні, задаючи їхні коефіцієнти рівними нулю. Це значно спрощує модель, покращує її інтерпретованість і зменшує ризик перенавчання. Крім того, вона менш чутлива до викидів у порівнянні зі звичайною лінійною регресією. Приклад імплементації моделі лінійної регресії продемонстрований на рис. 2.3.

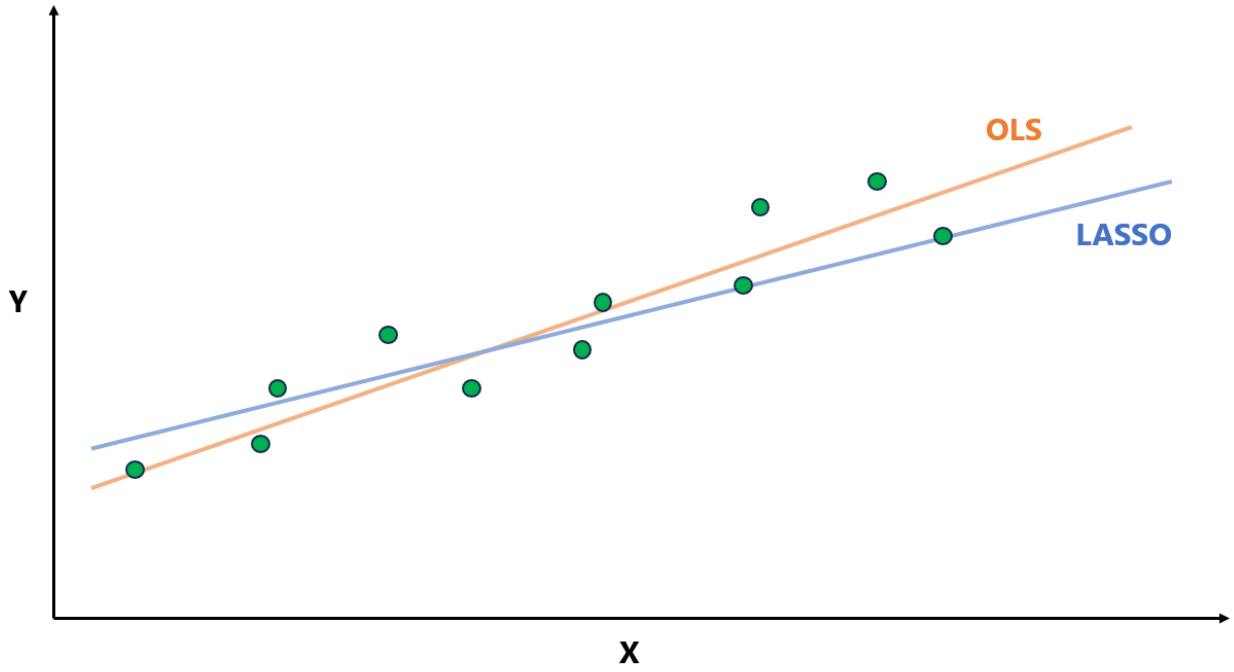

Рис. 2.3. Ілюстрація використання Lasso-регресії

Джерело: розрахунки автора

Random Forest - це метод машинного навчання, який показує багатообіцяючі результати у прогнозуванні акцій. Ця модель використовує набір дерев рішень для прогнозування майбутньої ціни акцій.

Основна ідея Random Forest полягає у створенні декількох дерев рішень, кожне з яких навчається на випадковій підмножині наявних даних заданого розміру, відібраних методом вибору з повтореннями (bootstrap sampling) [51]. Агрегуючи прогнози цих дерев, модель може зменшити варіацію і підвищити точність своїх прогнозів.

Однією з переваг моделі Random Forest є її здатність обробляти нелінійні зв'язки між змінними. Це робить її добре пристосованою до складної та динамічної природи даних фондового ринку, які часто демонструють нелінійні та нестаціонарні зв’язки.

Загалом, Random Forest є перспективним для прогнозування цін на акції, але його ефективність залежить від якості даних та ретельного відбору змінних і гіперпараметрів.

На рис. 2.4. продемонстрована архітектура Random Forest з параметром кількості естиматорів (дерев рішень) рівному 600.

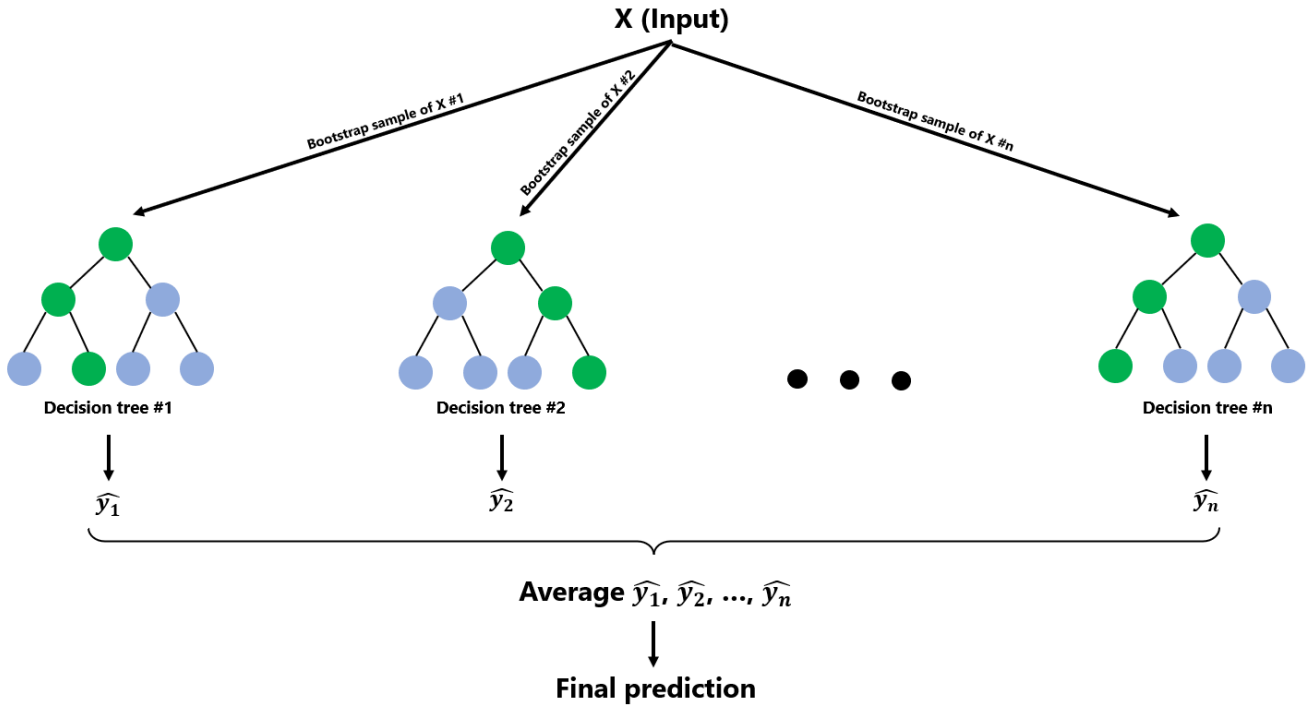

Рис. 2.4. Архітектура моделі Random Forest

Джерело: розрахунки автора

Градієнтний бустинг — це ансамблевий метод машинного навчання, який поєднує багато слабких моделей (зазвичай дерев рішень) для створення сильної прогностичної моделі [52]. Основна ідея цієї техніки полягає в тому, що моделі тренуються послідовно, і кожна нова модель коригує помилки попередньої. Процес починається з простої базової моделі, яка прогнозує середнє значення цільової змінної. Після цього обчислюється градієнт функції втрат (в задачі регресії, фактично, залишки, тобто різниця між фактичними значеннями та прогнозами моделі). Наступні моделі будуються

для передбачення цих залишків, і результати додаються до загального прогнозу. Кожна ітерація модифікує модель так, щоб вона все краще наближалася до реальних значень. Ця поетапна оптимізація триває доти, доки модель не досягне заданої точності або не буде виконано задану кількість ітерацій.

У задачах прогнозування цін на акції градієнтний бустинг корисний тим, що дозволяє ефективно враховувати нелінійні залежності між ринковими індикаторами, такими як попередні ціни, обсяги торгів, макроекономічні показники тощо. Це робить метод особливо цінним для складних фінансових систем із великою кількістю змінних.

LightGBM (Light Gradient Boosting Machine або LGBM) – це алгоритм градієнтного бустингу, який використовує побудову дерев рішень і відзначається високою швидкістю та ефективністю обчислень. Його головна особливість полягає у застосуванні технологій Histogram-based Learning та Leaf-wise Growth Strategy, що відрізняє його від традиційних реалізацій градієнтного бустингу, зокрема XGBoost.

Головний принцип роботи LightGBM полягає в побудові ансамблю дерев рішень, де кожне нове дерево намагається мінімізувати залишкову похибку попередніх. На відміну від традиційного підходу, де розгалуження здійснюється рівномірно на рівнях дерева (level-wise), LightGBM використовує стратегію розширення дерев у глибину (leaf-wise). Це означає, що на кожному кроці вибирається найкращий за приростом інформації лист для подальшого розгалуження. Такий підхід забезпечує вищу точність порівняно з рівневим зростанням, оскільки дозволяє краще адаптуватися до складних залежностей у даних [53].

Ще однією важливою особливістю LightGBM є використання групування (бінінг) ознак за допомогою гістограм, що дозволяє суттєво зменшити кількість операцій порівняння і прискорити обчислення. Це робить алгоритм значно швидшим за класичні реалізації градієнтного бустингу, особливо при роботі з великими вибірками. До того ж LightGBM ефективно працює з розрідженими даними, автоматично враховуючи відсутні або нульові значення у процесі побудови дерев.

Застосування LightGBM у прогнозуванні цін на акції пояснюється його здатністю ефективно працювати з великими обсягами даних і складними нелінійними залежностями. Фінансові часові ряди часто містять значну кількість ознак, включаючи фундаментальні, технічні та макроекономічні показники, а також демонструють високу мінливість і залежності, які важко моделювати стандартними методами регресії. LightGBM дозволяє швидко обробляти великі масиви фінансових даних, ідентифікувати приховані патерни та забезпечувати високу точність прогнозування. Завдяки цьому його активно використовують у фінансовій аналітиці для завдань передбачення руху котирувань, оцінки ризиків та побудови торгових стратегій. Візуалізація архітектури нейронних мереж представлена на рис. 2.5.

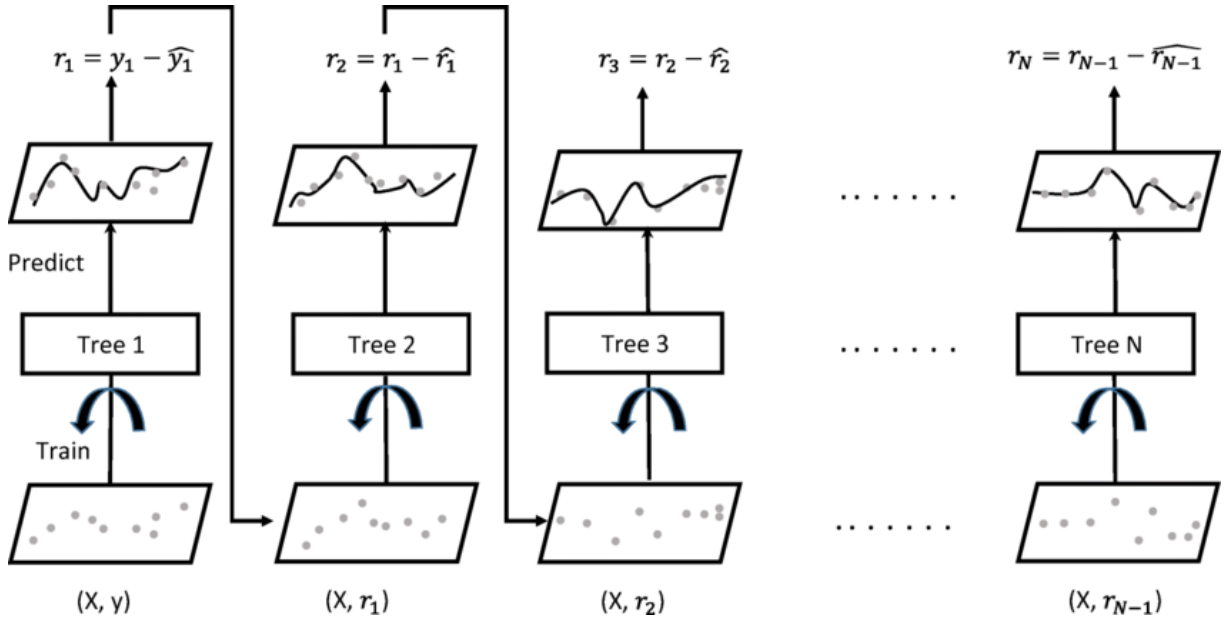

Рис. 2.5. Візуалізація алгоритму навчання моделі градієнтного бустингу

Джерело: [54]

На момент написання роботи градієнтний бустинг та його модифікації (XGBoost, CatBoost, використана в роботі модель LightGBM та ін.) багатьма науковцями та практиками вважаються одними з найсильніших моделей для табличних даних на рівні з передовими архітектурами нейронних мереж. Так, до прикладу, за матеріалами [55] на 176 наборах табличних даних порівнювались архітектури нейронних мереж, модифікації градієнтного бустингу та класичні моделі машинного навчання. В цій роботі дослідники

дійшли висновку – в середньому градієнтний бустинг на основі дерев рішень переграє нейронні мережі.

Ансамбль моделей (Ensemble) - це метод комбінування кількох базових моделей для покращення точності та стійкості прогнозів. Логіка ансамблів базується на припущенні, що окремі моделі, хоч і можуть мати похибки, надають корисну інформацію. Об'єднання цих моделей дозволяє зменшити вплив помилок і посилити важливі сигнали, які кожна з них виявляє.

В даній роботі для побудови ансамблю був використаний підхід бегінгу [56] при якому моделі тренуються незалежно одна від одної, після чого їхні результати зважено усереднюються. Для побудови ансамблю були обрані моделі k-NN, Lasso, Random Forest та LightGBM, оскільки вони представляють різні підходи до моделювання. k-NN враховує локальні подібності між спостереженнями, Lasso забезпечує лінійне моделювання з відбором релевантних ознак, Random Forest ефективно виявляє нелінійні зв’язки, а LightGBM поєднує високу точність із швидкістю обчислень. Їхня різноманітність дозволяє збалансувати похибки окремих моделей і підвищити загальну якість прогнозу. Вагові коефіцієнти моделей вважались гіперпараметрами моделі і налаштовувались аналогічним до гіперпараметрів інших моделей чином (розділ 2.2).

Сентимент-аналіз, також відомий як «**аналіз настроїв**», - це метод, що використовується для визначення настроїв або емоційного тону, виражених у тексті, наприклад, у повідомленнях у соціальних мережах, відгуках клієнтів, статтях у новинах або онлайн-дискусіях [7]. Він передбачає аналіз тексту для виявлення та класифікації суб'єктивної інформації на позитивну, негативну або нейтральну та визначення ступеня впевненості в своєму рішенні.

Мета аналізу настроїв - отримати цінну інформацію з великих обсягів неструктурованих текстових даних. До прикладу, зрозумівши настрої, що стоять за текстом, організації можуть глибше зрозуміти громадську думку, відгуки клієнтів, ринкові тенденції та сприйняття бренду або, як в нашому випадку, дослідники можуть

отримати цінний фактор для подальшого прогнозування цін на акції з врахуванням наявних настроїв на ринку [57].

Одним із способів отримання бази даних для подальшого сентимент-аналізу є моніторинг соціальних мереж. Такий спосіб є одним з найбільш популярних на даний момент через велику кількість людей, які регулярно відвідують та залишають повідомлення на таких платформах як Twitter, Facebook та Instagram, і наявність хештегів – спеціальних символів, за якими без зайвих зусиль можна знайти всі відкриті для загального доступу повідомлення, які стосуються обраної тематики. Проте в таких середовищах, як соціальні мережі, іноді складно отримати репрезентативну оцінку настроїв як окремих повідомлень (через використання користувачами сарказму, скорочень, «сленгу» тощо), так і об’єкту дослідження загалом (через повідомлення «фейкових» користувачів або ботів).

Загалом, сентимент-аналіз, при правильному використанні, може стати потужним методом, здатним суттєво покращити точність прогнозів та зменшити похибки.

2.2. Принципи налаштування гіперпараметрів у прогнозних моделей

В даній роботі гіперпараметри підбирались за допомогою крос-валідації, що являє собою метод оцінки ефективності моделі і який активно використовується в машинному навчанні. Ідея крос-валідації полягає у використанні підмножини наявних даних для навчання моделі, а іншої підмножини - для перевірки її роботи. Повторюючи цей процес з різними підмножинами даних, крос-валідація дає більш надійну оцінку ефективності моделі, ніж перевірка на тренувальних даних або розділення тренувальної множини на єдину навчальну і єдину валідаційну вибірки.

