УДК 004.89 DOI: 10.18372/2073-4751.76.18245

Саттарова М.Л.,
Тхорик В.Б.,
Савченко А.С., д.т.н.,
orcid.org/0000-0001-8205-8852,
Моденов Ю.Б., к.т.н.,
orcid.org/0000-0003-3898-4159

ПРОГНОЗУВАННЯ ЦІН ФІНАНСОВИХ ІНСТРУМЕНТІВ ЗА ДОПОМОГОЮ ЗАСОБІВ ГЛИБОКОГО НАВЧАННЯ

Національний авіаційний університет

sattarova.mrgrt@gmail.com,
v.tkhoryk@icloud.com,
alina.savchenko@npp.nau.edu.ua,
yurii.modenov@npp.nau.edu.ua

Вступ

У сучасному світі, де фінансові ринки відіграють ключову роль у глобальній економіці, точність прогнозування цін на акції є життєво важливою не тільки для інвесторів та трейдерів, але й для економічної стабільності у її ширшому розумінні. З появою розширених методів машинного та глибокого навчання відкрилися нові можливості для покращення прогнозування фінансових часових рядів.

Проте, багато існуючих підходів до прогнозування цін на акції зазнають важливих обмежень. Зокрема, в переважній більшості функціонал існуючих рішень обмежується механізмами технічного аналізу, які є простішими в реалізації, ніж засоби фундаментального аналізу. В той же час, останній надає не менш важливу інформацію та враховує вплив багатьох факторів, які є