Існує декілька типів крос-валідації. В цій роботі була обрана крос-валідація у вікні, що розширюється. Її суть полягає в тому, що дані розбиваються таким чином, щоб на часовій осі валідаційні дані завжди були після тренувальних. З кожною наступною ітерацією попередня валідаційна вибірка додається до тренувальної, а нові валідаційні дані беруться у майбутньому. Цей процес повторюється k разів допоки в оцінюванні не

будуть задіяні всі тренувальні дані. Потім результати усереднюються за k ітерацій, щоб отримати єдину оцінку ефективності моделі.

Причина вибору такого типу крос-валідації полягає в тому, що при постановці завдання прогнозування майбутніх значень, справедливу оцінку якості моделей можуть давати лише ті дані, що знаходяться в майбутньому відносно тренувальних. В іншому ж випадку можна отримати занадто оптимістичні оцінки, які до того ж можуть призводити до вибору неоптимальних гіперпараметрів.

Схема крос-валідації у вікні, що розширюється зображена на рис. 2.6.

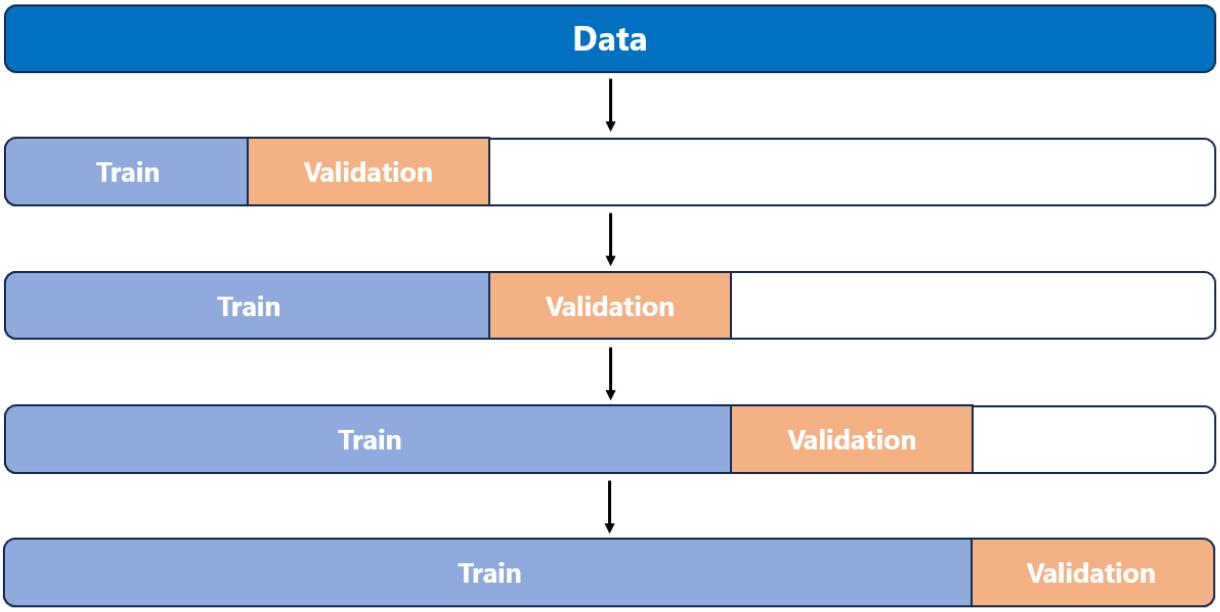

Рис. 2.6. Крос-валідація, у вікні що розширюється при розбитті на 4 підмножини

Джерело: розрахунки автора

Для оптимізації гіперпараметрів була використана бібліотека Optuna [58], яка реалізує підхід байєсової оптимізації. Логіка цього процесу базується на наступному алгоритмі. Спершу Optuna генерує початкові випадкові комбінації гіперпараметрів, щоб зібрати початкову інформацію про цільову функцію. Далі використовуються результати

оцінок попередніх ітерацій для побудови ймовірнісної моделі, яка апроксимує залежність між гіперпараметрами та значенням цільової метрики.

Загалом вибір гіперпараметрів в даній роботі здійснюється наступним чином:

- Обираються види гіперпараметрів та відповідні інтервали або набір можливих значень, в який кожен з гіперпараметрів може входити.
- Ітеративним чином алгоритмом з оптимізаційної бібліотеки optuna обираються значення кожного гіперпараметру.
- Моделі навчаються для кожного набору гіперпараметрів, після чого проходять крос-валідацію та повертають відповідну похибку.
- Гіперпараметри, що дають найменшу похибку моделей, стають гіперпараметрами моделі.

РОЗДІЛ 3. ПРАКТИЧНЕ ЗАСТОСУВАННЯ МЕТОДІВ ТА МОДЕЛЕЙ В ПРОГНОЗУВАННІ ЦІН НА АКЦІЇ

3.1. Опис та попередня обробка даних для прогнозування цін акцій

Дослідження акційних даних

В цьому дослідженні в якості даних були використані ціни на акції компанії Tesla (TSLA) та Apple (AAPL) в період з 30.09.2021 по 29.09.2022 [59]. В наборі даних наявні 252 рядки та 6 показників:

- *Open* – ціна на момент відкриття торгів.
- *High* – найвища ціна за день.
- *Low* – найменша ціна за день.
- *Close* – ціна на момент закриття торгів.
- *Adj Close* – ціна на момент закриття торгів після виплати усіх дивідендів.
- *Volume* – обсяг проданих акцій за день.

Фактор Adj Close був видалений, тому що для цього набору даних його значення є ідентичними значенням фактору Close, відповідно додаткової інформації від цього поля отримати не можна.

На рис. 3.1. зображені Close ціни на акції компаній Tesla та Apple за весь досліджуваний період.

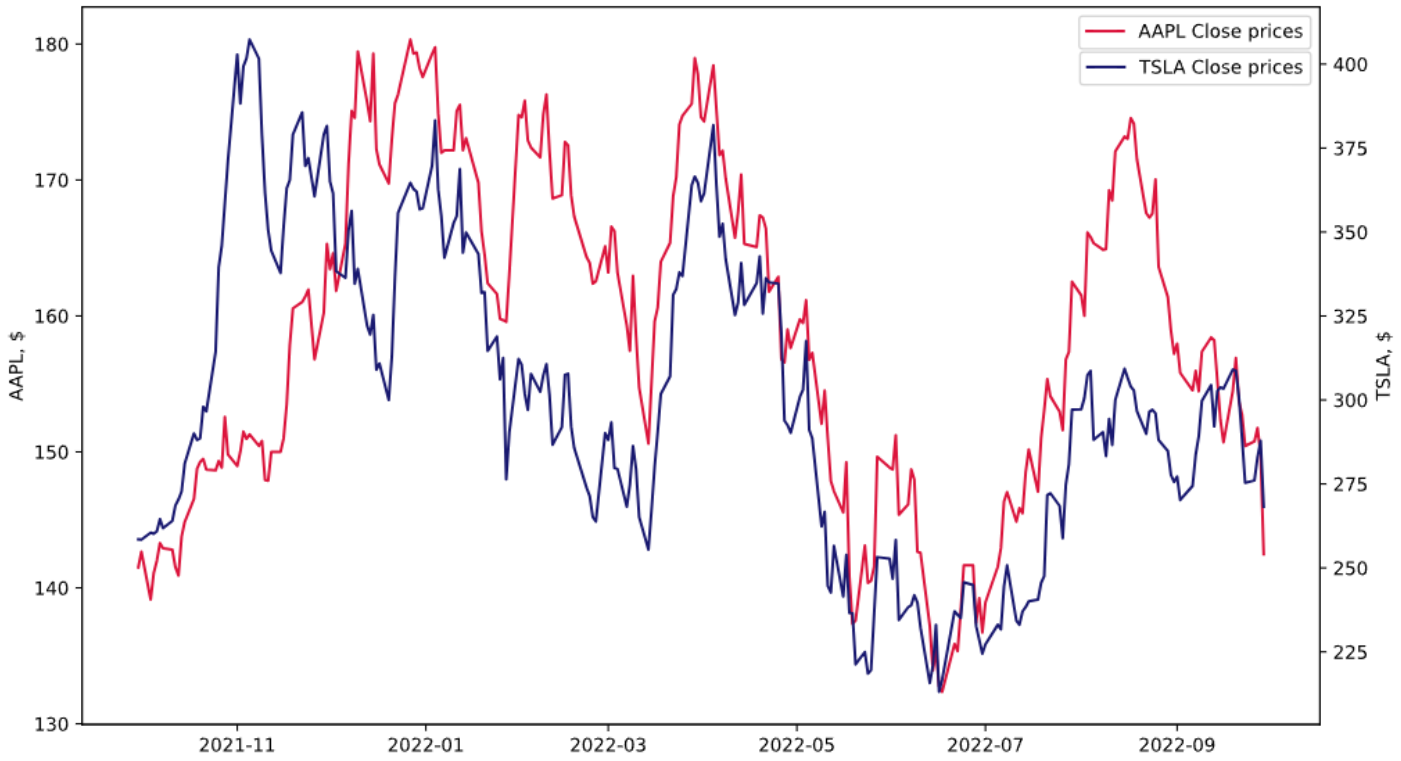

Рис. 3.1. Close ціни на акції TSLA та AAPL в період з 30.09.2021 по 29.09.2022

Джерело: розрахунки автора

Наведені ціни мають доволі сильний розкид і часто змінюють напрямок з падіння на зростання і навпаки. Напрямок руху цін є доволі схожим між TSLA та AAPL. Обидві акції зростають під кінець 2021 року, мають падіння після початку повномасштабного вторгнення Російської Федерації на територію України, стрімко зростають в квітні 2022 року, після цього спадають влітку 2022 року. Врешті-решт в кінці досліджуваного періоду ціни на акції обох компаній приблизно дорівнюють своєму початковому стану.

Найбільших значень ціни на акції компанії Tesla набули 05.11.2021 ($407.32), а найменших – 16.06.2022 ($213.10). Акції Apple коштували найдорожче 27.12.2021 ($180.33), а найдешевше – 17.06.2022 ($132.35).

Дослідження твітів

Щоб додати в моделі фактор почуттів акціонерів, який потенційно може впливати на ціни акцій, був завантажений набір даних [60] з повідомленнями людей щодо цін на акції з соціальної мережі X (раніше Twitter) в період з 30.09.2021 по 29.09.2022.

База даних з «твітами» містить 80793 твітів, які стосуються 25 компаній – з них 37422 повідомлення стосуються Tesla та 5056 – Apple.. Відбір «твітів» до бази даних відбувається за ознакою наявності фінансового хештегу, який стосується відповідної компанії. Для Tesla – це наявність $TSLA в «твіті», а для Apple - $AAPL. Таким чином всі повідомлення, які були зроблені в період з 30.09.2024 по 29.09.2022 та містили фінансовий хештег однієї з 25 компаній, потрапляли до бази даних.

Наступною дією стала обробка «твітів» користувачів стосовно акцій TSLA та AAPL. Для оцінки настроїв користувачів в роботі була використана бібліотека VADER, яка здатна по заданому реченню визначити чи є воно позитивним (0 < оцінка <= 1), нейтральним (оцінка = 0) або негативним (-1 <= оцінка < 0). Таким чином були розраховані оцінки для кожного «твіта». Так як повідомлень користувачів було дуже багато в кожен день дослідження, для того, щоб розрахувати загальну оцінку, усі «твіти» були згруповані по дням, і для кожного з них взяте медіанне значення, щоб визначити загальний настрій користувачів за день (sentiment_score). Після цього ці дані були занесені до загальної бази даних.

Графічно sentiment_score для акцій Tesla зображений на рис. 3.2. Вертикальна лінія всередині зеленого прямокутника означає медіану, вертикальні сторони – 0,25 квантиль (ліва) та 0,75 квантиль (права). «Бакенбарди» (вертикальні риски на кінцях «вус») означають краї вибірки, в середину яких потрапляють значення, які входять у інтервал $[Q_1 - 1{,}5IQR; Q_3 + 1{,}5IQR]$, де:

- Q_1 – 0,25 квантиль вибірки.
- Q_3 – 0,75 квантиль вибірки.
- IQR – міжквартильний розмах, який визначається за формулою $IQR = Q_3 - Q_1$.

Викиди (значення, що не потрапили в межі «бакенбардів») представлені у вигляді ромбів.

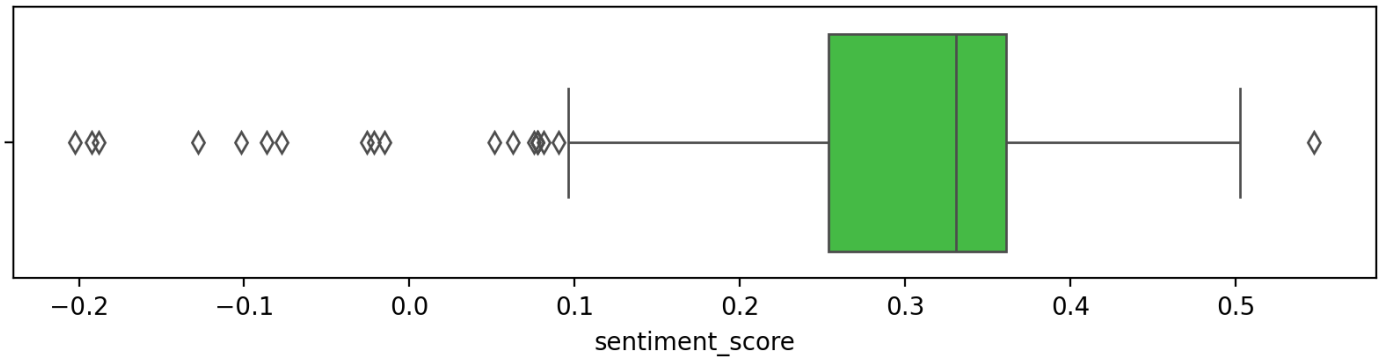

Рис. 3.2. Показник sentiment_score для акцій TSLA, зображений на графіку «ящик з вусами»

Джерело: розрахунки автора

Медіанне значення sentiment_score становить 0,331, 0,25 та 0,75 квантилі дорівнюють 0,254 та 0,361 відповідно. Викидів всього наявно 20 од. – 19 од. ліворуч та 1 од. праворуч. Найчастіше настрої були позитивні, проте іноді все ж були нижчими за 0.

Для того, щоб проаналізувати, чи були викиди sentiment_score зумовленими різкими подіями з цінами на акції, на рис. 3.3. було зображено Open ціни та всі викиди sentiment_score у вигляді точок, де червоні точки уособлюють негативні викиди, а зелені – позитивні.

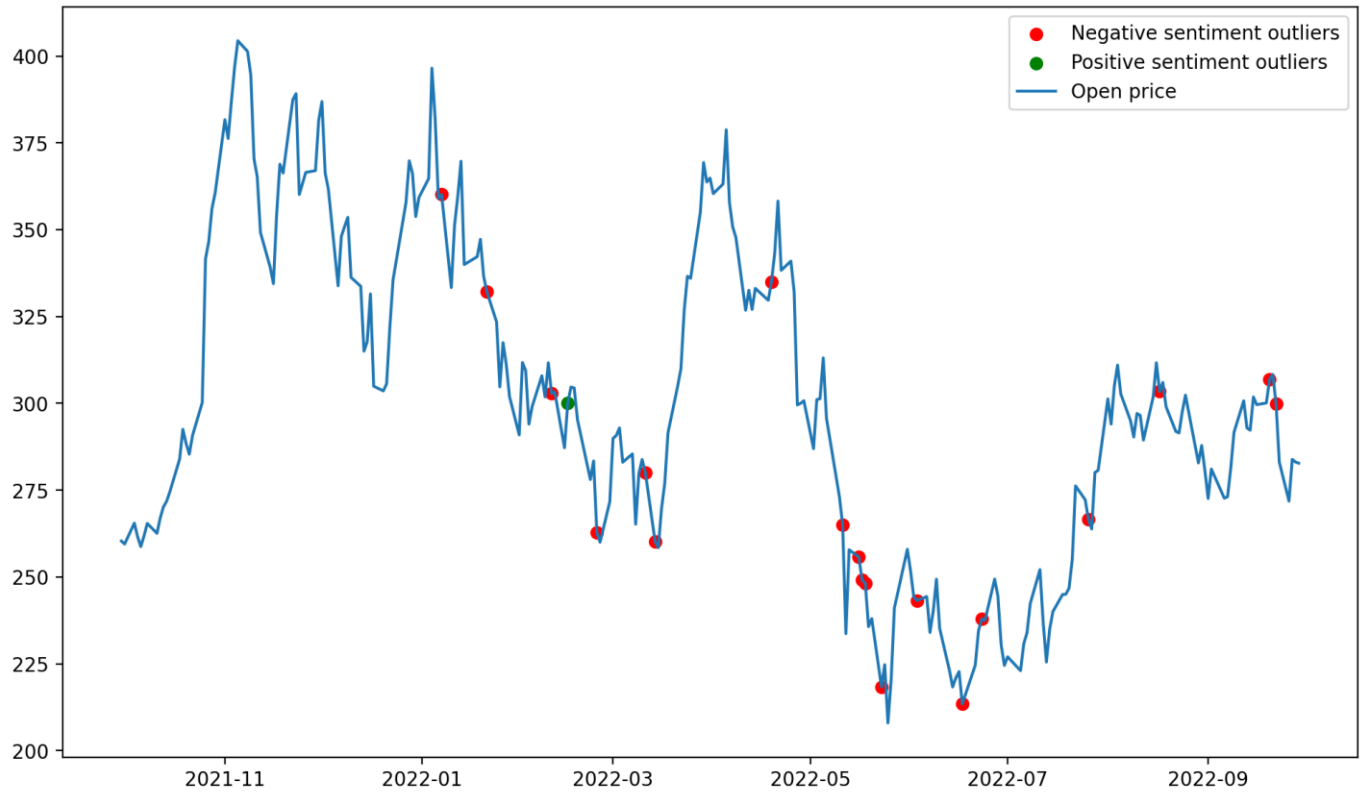

Рис. 3.3. Викиди sentiment_score на графіку Open цін TSLA

Джерело: розрахунки автора

Рис 3.3 демонструє моменти, що заслуговують уваги. Часто в день, коли ціна на акції TSLA досягала локального мінімуму, маємо викиди низьких значень оцінок настроїв – це, наприклад, 14 березня 2022 року та 17 червня 2022 року. Різке падіння в квітні-червні 2022 року, коли було досягнуто найменшого значення цін за весь досліджуваний період, супроводжувалось лише негативними реакціями і у великому обсязі. Також варто зазначити цікавий факт, що друге за негативністю значення настроїв припало на 24.02.2022 – початок повномасштабної війни Росії проти України. В той же день ціни доволі різко впали, досягнувши локального мінімуму 25.02.2022.

Проте, в цілому не можна повним чином стверджувати, що викиди значень sentiment_score для акцій Tesla дуже точно описують поведінку цін. Те саме можна зазначити і для акцій Apple – в деяких моментах прослідковується, що негативні настрої акціонерів співпадають з суттєвим зниженням ціни, проте не завжди.

Обробка викидів

Після того, як були проаналізовані настрої користувачів, почалась обробка викидів в цінових даних. Для цього були проведені 5 ітерацій за допомогою алгоритму Isolation Forest, в кожному з яких використовувались по 2 ознаки:

- High price та Low price
- Open Price за день t - Open Price за день t-1 та Open Price за день t+1 - Open Price за день t
- High Price за день t - High Price за день t-1 та High Price за день t+1 - High Price за день t
- Low Price за день t - Low Price за день t-1 та Low Price за день t+1 - Low Price за день t
- Close Price за день t - Close Price за день t-1 та Close Price за день t+1 - Close Price за день t

Результати цього алгоритму для акції TSLA продемонстровані на рис. 3.4.-3.8.

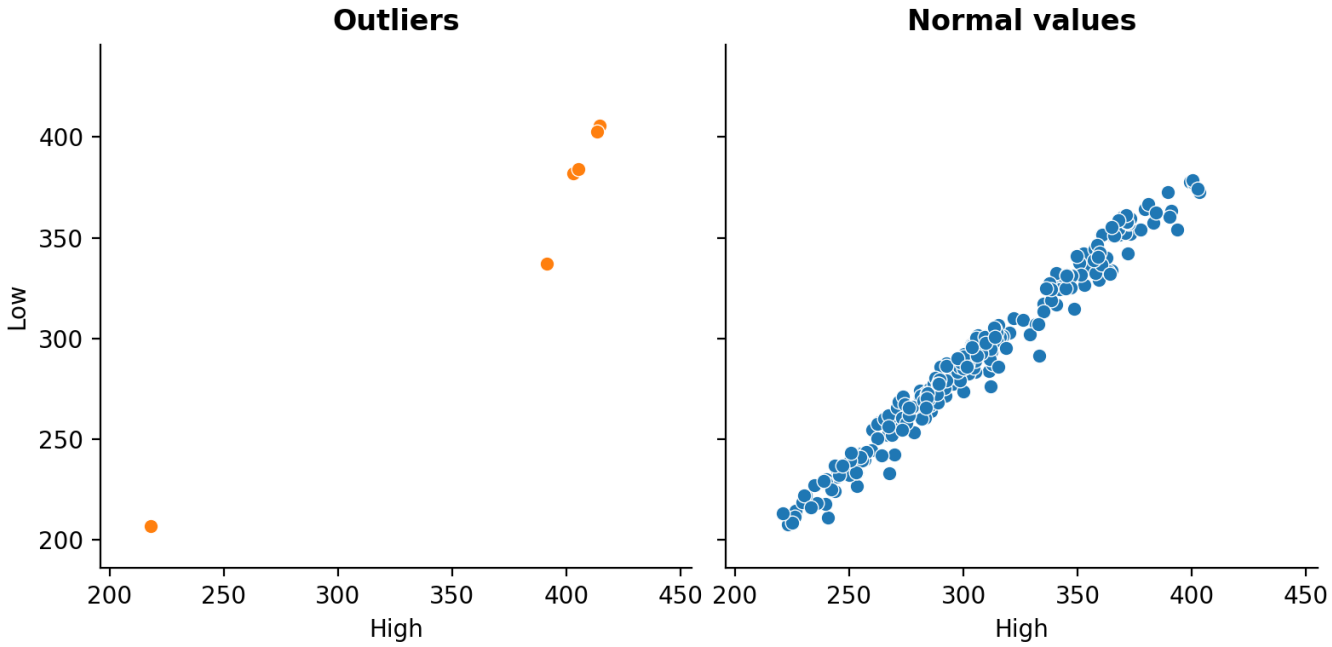

Рис. 3.4. Перевірка на викиди методом Isolation Forest №1 для акцій TSLA

Джерело: розрахунки автора

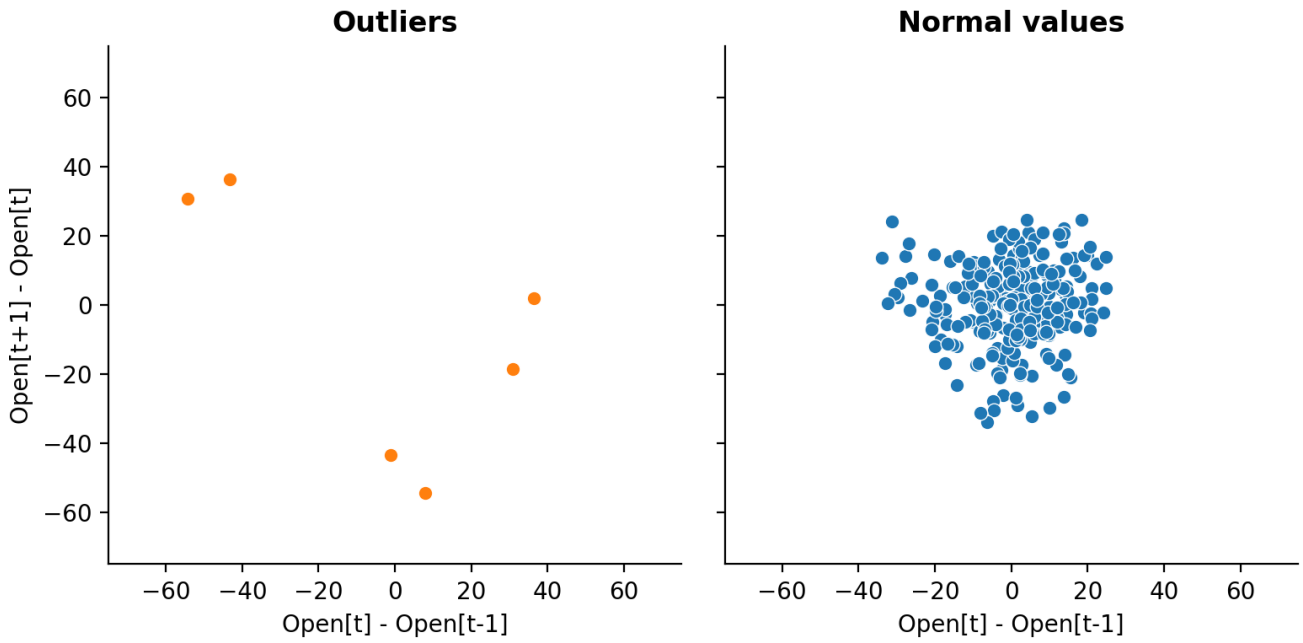

Рис. 3.5. Перевірка на викиди методом Isolation Forest №2 для акцій TSLA

Джерело: розрахунки автора

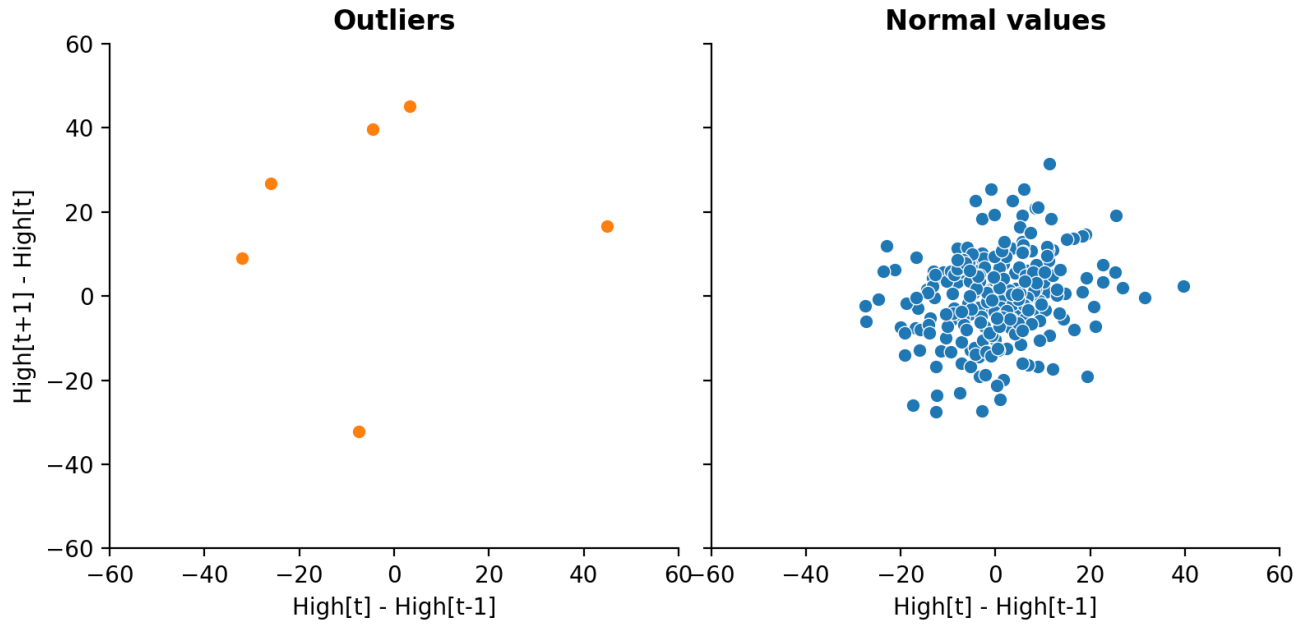

Рис. 3.6. Перевірка на викиди методом Isolation Forest №3 для акцій TSLA

Джерело: розрахунки автора

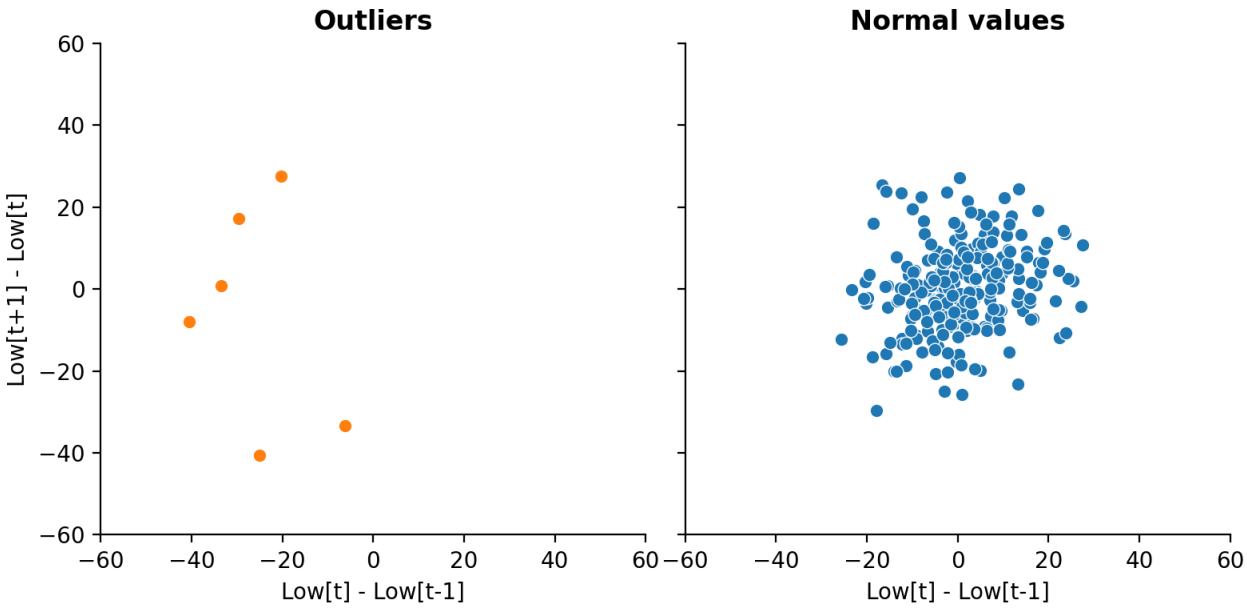

Рис. 3.7. Перевірка на викиди методом Isolation Forest №4 для акцій TSLA

Джерело: розрахунки автора

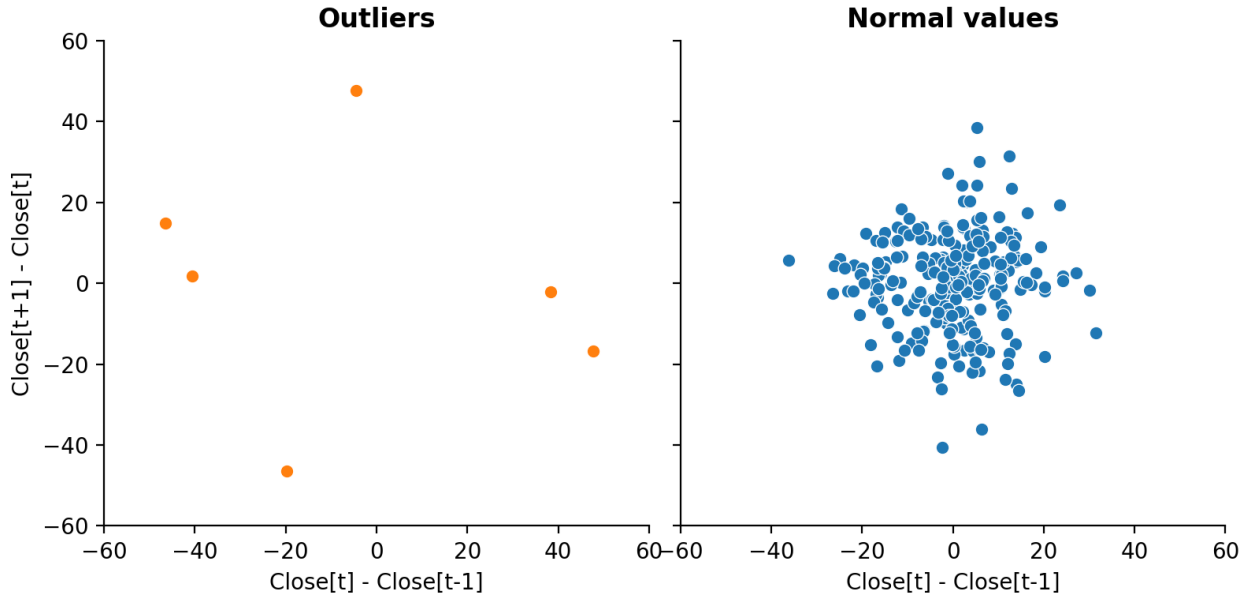

Рис. 3.8. Перевірка на викиди методом Isolation Forest №5 для акцій TSLA

Джерело: розрахунки автора

Як можна побачити, на рис. 3.4. нормальні значення збираються в умовний кластер, що геометрично розташований приблизно вздовж прямої. Значення, які знаходяться на кінцях відрізка, були визначені алгоритмом Isolation Forest як аномальні. На рис. 3.5-3.8

нормальні значення здебільшого збираються в один кластер, який утворює кульовий окіл центроїду. Аномальні значення виходять за межі цих кластерів.

Далі, якщо принаймні один розподіл ознак в день t був класифікований алгоритмом як аномальний, то відповідне P_t було замінене на $\frac{(Pt-2 + Pt-1 + Pt)}{3}$, де P_t – ціна на акцію в день t.

Нормалізація даних

При роботі з моделями BiLSTM, k-NN та Lasso всі тренувальні X_{train} та тестові X_{test} значення були нормалізовані[1]. Для нормалізації тестових значень використовувались отримані параметри нормалізації тренувальної вибірки. Конкретний тип нормалізації виступав гіперпараметром та обирався за допомогою крос-валідації. Всього в пошуку знаходилось 3 варіанти:

- Min-max нормалізація, що переносить всі значення у діапазон від 0 до 1 включно. Розраховуються нормалізовані значення за наступною формулою:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

де x_{min} – мінімальне значення вибірки;

x_{max} – максимальне значення вибірки.

- Standard нормалізація приводить дані до стандартного нормального розподілу з середнім рівним 0 і стандартним відхиленням рівним 1 за наступною формулою:

$$x' = \frac{x - \mu}{\sigma}$$

де μ – середнє вибірки;

σ – стандартним відхилення вибірки.

- Robust нормалізація масштабує дані з урахуванням медіани та міжквартильного розмаху, що робить її стійкою до викидів, за наступною формулою:

$$x' = \frac{x - \mathrm{m}}{\mathrm{IQR}},$$

[1] В подальшому нормалізовані значення додатково будуть позначатись апострофом (’).

де m – медіана вибірки;

IQR – міжквартильний розмах вибірки.

Створення нових факторів

Для покращення прогностичних здатностей моделей були розраховані нові фактори з вже наявних, що вносили б додаткову інформацію для моделей.

- Згладжені значення[2] цін з використанням перетворення Фур'є. Щоб отримати значення цін зі зменшеним впливом «викидів», тренувальні значення по кожній з видів цін були розвинуті (Open, High, Low, Close) в ряд Фур'є за допомогою функції FFT (Швидке перетворення Фур'є), а далі взяті перші n членів ряду. На рис. 3.9 можна побачити графіки Open цін на акції TSLA за увесь період та відповідні апроксимації Фур'є з 2-ма, 8-ма, 12-ма, 16-ма та 32-ма членами.

Рис. 3.9. Апроксимація Фур'є для Open цін на акції TSLA

Джерело: розрахунки автора

[2] В подальшому такі значення в день t будуть позначитись F_t

З графіку видно, що менша кількість компонентів більше згладжує ряд, а більша наближається до початкового графіка.

Для подальшого прогнозування як додаткові фактори були обрані значення з апроксимацій з 12-ти компонентів як для акцій TSLA, так і AAPL як збалансоване значення між згладжуванням та збереженням інформації про швидкі зміни в значеннях. При цьому перші та останні 5 значень апроксимації були замінені на справжні значення цін в день t, так як значення F_t набувають аномально високих та аномально низьких значень спочатку та в кінці відповідно.

- day_change, що являє собою відсоткову зміну вартості акції за день.

$$day\_change_t = \frac{Close\ P_t - Open\ P_t}{Open\ P_t}$$

- day_gap, що являє собою відсотковий розрив вартості акції між максимальним та мінімальним значення за день.

$$day\_gap_t = \frac{High\ P_t - Low\ P_t}{Low\ P_t}$$

За аналогією з sentiment_score, day_change та day_gap зображені графічно на рис. 3.10 та 3.11. Викиди представлені відповідно у вигляді квадратів та зірок.

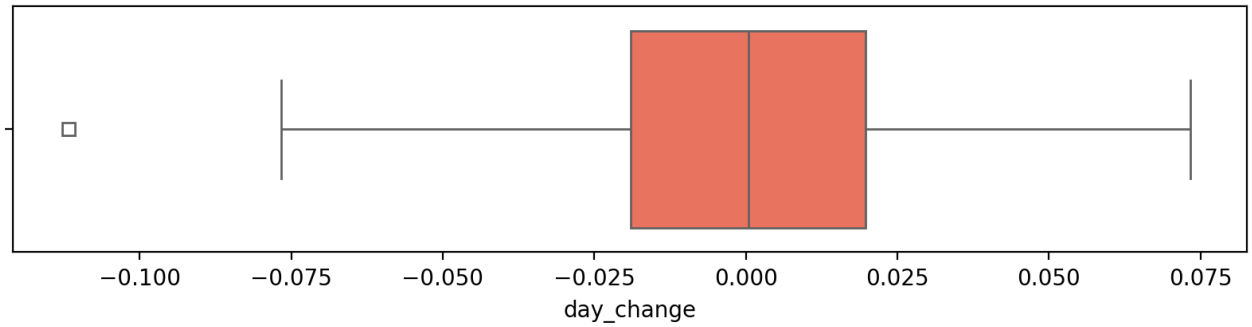

Рис. 3.10. Показник day_change, зображений на графіку «ящик з вусами»

Джерело: розрахунки автора

Аналізуючи рис. 3.10, можна зробити висновки щодо статистичних показників фактору day_change. Медіана вибірки day_change приблизно дорівнює 0. 0,25 та 0,75 квантилі становлять відповідно -0.019 та 0.02. Наявний лише 1 викид – в меншу сторону.

Можна зробити висновок, що за день ціни як зменшувались, так і підвищувались приблизно однаково часто та приблизно з однаковою амплітудою.

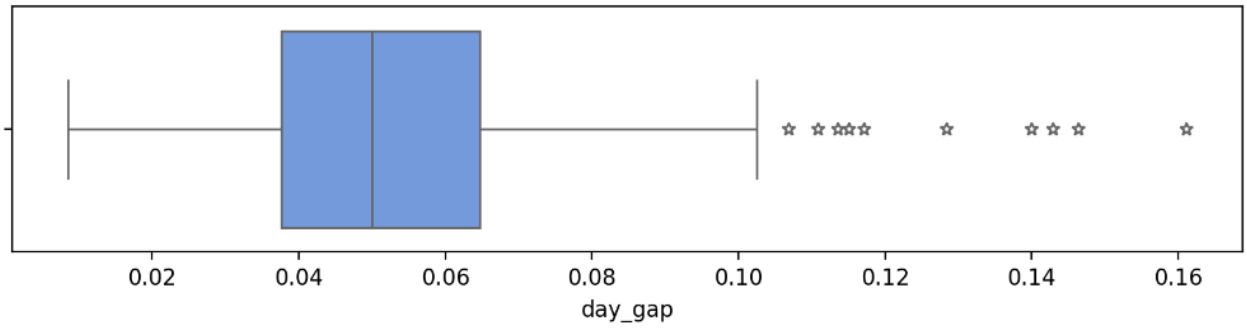

Рис. 3.11. Показник day_gap, зображений на графіку «ящик з вусами»

Джерело: розрахунки автора

Медіана day_gap становить 0,05, 0,25-квантиль дорівнює 0,038, а 0,75-квантиль – 0,066. Викидів наявно 5 одиниць (всі в більшу сторону). Середнє значення day_gap становить 0,053 – це означає, що в середньому максимальне значення ціни на акцію за день було більшим за найменше на 5,3%.

- Лагові значення[3] цін з кроком в 1 день для того, щоб моделі могли враховувати контекст співвідношення сьогоднішньої ціни до вчорашньої .

$$L_t = P_{t-1}$$

- MACD - індикатор імпульсу, який показує різницю між двома експоненційними ковзкими середніми (EMA): короткостроковими – вікно 12 днів та довгостроковими – вікно 26 днів [47]. Його завдання полягає у виявленні змін у тренді ціни. Цей індикатор допомагає виявити точки можливого розвороту тренду. Наприклад, коли MACD перетинає нульову лінію вгору, це може свідчити про початок висхідного тренду, а перетин вниз — про спадний. Це дозволяє моделі враховувати динаміку змін тренду.

$$MACD = EMA_{12}(Close\ P) - EMA_{26}(Close\ P)$$

[3] В подальшому такі значення в день t будуть позначатись L_1

- Signal Line (SL) — це згладжена версія MACD, яка обчислюється як експоненційне згладжене значення MACD з вікном в 9 днів [61]. Вона використовується для порівняння з MACD для генерації сигналів купівлі чи продажу. Сигнальна лінія посилює здатність моделі розпізнавати короткострокові зміни в імпульсі, оскільки перетин MACD і Signal Line вказує на потенційні точки входу чи виходу. Це дозволяє прогнозувати моменти, коли ціна, ймовірно, змінить напрямок.

$$SL = EMA_9(MACD)$$

MACD та сигнальна лінія для акцій TSLA візуалізовані на рис. 3.12.

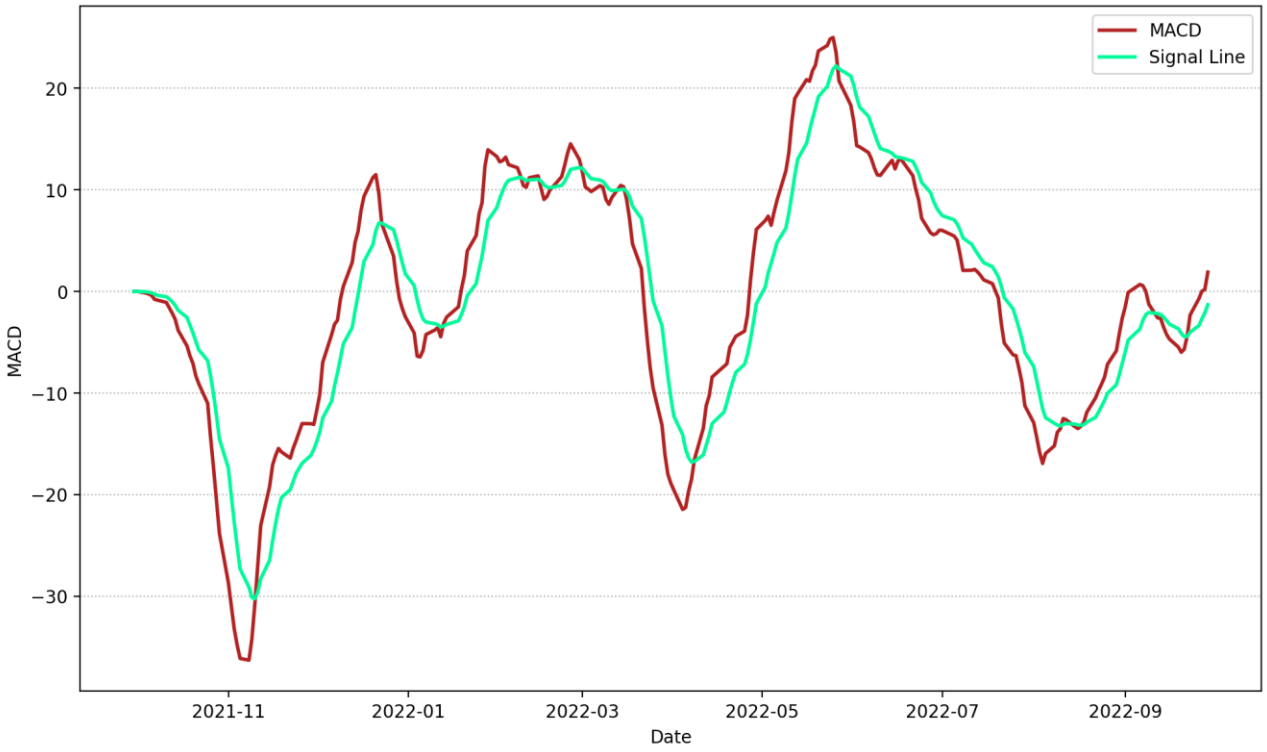

Рис. 3.12. MACD та сигнальна лінія для акцій TSLA

Джерело: розрахунки автора

- Лінії Боліндже́ра (BL) включають значення ковзкого середнього з вікном 20 днів (MA) в двох межах (верхня – upper та нижня – lower), які віддаляються від ковзкого середнього на величину 2-х стандартних відхилень [62]. Ці межі відображають рівень волатильності ціни. Лінії Боліндже́ра допомагають виявити періоди низької або високої волатильності. Коли ціна наближається до верхньої лінії, це може

свідчити про перекупленість активу, а нижня лінія може вказувати на перепроданість. Врахування цих факторів дозволяє моделі прогнозувати можливі розвороти або продовження тренду.

$$BL_{lower} = MA_{20}(Close\ P) - 2\sigma$$

$$BL_{upper} = MA_{20}(Close\ P) + 2\sigma$$

Графічно Close ціни на акції TSLA та лінії Боліндже зображені на рис. 3.13.

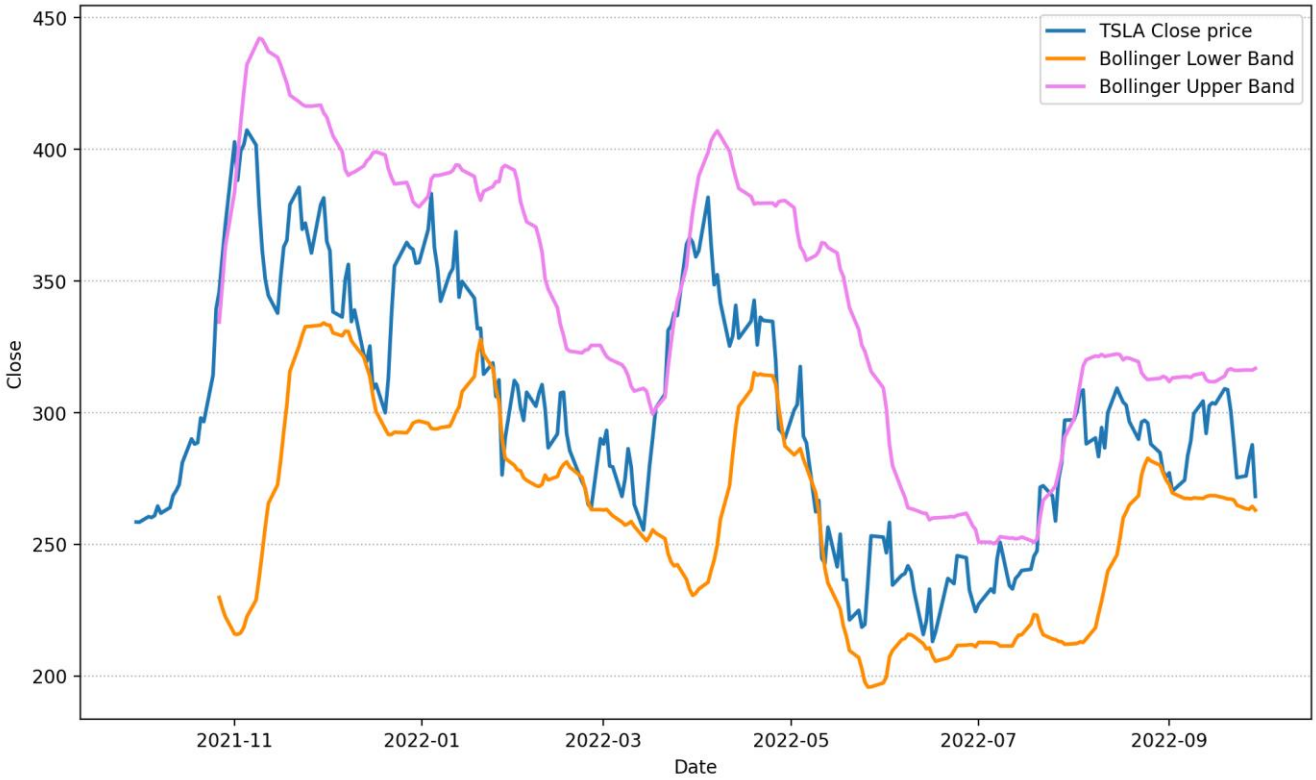

Рис. 3.13. Close ціни на акції TSLA та лінії Боліндже

Джерело: розрахунки автора

3.2. Процес налаштування гіперпараметрів та вхідні дані моделей

Підбір гіперпараметрів

Практично процес підбору гіперпараметрів для моделей, в яких вони наявні, здійснювався наступним чином:

Для BiLSTM підбирались 5 гіперпараметрів (один з яких, norm_layer, визначав архітектуру моделі):

- *scaler* (тип нормалізації даних – min-max, standard або robust)

- *norm_layer* (чи додавати нормалізуючий шар після шару BiLSTM – True або False)
- *units* (скільки нейронів повинно бути в шарі BiLSTM – степені двійки від 2 до 256)
- *dropout* (яку частку нейронів потрібно прирівнювати до 0 після шару LSTM - від 0 до 0.5)
- *learning_rate* (швидкість навчання – дійсні числа від 0.001 до 0.1)

Для k-NN підбирались 4 гіперпараметри:

- *scaler* (тип нормалізації даних – min-max, standard або robust)
- *n_neighbors* (кількість найближчих сусідів – цілі числа від 1 до 20)
- *weights* (чи потрібно зважувати середнє значення ціни сусідів, враховуючи відстань до кожного з них – uniform, якщо не потрібно, та distance, якщо потрібно).
- *p* (параметр потужності для метрики Мінковського – 1 або 2)

Для ARIMA підбирались 3 гіперпараметри:

- p (авторегресійний компонент)
- d (інтегральний компонент)
- q (компонента ковзкого середнього).

Для Lasso підбирались 2 гіперпараметри:

- *scaler* (тип нормалізації даних – min-max, standard або robust)
- *alpha* (коефіцієнт L_1-регуляризації – дійсні числа від 0,001 до 1000)

Для Random Forest підбирались 3 гіперпараметри:

- *max_depth* (максимальна глибина кожного дерева – цілі числа від 1 до 12)
- *min_samples_leaf* (мінімально можлива кількість екземплярів даних на кожному листі дерева – цілі числа від 1 до 3)
- *min_samples_split* (мінімальна кількість екземплярів даних на листі, щоб їх можна було розділяти – цілі числа від 2 до 4).

- *max_features* (частка факторів, що будуть використовуватись при кожному розділенні дерева – дійсні числа 0 до 1)

Для LightGBM підбирались 9 гіперпараметрів:

- *max_depth* (максимальна глибина кожного дерева – цілі числа від 2 до 7)
- *num_leaves* (максимальна кількість листків в дереві – цілі числа від 2 до 31)
- *min_child_samples* (мінімальна сума ваг екземплярів (гессіану) у листку, в задачі регресії – мінімальна кількість екземплярів у листку – цілі числа від 1 до 100)
- *reg_alpha* (коефіцієнт L_1-регуляризації, який додає штраф за абсолютні значення ваг моделі – дійсні числа від 0.001 до 10)
- *reg_lambda* (коефіцієнт L_2-регуляризації, який додає штраф за абсолютні значення ваг моделі – дійсні числа від 0.001 до 10)
- *subsample* (частка випадково вибраних даних для побудови кожного дерева – дійсні числа від 0.25 до 1)
- *subsample_freq* (частота застосування гіперпараметра subsample, раз в скільки дерев використовувати гіперпараметр subsample – від 1 до 10)
- *colsample_bytree* (частка вибраних ознак для побудови кожного дерева – значення від 0.25 до 1)
- *learning_rate* (швидкість навчання – дійсні числа від 0.005 до 0.05)

Для ансамблю підбирались ваги кожної з 4-х моделей, при цьому в сумі вони дорівнювали одиниці:

- *w_knn* (вага моделі k-NN в ансамблі – дійсні числа від 0 до 1)
- *w_lasso* (вага моделі Lasso в ансамблі – дійсні числа від 0 до 1)
- *w_rf* (вага моделі Random Forest в ансамблі – дійсні числа від 0 до 1)
- *w_lgbm* (вага моделі LightGBM в ансамблі – дійсні числа від 0 до 1)

Результати підбору гіперпараметрів для кожного виду цін на акції TSLA висвітлені у табл. 3.1.

Таблиця 3.1

Гіперпараметри моделей

Модель	**Гіперпараметр**	**Open**	**High**	**Low**	**Close**
BiLSTM	scaler	min-max	min-max	min-max	min-max
	norm_layer	True	False	True	False
	units	4	256	2	128
	dropout	0,239	0,407	0,356	0,296
	learning_rate	0,011	0,045	0,03	0,052
KNN	scaler	standard	min-max	standard	min-max
	n_neighbors	3	4	3	5
	weights	uniform	distance	distance	distance
	p	1	1	1	1
ARIMA	p	1	2	1	1
	d	1	1	1	1
	q	0	2	0	0
Lasso	scaler	min-max	min-max	min-max	min-max
	alpha	0,058	0,02	0,039	0,012
Random Forest	max_depth	20	18	22	24
	min_samples_leaf	1	1	1	1
	min_samples_split	2	4	4	2
	max_features	0,42	0,966	0,619	0,957
LightGBM	max_depth	4	2	5	2
	num_leaves	8	30	11	18
	min_child_samples	1	3	3	3
	reg_alpha	0,028	0	0,002	0,744
	reg_lambda	0,01	0,044	0,024	0,013
	subsample	0,63	0,78	0,589	0,598
	subsample_freq	10	10	5	6
	colsample_by_tree	0,956	0,512	0,42	0,66
	learning_rate	0,002	0,002	0,004	0,002
Ensemble	w_knn	0	0	0	0,001
	w_lasso	0,385	0,166	0,201	0
	w_rf	0	0,57	0,55	0,998
	w_lgbm	0,615	0,264	0,249	0,001

Джерело: розрахунки автора

У табл. 3.2. наведена зібрана таблиця «входів» та «виходів» усіх моделей, використаних у дослідженні.

Таблиця 3.2

Показники, використані в моделях

Модель	**Тип**	**Open**	**High**	**Low**	**Close**
BiLSTM	Вхід	Open P', Open F', Open L', day_change', day_gap', MACD', SL', BL_upper', BL_lower', sentiment_score' (t-4, t-3, …, t)	High P', High F', High L', day_change', day_gap', MACD', SL', BL_upper', BL_lower', sentiment_score' (t-4, t-3, …, t)	Low P', Low F', Low L', day_change', day_gap', MACD', SL', BL_upper', BL_lower', sentiment_score' (t-4, t-3, …, t)	Close P', Close F', Close L', day_change', day_gap', MACD', SL', BL_upper', BL_lower', sentiment_score' (t-4, t-3, …, t)
	Вихід	Open $\widehat{P}_{t+1}$	High $\widehat{P}_{t+1}$	Low $\widehat{P}_{t+1}$	Close $\widehat{P}_{t+1}$
k-NN	Вхід	Open P'_t, Open F'_t, Open L'_t, $Volume'_t$ $day\_change'_t$, $day\_gap'_t$, $MACD'_t$, SL'_t, $BL\_upper'_t$, $BL\_lower'_t$, $sentiment\_score'_t$	High P'_t, High F'_t, High L'_t, $Volume'_t$ $day\_change'_t$, $day\_gap'_t$, $MACD'_t$, SL'_t, $BL\_upper'_t$, $BL\_lower'_t$, $sentiment\_score'_t$	Low P'_t, Low F'_t, Low L'_t, $Volume'_t$ $day\_change'_t$, $day\_gap'_t$, $MACD'_t$, SL'_t, $BL\_upper'_t$, $BL\_lower'_t$, $sentiment\_score'_t$	Close P'_t, Close F'_t, Close L'_t, $Volume'_t$ $day\_change'_t$, $day\_gap'_t$, $MACD'_t$, SL'_t, $BL\_upper'_t$, $BL\_lower'_t$, $sentiment\_score'_t$
	Вихід	Open $\widehat{P}_{t+1}$	High $\widehat{P}_{t+1}$	Low $\widehat{P}_{t+1}$	Close $\widehat{P}_{t+1}$
ARIMA	Вхід	(Open) P_1, P_2, …, P_t	(High) P_1, P_2, …, P_t	(Low) P_1, P_2, …, P_t	(Close) P_1, P_2, …, P_t
	Вихід	Open $\widehat{P}_{t+1}$	High $\widehat{P}_{t+1}$	Low $\widehat{P}_{t+1}$	Close $\widehat{P}_{t+1}$
Lasso	Вхід	Open P'_t, Open F'_t, Open L'_t, $Volume'_t$ $day\_change'_t$, $day\_gap'_t$, $MACD'_t$, SL'_t, $BL\_upper'_t$, $BL\_lower'_t$, $sentiment\_score'_t$	High P'_t, High F'_t, High L'_t, $Volume'_t$ $day\_change'_t$, $day\_gap'_t$, $MACD'_t$, SL'_t, $BL\_upper'_t$, $BL\_lower'_t$, $sentiment\_score'_t$	Low P'_t, Low F'_t, Low L'_t, $Volume'_t$ $day\_change'_t$, $day\_gap'_t$, $MACD'_t$, SL'_t, $BL\_upper'_t$, $BL\_lower'_t$, $sentiment\_score'_t$	Close P'_t, Close F'_t, Close L'_t, $Volume'_t$ $day\_change'_t$, $day\_gap'_t$, $MACD'_t$, SL'_t, $BL\_upper'_t$, $BL\_lower'_t$, $sentiment\_score'_t$
	Вихід	Open $\widehat{P}_{t+1}$	High $\widehat{P}_{t+1}$	Low $\widehat{P}_{t+1}$	Close $\widehat{P}_{t+1}$
Random Forest	Вхід	Open P_t, Open F_t, Open L_t, $Volume_t$ day_change_t, day_gap_t, $MACD_t$, SL_t, BL_upper_t, BL_lower_t, $sentiment\_score_t$	High P_t, High F_t, High L_t, $Volume_t$ day_change_t, day_gap_t, $MACD_t$, SL_t, BL_upper_t, BL_lower_t, $sentiment\_score_t$	Low P_t, Low F_t, Low L_t, $Volume_t$ day_change_t, day_gap_t, $MACD_t$, SL_t, BL_upper_t, BL_lower_t, $sentiment\_score_t$	Close P_t, Close F_t, Close L_t, $Volume_t$ day_change_t, day_gap_t, $MACD_t$, SL_t, BL_upper_t, BL_lower_t, $sentiment\_score_t$
	Вихід	Open $\widehat{P}_{t+1}$	High $\widehat{P}_{t+1}$	Low $\widehat{P}_{t+1}$	Close $\widehat{P}_{t+1}$

Продовження табл. 3.2

Модель	**Тип**	**Open**	**High**	**Low**	**Close**
LightGBM	Вхід	Open P_t, Open F_t, Open L_t, $Volume_t$ day_change_t, day_gap_t, $MACD_t$, SL_t, BL_upper_t, BL_lower_t, $sentiment\_score_t$	High P_t, High F_t, High L_t, $Volume_t$ day_change_t, day_gap_t, $MACD_t$, SL_t, BL_upper_t, BL_lower_t, $sentiment\_score_t$	Low P_t, Low F_t, Low L_t, $Volume_t$ day_change_t, day_gap_t, $MACD_t$, SL_t, BL_upper_t, BL_lower_t, $sentiment\_score_t$	Close P_t, Close F_t, Close L_t, $Volume_t$ day_change_t, day_gap_t, $MACD_t$, SL_t, BL_upper_t, BL_lower_t, $sentiment\_score_t$
	Вихід	Open $\widehat{P}_{t+1}$	High $\widehat{P}_{t+1}$	Low $\widehat{P}_{t+1}$	Close $\widehat{P}_{t+1}$
Ensemble	Вхід	(k-NN) Open $\widehat{P}_{t+1}$, (Random Forest) Open $\widehat{P}_{t+1}$, (Lasso) Open $\widehat{P}_{t+1}$, (LightGBM) Open $\widehat{P}_{t+1}$	(k-NN) High $\widehat{P}_{t+1}$, (Random Forest) High $\widehat{P}_{t+1}$, (Lasso) High $\widehat{P}_{t+1}$, (LightGBM) High $\widehat{P}_{t+1}$	(k-NN) Low $\widehat{P}_{t+1}$, (Random Forest) Low $\widehat{P}_{t+1}$, (Lasso) Low $\widehat{P}_{t+1}$, (LightGBM) Low $\widehat{P}_{t+1}$	(k-NN) Close $\widehat{P}_{t+1}$, (Random Forest) Close $\widehat{P}_{t+1}$, (Lasso) Close $\widehat{P}_{t+1}$, (LightGBM) Close $\widehat{P}_{t+1}$
	Вихід	Open $\widehat{P}_{t+1}$	High $\widehat{P}_{t+1}$	Low $\widehat{P}_{t+1}$	Close $\widehat{P}_{t+1}$

Джерело: розрахунки автора

3.3. Формування прогнозів. Оцінка та порівняння ефективності моделей

В дослідженні вибірка даних розділялась на 2 частини: тренувальну (дані в період з 30.09.2021 по 31.08.2022 – 232 екземпляри даних) та тестову (дані в період з 01.09.2022 по 29.09.2022 – 20 екземплярів даних). Після обробки даних гіперпараметри моделей підбирались за допомогою крос-валідації на тренувальних даних, на них же далі навчались моделі. Потім на основі тестових даних будувались прогнози, які порівнювались з реальними значеннями.

На рис. 3.14-3.17 зображені реальні ціни на акції TSLA (усі види: Open, High, Low, Close) та відповідні прогнози кожної з моделей.

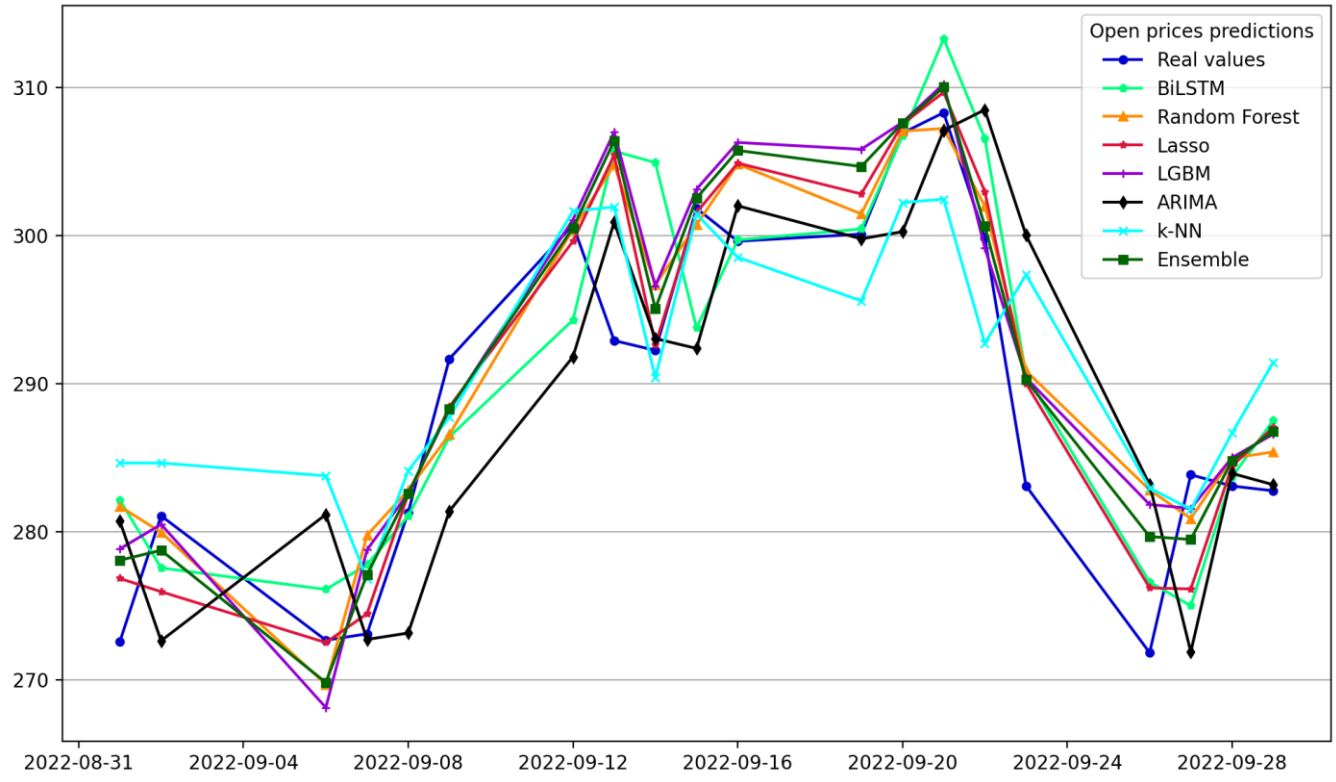

Рис. 3.14. Open ціни на акції TSLA та відповідні прогнози усіма методами

Джерело: розрахунки автора

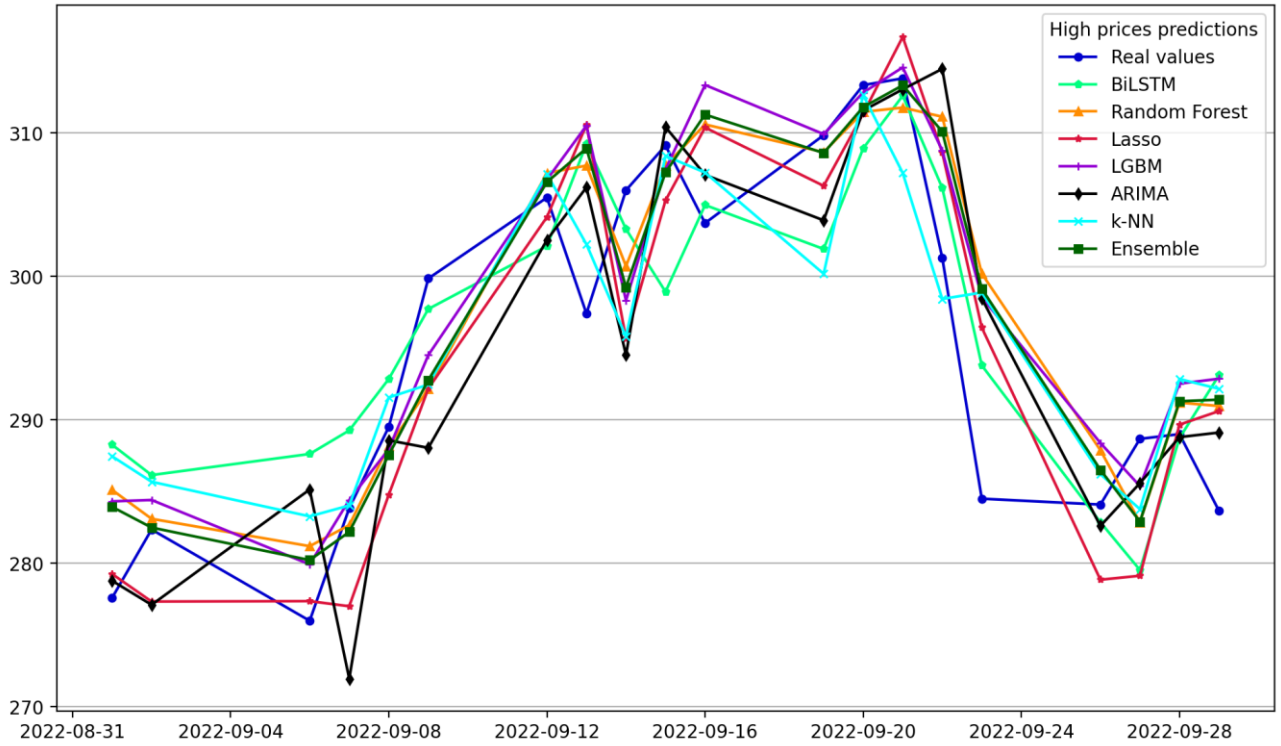

Рис. 3.15. High ціни на акції TSLA та відповідні прогнози усіма методами

Джерело: розрахунки автора

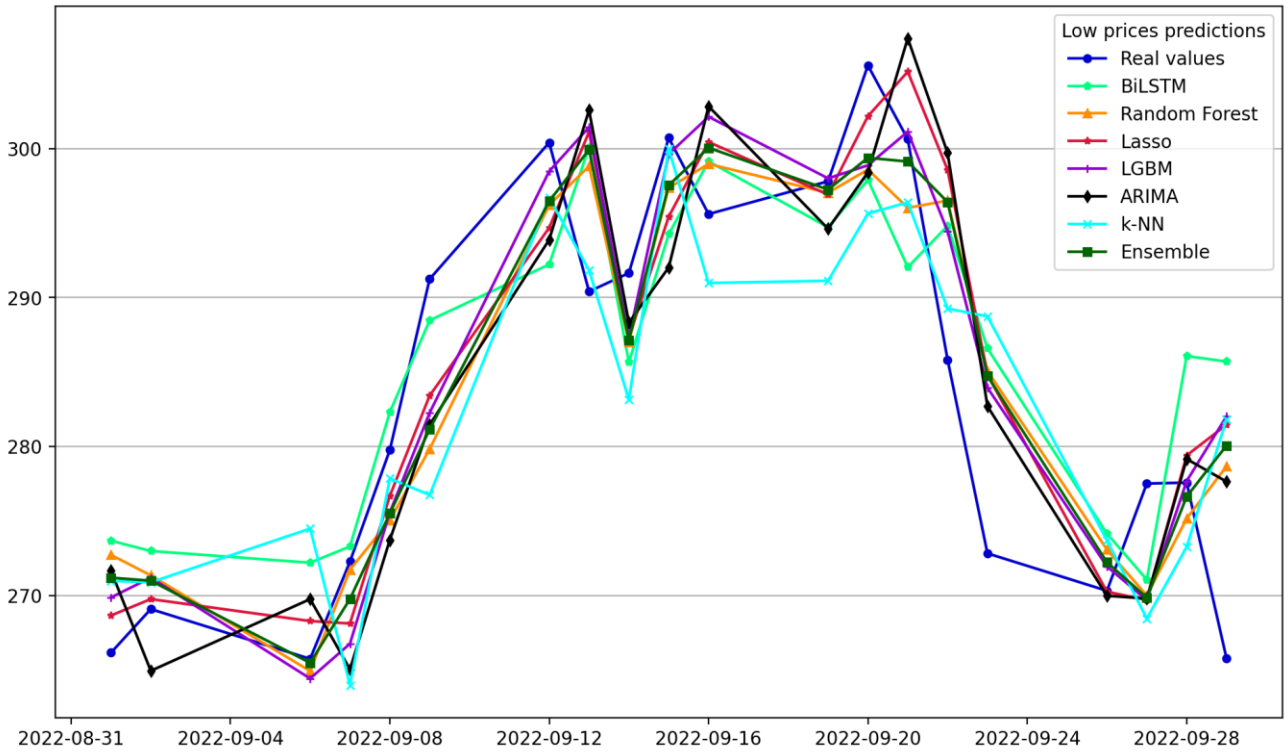

Рис. 3.16. Low ціни на акції TSLA та відповідні прогнози усіма методами

Джерело: розрахунки автора

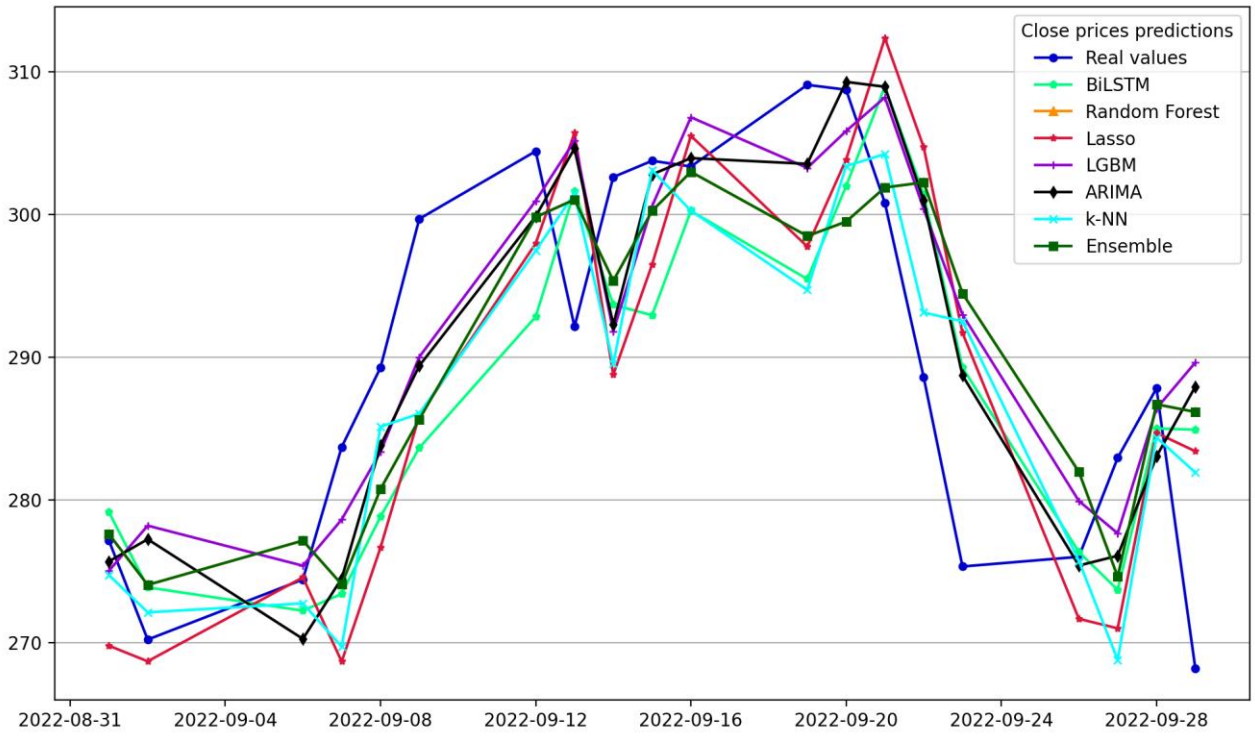

Рис. 3.17. Close ціни на акції TSLA та відповідні прогнози усіма методами

Джерело: розрахунки автора

Щоб перевірити якість прогнозів кожною моделлю, використовувались показники MAPE, MSE, RMSPE. На першому етапі були обраховані похибки моделей на тренувальній вибірці. Результати подано у табл. 3.3-3.5.

Таблиця 3.3

Похибки тренування моделей на даних цін на акції TSLA (MAPE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	1,5097%	1,8528%	2,0652%	2,3196%	1,9368%
k-NN	1,6989%	0,0%	0,0%	0,0%	0,4247%
ARIMA	3,2732%	2,6713%	2,8509%	3,1122%	2,9769%
Lasso	1,5188%	1,7912%	2,0409%	2,4200%	1,9427%
Random Forest	1,2561%	1,2948%	1,0680%	1,7923%	1,3528%
LightGBM	**0,5762%**	**1,2082%**	**0,2007%**	**1,7459%**	**0,9328%**
Ensemble	0,8913%	1,3114%	1,0025%	1,8611%	1,2666%

Джерело: розрахунки автора

Таблиця 3.4

Похибки тренування моделей на даних цін на акції TSLA (MSE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	39,2017	52,9881	56,8510	76,2585	56,3248
k-NN	41,9933	0,0	0,0	0,0	10,4983
ARIMA	141,9487	102,0373	103,5606	134,9679	120,6286
Lasso	39,9650	51,2579	56,6908	82,7249	57,6596
Random Forest	26,0028	27,8092	14,9611	47,9064	29,1699
LightGBM	**4,3580**	**22,3257**	**0,5182**	**41,9230**	**17,2812**
Ensemble	12,4715	27,1066	13,1377	50,7973	25,8783

Джерело: розрахунки автора

Таблиця 3.5

Похибки тренування моделей на даних цін на акції TSLA (RMSPE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	2,1015%	2,2971%	2,6013%	2,9346%	2,4836%
k-NN	2,1501%	0,0%	0,0%	0,0%	0,5375%
ARIMA	4,1097%	3,3684%	3,6229%	4,0479%	3,7872%
Lasso	2,0874%	2,2955%	2,6410%	3,0553%	2,5198%
Random Forest	1,6875%	1,6758%	1,3367%	2,3109%	1,7527%
LightGBM	**0,7118%**	**1,5250%**	**0,2537%**	**2,1992%**	**1,1724%**
Ensemble	1,1795%	1,6673%	1,2618%	2,3773%	1,6215%

Джерело: розрахунки автора

Нульові похибки тренування k-NN пов’язані з принципом функціонування даної моделі при використанні гіперпараметра weights рівному distance. Модель, шукаючи найближчих сусідів для будь-якого екземпляра з тренувальної вибірки буде завжди знаходити той самий екземпляр в якості одного з сусідів при будь-якому гіперпараметрі n_neighbors. Так як відстань до однакової точки дорівнює 0, найближчому сусіду (тій самій точці) буде надаватись нескінченна вага, а отже прогноз для обраної точки буде завжди дорівнювати значенню цієї точки, що буде давати нульове відхилення будь-якої метрики.

Не враховуючи k-NN, найменші похибки тренування з відривом має модель LightGBM ($MAPE_{avg}$ = 0,9328%). На другому місці знаходиться ансамбль моделей ($MAPE_{avg}$ = 1,6215%). Далі йде Random Forest ($MAPE_{avg}$ = 1,3528%). Приблизно

однакові результати мають моделі BiLSTM та Lasso ($MAPE_{avg}$ = 1,9368% та 1,9427% відповідно). На останньому місці знаходиться ARIMA з $MAPE_{avg}$ = 2,9769%.

На другому етапі були пораховані похибки прогнозів (між тестовими даними та реальними цінами за останні 20 днів). Результати відображені у табл. 3.6-3.8.

Таблиця 3.6

Похибки тестування моделей на даних цін на акції TSLA (MAPE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	1,8294%	1,9563%	2,4649%	2,9721%	2,3057%
k-NN	1,9779%	1,7863%	2,3491%	2,5440%	2,1643%
ARIMA	2,2955%	1,9427%	2,4120%	**2,4197%**	2,2675%
Lasso	**1,1727%**	1,9159%	1,9553%	3,2576%	2,0754%
Random Forest	1,4152%	1,6942%	1,9657%	2,6239%	1,9248%
LightGBM	1,4526%	1,6553%	**1,8438%**	2,5011%	1,8632%
Ensemble	1,3196%	**1,6255%**	1,8634%	2,6233%	**1,8580%**

Джерело: розрахунки автора

Таблиця 3.7

Похибки тестування моделей на даних цін на акції TSLA (MSE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	42,2595	47,0774	65,8712	97,6486	63,2142
k-NN	47,6708	41,5663	64,5732	84,4948	59,5763
ARIMA	65,0217	53,7147	58,9682	**73,1361**	62,7102
Lasso	**20,5071**	44,3632	48,4098	116,1320	57,3530
Random Forest	28,1365	39,6121	45,2509	86,5032	49,8757
LightGBM	29,1911	40,2386	45,3858	79,7333	48,6372
Ensemble	23,7987	**37,7830**	**43,6881**	86,4630	**47,9332**

Джерело: розрахунки автора

Таблиця 3.8

Похибки тестування моделей на даних цін на акції TSLA (RMSPE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	2,2522%	2,3668%	2,9200%	3,4099%	2,7372%
k-NN	2,4515%	2,2119%	2,8809%	3,1917%	2,6840%
ARIMA	2,8232%	2,5033%	2,7055%	**3,0238%**	2,7640%
Lasso	**1,5742%**	2,2672%	2,4880%	3,7397%	2,5173%
Random Forest	1,8818%	2,1649%	2,3975%	3,2682%	2,4281%
LightGBM	1,8944%	2,1728%	2,4183%	3,1719%	2,4143%
Ensemble	1,7038%	**2,1070%**	**2,3641%**	3,2675%	**2,3606%**

Джерело: розрахунки автора

Похибки тестування суттєво відрізняються від похибок тренування. Так, при тестуванні найменші похибки показав ансамбль моделей ($MAPE_{avg}$ = 1,8580%). Ненабагато більші похибки має модель LightGBM ($MAPE_{avg}$ = 1,8632%). На третьому місці знаходиться модель Random Forest ($MAPE_{avg}$ = 1,9248%). Наступними йдуть Lasso та k-NN ($MAPE_{avg}$ = 2,0754% та 2,1643% відповідно). Найбільші похибки демонструють моделі ARIMA та BiLSTM ($MAPE_{avg}$ = 2,2675% та 2,3057% відповідно).

Також важливо дослідити, як вплинув показник настроїв користувачів на якість прогнозів. З цією метою, прогнози були повторені, але в якості вхідних даних вже не передавався фактор sentiment_score[4]. Результати подані у табл. 3.9-3.11.

[4] В табл. 3.9-3.11 не наявні результати моделі ARIMA, так як вона в якості вхідних даних приймає лише ціни, без sentiment_score.

Таблиця 3.9

Похибки тестування моделей на даних цін на акції TSLA без настроїв користувачів у якості вхідного параметра (MAPE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	1,8411%	1,9820%	2,5379%	3,0212%	2,3456%
k-NN	1,9287%	**1,6374%**	2,3410%	2,6652%	2,1431%
Lasso	**1,2689%**	1,9713%	2,0451%	3,2683%	2,1384%
Random Forest	1,4245%	1,6634%	1,9866%	2,6276%	1,9255%
LightGBM	1,4598%	1,6582%	**1,8314%**	**2,5197%**	**1,8673%**
Ensemble	1,3844%	1,7116%	1,9588%	2,6277%	1,9206%

Джерело: розрахунки автора

Таблиця 3.10

Похибки тестування моделей на даних цін на акції TSLA без настроїв користувачів у якості вхідного параметра (MSE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	43,6634	50,0508	67,7646	99,1285	65,1518
k-NN	45,4739	39,4320	63,4669	83,3409	57,9284
Lasso	**21,5061**	46,5722	50,4053	120,8419	59,8314
Random Forest	28,2256	**38,1706**	45,4794	87,5078	49,8459
LightGBM	29,9720	40,0629	**44,7424**	**79,3669**	**48,5361**
Ensemble	26,6280	40,0134	46,2850	87,4955	50,1055

Джерело: розрахунки автора

Таблиця 3.11

Похибки тестування моделей на даних цін на акції TSLA без настроїв користувачів у якості вхідного параметра (RMSPE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	2,3340%	2,4717%	2,9305%	3,4416%	2,7945%
k-NN	2,3511%	2,1903%	2,8518%	3,2132%	2,6516%
Lasso	**1,6941%**	2,3122%	2,5179%	3,7789%	2,5758%
Random Forest	1,8847%	**2,1260%**	2,4085%	3,2847%	2,4260%
LightGBM	1,8934%	2,1650%	**2,3812%**	**3,2054%**	**2,4113%**
Ensemble	1,8147%	2,1661%	2,4236%	3,2845%	2,4222%

Джерело: розрахунки автора

Аналізуючи наведені результати, можна побачити, що ефективність фактору sentiment_score залежить від моделі. Так, для BiLSTM, Lasso та ансамблю моделей використання настроїв акціонерів зменшило похибки. Для інших моделей похибки або незначно збільшились, або майже не змінились.

Найвищі похибки тестування моделі демонструють для Close цін (MAPE від 2,4197% до 3,2576%) – це сигналізує про схильність до різких змін наприкінці торгової сесії, в той же час Open ціни демонструють нижчі похибки (MAPE від 1,1727% до 2,2955%), що є індикатором меншої схильності до короткострокових ринкових потрясінь на початку торгового дня.

Загалом модель BiLSTM добре вловлює послідовні залежності, але є чутливою до шуму та перенавчання. Вища похибка має місце в першу чергу через невелику кількість даних для тренування, адже BiLSTM є комплексною рекурентною нейронною мережею, яка потребує великої кількості тренувальних даних для якісного моделювання справжнього розподілу даних.

k-NN демонструє середні результати. Враховуючи принцип функціонування моделі, це означає що вона здатна вловлювати схожості між історичними спостереженнями на основі схожості вхідних даних, проте волатильність акцій TSLA та непараметричність архітектури знижують її результати.

Моделі ARIMA та Lasso передбачають лінійність взаємозв’язків між даними. Їх посередні результати є індикатором, що справжній розподіл даних цін на акції Tesla має природу відмінну від лінійного, через що лінійні обмеження моделей підвищують їх похибки прогнозування.

Сильні результати ансамблю моделей, LightGBM та Random Forest свідчать про те, що комбінування моделей здатне вловлювати різні взаємозв’язки в даних та, агрегуючи результат, приходити до якісного прогнозу. Модель LightGBM, що традиційно вважається однією з найсильніших моделей для табличних даних і гнучка у моделюванні складних взаємозалежностей, знаходиться на 2-му місці за результатами. Ансамбль моделей має найнижчі похибки з усіх моделей. Він агрегує кілька різних за своєю природою моделей і тому може вловлювати різні за своєю суттю залежності та патерни, демонструючи високі результати.

Для додаткової валідації використаних підходів та моделей, були розглянуті результати прогнозування цін на акції AAPL.

Процес обробки даних акцій та твітів AAPL, налаштування гіперпараметрів та тренування моделей є аналогічним до представленого для акцій TSLA.

Метрики прогнозування тренувальної вибірки з використанням даних про настрої акціонерів зображені у табл. 3.12-3.14.

Таблиця 3.12

Похибки тренування моделей на даних цін на акції AAPL (MAPE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	2,0891%	2,2118%	1,0210%	2,1386%	1,8651%
k-NN	0,8579%	0,0%	0,0%	0,0%	0,2145%
ARIMA	1,5620%	1,2550%	1,3234%	1,5155%	1,4140%
Lasso	0,7762%	0,9016%	0,9600%	1,3063%	0,9860%
Random Forest	**0,3247%**	0,6642%	0,5251%	1,0260%	0,6350%
LightGBM	0,5195%	**0,0006%**	**0,0105%**	**0,7038%**	**0,3086%**
Ensemble	0,4227%	0,5658%	0,2364%	1,0541%	0,5698%

Джерело: розрахунки автора

Таблиця 3.13

Похибки тренування моделей на даних цін на акції AAPL (MSE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	14,4192	16,0846	4,034	17,2422	12,945
k-NN	3,0407	0,0	0,0	0,0	0,7602
ARIMA	10,1104	6,7442	7,0425	9,4455	8,3357
Lasso	2,5559	3,4456	3,779	6,6531	4,1084
Random Forest	**0,4443**	1,8867	1,0258	4,007	1,841
LightGBM	1,1245	**0,0001**	**0,0005**	**1,8929**	**0,7545**
Ensemble	0,7513	1,3616	0,2086	4,1901	1,6279

Джерело: розрахунки автора

Таблиця 3.14

Похибки тренування моделей на даних цін на акції AAPL (RMSPE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	2,3204%	2,4531%	1,2811%	2,5099%	2,1411%
k-NN	1,1147%	0,0%	0,0%	0,0%	0,2787%
ARIMA	1,9915%	1,6011%	1,7057%	1,9316%	1,8075%
Lasso	1,0166%	1,1536%	1,2483%	1,6149%	1,2584%
Random Forest	**0,4257%**	0,8534%	0,6538%	1,2568%	0,7974%
LightGBM	0,6742%	**0,0008%**	**0,0145%**	**0,8728%**	**0,3906%**
Ensemble	0,5497%	0,7233%	0,2938%	1,2822%	0,7123%

Джерело: розрахунки автора

Так як тільки для Open цін найкращим значенням гіперпараметра weights було визначено uniform, лише там похибки тренування не дорівнюють 0.

Якщо не враховувати результати моделі k-NN, найменші похибки тренування демонструє LightGBM (MAPE avg = 0,3904%). За рахунок своєї комбінованої архітектури вона майже ідеально прогнозує тренувальні значення цін High та Low. На 2-му та 3-му місці відповідно знаходяться Ансамбль (MAPE avg = 0,7123%) та Random Forest (MAPE avg = 0,7974%). На 4-му та 5-му місці знаходяться моделі Lasso (MAPE avg = 1,2584%) та ARIMA (MAPE avg = 1,8075%). Через лінійність взаємозв'язків в цих моделях вони мають схильність більше згладжувати результати, а не моделювати крайні випадки, через що мають вищу похибку тренування. На останньому місці можна побачити модель BiLSTM (MAPE avg = 2,1411%). Високі похибки тренування означають, що двонаправлена LSTM не мала достатньо даних, щоб повністю реалізувати свою комплексну архітектуру.

Розглянемо результати прогнозування тестової вибірки з використанням фактору настроїв акціонерів.

Таблиця 3.15

Похибки тестування моделей на даних цін на акції AAPL (MAPE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	1,5814%	1,6142%	1,4575%	1,7198%	1,5932%
k-NN	1,4225%	1,7170%	1,4008%	1,5778%	1,5295%
ARIMA	1,6394%	1,4556%	1,1686%	**1,3924%**	1,4140%
Lasso	**0,8166%**	**1,2366%**	**1,1165%**	1,4450%	**1,1537%**
Random Forest	1,2288%	1,5939%	1,5841%	1,7530%	1,5400%
LightGBM	1,1524%	1,7854%	1,6405%	1,5868%	1,5413%
Ensemble	1,1809%	1,6260%	1,6116%	1,7477%	1,5416%

Джерело: розрахунки автора

Таблиця 3.16

Похибки тестування моделей на даних цін на акції AAPL (MSE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	7,4552	10,6877	7,4893	12,1496	9,4454
k-NN	5,9174	9,0639	6,0902	8,0103	7,2704
ARIMA	8,4912	7,0442	**4,8277**	**6,8856**	6,8122
Lasso	**2,4899**	**4,7409**	5,0291	6,9323	**4,798**
Random Forest	5,0829	8,6613	8,8769	10,4611	8,2706
LightGBM	4,424	10,4406	8,8879	8,6376	8,0975
Ensemble	4,6203	8,828	8,6091	10,3909	8,1121

Джерело: розрахунки автора

Таблиця 3.17

Похибки тестування моделей на даних цін на акції AAPL (RMSPE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	1,7834%	2,0766%	1,8581%	2,2641%	1,9956%
k-NN	1,5900%	1,9465%	1,6638%	1,8587%	1,7648%
ARIMA	1,9023%	1,7219%	**1,4781%**	1,7589%	1,7153%
Lasso	**1,0447%**	**1,4256%**	1,5296%	**1,7475%**	**1,4369%**
Random Forest	1,4813%	1,9072%	2,0201%	2,1348%	1,8858%
LightGBM	1,3788%	2,0977%	2,0017%	1,9394%	1,8544%
Ensemble	1,4112%	1,9266%	1,9794%	2,1277%	1,8612%

Джерело: розрахунки автора

Перейдемо до результатів прогнозування тестової вибірки без використання фактору сентименту акціонерів.

Таблиця 3.18

Похибки тестування моделей на даних цін на акції AAPL без настроїв користувачів у якості вхідного параметра (MAPE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	1,3599%	2,3486%	1,3854%	1,9201%	1,7535%
k-NN	1,4174%	1,7034%	1,3463%	1,6232%	1,5226%
Lasso	**0,8112%**	**1,2366%**	**1,1142%**	**1,4450%**	**1,1518%**
Random Forest	1,3132%	1,5399%	1,5526%	1,7103%	1,5290%
LightGBM	1,1586%	1,7971%	1,5720%	1,6144%	1,5355%
Ensemble	1,2252%	1,5671%	1,5527%	1,7066%	1,5129%

Джерело: розрахунки автора

Таблиця 3.19

Похибки тестування моделей на даних цін на акції AAPL без настроїв користувачів у якості вхідного параметра (MSE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	6,5534	17,8521	6,5136	13,7833	11,1756
k-NN	6,6867	8,7308	6,8633	8,9602	7,8102
Lasso	**2,4764**	**4,7405**	**4,9388**	**6,9323**	**4,772**
Random Forest	6,1557	8,5386	9,0395	9,9856	8,4298
LightGBM	4,4606	11,0908	8,4905	9,1577	8,2999
Ensemble	5,1334	8,7563	8,1747	9,9384	8,0007

Джерело: розрахунки автора

Таблиця 3.20

Похибки тестування моделей на даних цін на акції AAPL без настроїв користувачів у якості вхідного параметра (RMSPE)

Модель	Open	High	Low	Close	У середньому
BiLSTM	1,6742%	2,7193%	1,7250%	2,3923%	2,1277%
k-NN	1,6979%	1,9164%	1,7733%	1,9653%	1,8382%
Lasso	**1,0416%**	**1,4256%**	**1,5156%**	**1,7475%**	**1,4326%**
Random Forest	1,6364%	1,8932%	2,0411%	2,0909%	1,9154%
LightGBM	1,3887%	2,1674%	1,9560%	2,0057%	1,8794%
Ensemble	1,4928%	1,9193%	1,9307%	2,0861%	1,8572%

Джерело: розрахунки автора

Для набору даних цін на акції AAPL при використанні фактору настроїв акціонерів результати покращились для моделі BiLSTM.

Загалом як при використанні фактору сентименту, так і без нього, найкращі результати з відривом демонструє модель Lasso. На другому місці знаходиться економетрична модель ARIMA. Схожі результати мають моделі k-NN, Random Forest, LightGBM. На останньому місці по результатам йде модель BiLSTM.

Такі результати мають місце по 2-м причинам:

- Дані з цього набору мають взаємозв’язок, подібний до лінійного. Це прослідковується в тому, що найкращі результати демонструють моделі, які передбачають лінійні зв’язки між вхідними та вихідною змінною – Lasso та ARIMA.
- Комплексні моделі, такі як BiLSTM, потребують більшої кількості даних, щоб натренуватись та якісно моделювати взаємозв’язки між даними

ВИСНОВКИ

Застосування різних підходів для прогнозування, таких як машинне навчання, нейронні мережі та регресійний аналіз, демонструють хороші результати у сфері прогнозування акцій. Проте для їх максимально ефективного використання потрібно проводити детальний аналіз наявних даних, коригувати їх, знаходити нові важливі фактори, підбирати найкращі гіперпараметри у моделях тощо.

Ефективна обробка даних з логічним заповненням ненаявних даних і вдале опрацювання викидів з використанням різноманітних методів (наприклад, Isolation Forest та аналіз Фур'є) здатні сформувати «повну картину» для моделей та нівелювати можливий суттєвий вплив поодиноких екстремальних значень.

Використання сентимент-аналізу як джерела формування фактору настроїв акціонерів є доволі багатообіцяючим методом. Проте, для якісного аналізу почуттів потрібно проводити багатогранний аналіз першоджерела настроїв – текстів, відео, голосових повідомлень тощо. В сучасному світі це зробити вкрай важко через велику кількість скорочень, «сленгу» та використання слів у іншому контексті, так як це не дозволяє побудувати ефективні моделі оцінки настроїв «на всі випадки життя». Також засилля нереальних користувачів (ботів) в соціальних мережах ускладнює процес отримання адекватної та репрезентативної оцінки настроїв суспільства стосовно тих чи інших подій.

Вибір оптимальних гіперпараметрів в моделях для прогнозування також є дуже важливою складовою, так як це допомагає моделям краще виділяти основні закономірності в даних та не перенавчатись.

В цілому, реалізувавши вище наведені зауваження та використавши достатній спектр моделей для прогнозування, в даній роботі були отримані якісні рішення для прогнозування цін на акції. Для набору даних компанії Tesla найкраще себе показали ансамбль моделей та модель LightGBM. Це свідчить про те, що комбінування моделей та агрегація їх результатів дозволила знайти різні за своєю природою взаємозв'язки в даних цін на акції TSLA. Для Apple найсильнішими моделями виявились Lasso та ARIMA. Причиною цьому є наявність певної лінійності у взаємозв'язках в даних цін

на акції AAPL. Найгірші результати для обох компаній мала модель BiLSTM через малу кількість тренувальних даних – така комплексна модель потребує більшої кількості інформації, щоб генерувати якісні прогнози. Виділення фактору настроїв користувачів внаслідок сентимент-аналізу зменшило похибки для моделей BiLSTM, Lasso та ансамблю моделей. Загалом, аналіз настроїв залишається потужним методом, здатним покращити точність прогнозів цін на акції.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Lim, K.P., & Brooks, R. The evolution of stock market efficiency over time: A survey of the empirical literature. *Journal of Economic Surveys*. 2011. P. 69-108. URL: https://doi.org/10.1111/j.1467-6419.2009.00611.x

2. Kendall, M.G. The Analysis of Economic Time-Series-Part I: Prices. Journal of Royal Statistical Association. 1953. P. 11-34. URL: https://doi.org/10.2307/2980947

3. Ryll, L., & Seidens, S. Evaluating the performance of machine learning algorithms in financial market forecasting. 2019. 21 P. URL: http://dx.doi.org/10.48550/arXiv.1906.07786

4. Di Persio, L., & Honchar, O. Artificial Neural Networks Architectures for Stock Price Prediction: Comparisons and Applications. *International Journal of Circuits, Systems and Signal Processing*. 2016. P. 403-413. URL: https://core.ac.uk/download/pdf/217561207.pdf

5. Patel, J., Shah, S., Thakkar, P., & Kotecha, K. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*. 2015. P. 259-268. URL: https://doi.org/10.1016/j.eswa.2014.07.040

6. Nazário, R.T.F., e Silva, J.L., Sobreiro, V.A., & Kimura, H. A literature review of technical analysis on stock markets. *The Quarterly Review of Economics and Finance*. 2017. P. 115-126. URL: https://doi.org/10.1016/j.qref.2017.01.014

7. Atsalakis, G.S., & Valavanis, K.P. Surveying stock market forecasting techniques–Part II: Soft computing methods. *Expert Systems with Applications*. 2009. P. 5932-5941. URL: https://doi.org/10.1016/j.eswa.2008.07.006

8. Nti, I.K., Adekoya, A.F., & Weyori, B.A. A systematic review of fundamental and technical analysis of stock market predictions. *Artificial Intelligence Review*. 2019. P. 3007-3057. URL: https://link.springer.com/article/10.1007/s10462-019-09754-z

9. Bollen, J., Mao, H., & Zeng, X. Twitter mood predicts the stock market. *Journal of Computational Science*. 2011. P. 1-8. URL: https://doi.org/10.1016/j.jocs.2010.12.007

10. Pagolu, V.S., Reddy, K.N., Panda, G., & Majhi, Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. International Conference on Signal Processing, Communication, Power and Embedded System, Paralakhemundi, India. 2016. 6 P. URL: https://doi.org/10.1109/SCOPES.2016.7955659

11. Nguyen, T.H., Shirai, K., & Velcin, J. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*. 2015. P. 9603-9611. URL: https://doi.org/10.1016/j.eswa.2015.07.052

12. Ding, X., Zhang, Y., Liu, T., & Duan, J. Deep learning for event-driven stock prediction. *Proceedings of Twenty-fourth international joint conference on artificial intelligence*. 2015. P. 2327-2333. URL: https://www.ijcai.org/Proceedings/15/Papers/329.pdf

13. Mittal, A. & Goel, A. Stock Prediction Using Twitter Sentiment Analysis. Stanford. *International Journal of Scientific Research in Science and Technology*. 2021. P. 265-270. URL: http://dx.doi.org/10.32628/CSEIT217475

14. Юхименко Г., Лазаренко, І. Прогнозування цін акцій на фондовому ринку за допомогою генеративних змагальних мереж та сентимент-аналізу соціальних мереж. *MODELING THE DEVELOPMENT OF THE ECONOMIC SYSTEMS*. 2022. С. 26-32. URL: http://dx.doi.org/10.31891/mdes/2022-6-4

15. Izzeldin, M., Muradoğlu, Y.G., Pappas, V., Petropoulou, A., & Sivaprasad, S. The impact of the Russian-Ukrainian war on global financial markets. *International Review of Financial Analysis*. 2023. 13 P. URL: https://doi.org/10.1016/j.irfa.2023.102598

16. Liashenko, O., Kravets, T. & Kostovetskyi, Y. Machine Learning and Data Balancing Methods for Bankruptcy Prediction. *Ekonomika*. 2023. P. 28-46. URL: http://dx.doi.org/10.15388/Ekon.2023.102.2.2

17. Hutto C. J., Gilbert E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*. 2014. P. 216-225. URL: https://doi.org/10.1609/icwsm.v8i1.14550

18. Seabold S., Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. *Proc. of the 9th Python in science conf. (SCIPY 2010)*. 2010. P. 92-96. URL: https://doi.org/10.25080/MAJORA-92BF1922-011

19. Abadi M., Agarwal A., Barham P., Brevdo E., Chen Z., Citro C., Corrado G. S., Davis A., Dean J. and others TensorFlow: A System for Large-Scale Machine Learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*. 2016. P. 264-283. URL: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

20. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Müller A. and others Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2012. P. 2825-2830. URL: https://arxiv.org/abs/1201.0490

21. Liashenko O., Kravets T., Pliushchov V. Stock Price Forecasting using Sentiment Analysis of Stock Tweets. Materials of 14th International Conference on Advanced Computer Information Technologies (ACIT). České Budějovice, Czech Republic, 2024. URL: https://doi.org/10.1109/ACIT62333.2024.10712521

22. Ляшенко О., Кравець Т., Плющов В. Вплив настроїв біржових твітів на формування ціни акцій. Збірник матеріалів Першої Всеукраїнській наукової конференції «Когнітивні дослідження: результати, виклики та перспективи». Київ, Україна, 2024. С. 289-297. URL: https://drive.google.com/file/d/1bjioYxbdjkejCoLaH2G9DwflXp5PmlDw/view?usp=sharing

23. Malkiel B. A Random Walk Down Wall Street by Burton. New-York, 1973. 464 p.

24. Fama E. Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*. 1970. P. 383-417. URL: https://doi.org/10.2307/2325486

25. Siegel J. Stocks for the Long Run: The Definitive Guide to Financial Market Returns & Long-Term Investment Strategies. New-York, 1994. 512 P.

26. Wolfson N., Russo T. The stock exchange specialist: an economic and legal analysis. *Duke Law Journal.* 1970. P. 707-746. URL: https://scholarship.law.duke.edu/dlj/vol19/iss4/3/

27. Kaustia M., Conlin A., Luotonen N. What drives stock market participation? The role of institutional, traditional, and behavioral factors. *Journal of Banking & Finance*. 2023. 21 P. URL: https://doi.org/10.1016/j.jbankfin.2022.106743

28. Oseni J. Determinants of Equity Prices in the Stock Markets. *Paradigms: A Research Journal of Commerce, Economics and Social Sciences*. 2009. P. 98-114. URL: http://dx.doi.org/10.2139/ssrn.1326912

29. Kravets T., Sytienko A. Wavelet analysis of the crisis effects in stock index returns. *Ekonomika*. 2013. P. 78-96. URL: https://doi.org/10.15388/Ekon.2013.0.1133

30. Про цінні папери та фондовий ринок: Закон України від 29.05.2006 №3480-IV. URL: https://zakon.rada.gov.ua/laws/show/3480-15#Text

31. Ляшенко О., Кравець Т., Хрущ Л. Застосування прикладних програм в економетричному моделюванні фінансових часових рядів. *Економіко-математичне моделювання соціально-економічних систем*. 2017. С. 33-60. URL: http://jnas.nbuv.gov.ua/article/UJRN-0000837791

32. Rosenberg D. Stock market indexes: Tracking Wall Street in real time and long term. *Encyclopedia Britannica.* 2025. URL: https://www.britannica.com/money/stock-market-index

33. Khan M., Ali H. Impact of Macroeconomic Indicators on Stock Market Predictions: A Cross Country Analysis. 2024. URL: http://dx.doi.org/10.56979/801/2024

34. Christensen K., Timmermann A., Veliyev B. Warp Speed Price Moves:Jumps after Earnings Announcements. *SSRN*. 2023. URL: https://dx.doi.org/10.2139/ssrn.4422376

35. Bhuiyan E., Chowdhury M. Macroeconomic variables and stock market indices: Asymmetric dynamics in the US and Canada. *The Quarterly Review of Economics and Finance.* 2020. P. 62-74. URL: https://doi.org/10.1016/j.qref.2019.10.005

36. Lee W., Jiang Ch., Indro D. Stock market volatility, excess returns, and the role of investor sentiment. *Journal of Banking & Finance*. 2002. P. 2277-2299. URL: https://doi.org/10.1016/S0378-4266(01)00202-3

37. Xie W., Zhao W., Ding B. Empirical Research on the Impact of Technological Innovation on New Energy Vehicle Sales. *Sustainability*. 2024. URL: http://dx.doi.org/10.3390/su16208794

38. Коваленко О., Коберник К. Ілон Маск шокує світ дивними менеджерськими рішеннями і наказами. *Матеріали журналу Бабель*. 2025. URL: https://babel.ua/texts/89644-ilon-mask-kupiv-twitter-za-fantastichni-44-milyardi-i-zruynuvav-za-rekordni-3-misyaci-istoriya-velikoji-nenavisti-perekaz-the-verge-ta-new-york-magazine

39. Kerin J. How Product Releases Affect Apple's Stock Price. *Investopedia*. 2024. URL: https://www.investopedia.com/articles/stocks/12/history-apple-stock-increases.asp

40. Levine A. Apple Has No Place to Hide. Tariffs on China Are Just 1 Problem. *Barron's*. 2025. URL: https://www.barrons.com/articles/apple-stock-tariffs-china-64b5319e

41. Wu A. The Story Behind Tesla's Success (TSLA). *Investopedia*. 2025. URL: https://www.investopedia.com/articles/personal-finance/061915/story-behind-teslas-success.asp

42. Reed E., Diongson D. History of Tesla & its stock: Timeline, facts & milestones. TheStreet. 2024. URL: https://www.thestreet.com/technology/history-of-tesla-15088992

43. Nicas J. Apple Becomes First Company to Hit $3 Trillion Market Value. *The New York Times*. 2022. URL: https://www.nytimes.com/2022/01/03/technology/apple-3-trillion-market-value.html

44. Han Ch., Fu X. Challenge and Opportunity: Deep Learning-Based Stock Price Prediction by Using Bi-Directional LSTM Model. Frontiers in Business, Economics and Management. 2023. P. 51-54. URL: https://doi.org/10.54097/fbem.v8i2.6616

45. Gao F., Zhang J., Zhang Ch., Shuang X. Long Short-Term Memory Networks with Multiple Variables for Stock Market Prediction. Neural Processing Letters. 2022. P. 4211-4229. URL: http://dx.doi.org/10.1007/s11063-022-11037-8

46. Yathish V. Loss Functions and Their Use In Neural Networks. *Towards Data Science*. 2022. URL: https://towardsdatascience.com/loss-functions-and-their-use-in-neural-networks-a470e703f1e9/

47. Altman N. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*. 1992. P. 175-185. URL: https://sites.stat.washington.edu/courses/stat527/s13/readings/Altman_AmStat_1992.pdf

48. Ляшенко О., Молоканова К. Аналіз та прогнозування дохідності акцей Microsoft та Pfizer за допомогою ARIMA-GARCH моделей. *Вісник Київського національного університету імені Тараса Шевченка*. 2023. С. 76-87. https://doi.org/10.17721/1728-2667.2023/222-1/10

49. Maulud D. H., Abdulazeez A. M. A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*. 2020. P. 140-147. URL: http://dx.doi.org/10.38094/jastt1457

50. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*. 1994. P. 267-288. URL: https://www.jstor.org/stable/2346178

51. Loupe G. Understanding Random Forests: From Theory to Practice. Preprint. 2014. P. 173. URL: http://dx.doi.org/10.13140/2.1.1570.5928

52. Natekin A., Knoll A. Gradient Boosting Machines, A Tutorial. *Frontiers in Neurorobotics*. 2013. 21 P. URL: http://dx.doi.org/10.3389/fnbot.2013.00021

53. Guolin K., Meng Q., Finley T., Wang T., Chen W., Ma W., Liu T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017. 9 P. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

54. Chakrabarti B., Jain A., Singh Nagpal P., Rout J. A spatiotemporal context aware hierarchical model for corporate bankruptcy prediction. *Multimedia Tools and Applications*. 2023. P. 28281-28303. URL: https://link.springer.com/article/10.1007/s11042-023-15353-6

55. McElfresh D., Khandagale S., Valverde J., Prasad V., Feuer B., Hedge Ch., Ramakrishnan G., Goldblum M., White C. When Do Neural Nets Outperform Boosted Trees on Tabular Data? *Materials of 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*. 2023. 39 P. URL: http://dx.doi.org/10.48550/arXiv.2305.02997

56. Вербівський, Д. С., Карплюк, С. О., Фонарюк, О. В., Сікора, Я. Б. Бустінг і беггінг як методи формування ансамблей моделей. *Abstracts of the 7th International scientific and practical conference*. P. 163-169. URL: http://eprints.zu.edu.ua/32426/

57. Mittal A., Goel A. Stock Prediction Using Twitter Sentiment Analysis. *International Journal of Scientific Research in Science and Technology*. 2021. 6 P. URL: http://dx.doi.org/10.32628/CSEIT217475

58. Akiba T., Sano S., Yanase T., Ohta T., Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. *The 25th ACM SIGKDD International Conference*. 2019. P. 2623-2631. URL: https://arxiv.org/abs/1907.10902

59. Yahoo! Finance. URL: https://finance.yahoo.com/

60. Yuhkymenko H. Stock Tweets for Sentiment Analysis and Prediction. 2022. URL: https://www.kaggle.com/datasets/equinxx/stock-tweets-for-sentiment-analysis-and-prediction

61. Dolan B. What Is MACD? *Investopedia*. 2024. URL: https://www.investopedia.com/terms/m/macd.asp

62. Thompson C. Bollinger Bands: What They Are, and What They Tell Investors. *Investopedia*. 2024. URL: https://www.investopedia.com/terms/b/bollingerbands.asp

ДОДАТКИ

Додаток А

Важливі блоки програмного коду

```
def evaluate_sentiment(x): #функція для розрахунки оцінки настроїв акціонерів
    scores = []
    for tweet in x:
        tweet = unicodedata.normalize('NFKD', tweet)
        score = sia_obj.polarity_scores(tweet)['compound']
        if score != 0:
            scores.append(score)
    if len(scores) > 0:
        return np.median(scores)
    return 0

sentiment_scores = tweets.groupby(['Date'])['Tweet'].apply(evaluate_sentiment)
sentiment_scores.name = 'scores'

sentiment_scores.to_csv(f'sentiment_data/{stock_name.lower()}_sentiment_scores.csv')

def fourierExtrapolation(x, n_predict, n_harm): #функція апроксимації Фур'є
    n = x.size
    t = np.arange(0, n)
    p = np.polyfit(t, x, 1)
    x_notrend = x - p[0] * t
    x_freqdom = np.fft.fft(x_notrend)
    f = np.fft.fftfreq(n)
    indexes = list(range(n))
    indexes.sort(key = lambda i: np.absolute(f[i]))
```

```
        t = np.arange(0, n + n_predict)
        restored_sig = np.zeros(t.size)
        for i in indexes[:1 + n_harm * 2]:
            ampli = np.absolute(x_freqdom[i]) / n   # amplitude
            phase = np.angle(x_freqdom[i])          # phase
            restored_sig += ampli * np.cos(2 * np.pi * f[i] * t + phase)
        return restored_sig + p[0] * t

    def rmspe(y_real, y_pred): #функція метрики RMSPE
        return round(np.sqrt(np.mean(np.square((y_real - y_pred) / y_real))) * 100, 4)

    def mape(y_real, y_pred): #функція метрики MAPE
        return round(np.mean(np.abs((y_real - y_pred) / y_real)) * 100, 4)

    def fill_metrics_arrays(metric): #функція заповнення матриці метрик
        array_test = pd.DataFrame(index=labels[1:], columns=names + ['Avg'])
        array_train = pd.DataFrame(index=labels[1:], columns=names + ['Avg'])
        for j in range(5):
            for i in range(4):
                array_test.iloc[j, i] = round(metric(y_real.iloc[-20:, i], values[j + 1].iloc[-20:,
i]), 4)

                train_preds = values[j + 1].iloc[:, i]
                array_train.iloc[j, i] = round(metric(y_real.iloc[-len(train_preds):-20, i],
train_preds[:-20]), 4)

            array_test.iloc[j, 4] = round(np.average(array_test.iloc[j, :4]), 4)
            array_train.iloc[j, 4] = round(np.average(array_train.iloc[j, :4]), 4)
        return array_test, array_train
```

class CustomTimeSeriesSplit(BaseCrossValidator): #надбудова до базового крос-валідатора

```
def __init__(self, n_splits=3, validation_window_size=None):
    self.n_splits = n_splits
    self.validation_window_size = validation_window_size

def split(self, X, y=None, groups=None):
    n_samples = len(X)
    split_size = self.validation_window_size or n_samples // (self.n_splits + 1)
    for i in range(self.n_splits):
        val_start = n_samples - (self.n_splits - i) * split_size
        val_end = val_start + split_size
        train_indices = np.arange(0, val_start)
        val_indices = np.arange(val_start, val_end)
        yield train_indices, val_indices

def get_n_splits(self, X=None, y=None, groups=None):
    return self.n_splits
```

def lgbm_objective(*trial*): #функція, за якою будується модель та оптимізуються її гіперпараметри на прикладі моделі LightGBM

time_series_cv = CustomTimeSeriesSplit(*n_splits*=n_splits, *validation_window_size*=validation_window_size)

```
params = {
    'num_leaves': trial.suggest_int('num_leaves', 2, 31),
    'max_depth': trial.suggest_int('max_depth', 2, 7),
    'min_child_samples': trial.suggest_int('min_child_samples', 1, 100),
    'reg_alpha': trial.suggest_float('reg_alpha', 1e-4, 10, log=True),
```

```
        'reg_lambda': trial.suggest_float('reg_lambda', 1e-4, 10, log=True),
        'subsample': trial.suggest_float('subsample', 0.25, 1.0),
        'subsample_freq': trial.suggest_int('subsample_freq', 1, 10),
        'colsample_bytree': trial.suggest_float('colsample_bytree', 0.25, 1.0),
        'learning_rate': trial.suggest_float('learning_rate', 1e-2, 0.5, log=True)
    }

            model = LGBMRegressor(**params, n_estimators=n_estimators, random_state=29, n_jobs=1)

    score = cross_val_score(model, X_train, y_train, cv=time_series_cv, scoring='neg_mean_squared_error', n_jobs=-1).mean()
    return score
```

def lgbm_forecasting(*X_train*, *y_train*, *X_test*, *path*, *preds*, *preds_index*): #функція для побудови моделі з використанням оптимальних гіперпараметрів та прогнозування на прикладі моделі LightGBM

```
    params = pd.read_csv(path, index_col=0)
    params = params.iloc[:, 0].to_dict()

    for i in params:
        if int(params[i]) == params[i]:
            params[i] = int(params[i])

    model = LGBMRegressor(**params, random_state=84, n_jobs=-1, verbose=-1)
    model.fit(X_train, y_train)

    predicted = model.predict(np.vstack([X_train, X_test]))
    preds.iloc[:, preds_index] = predicted
```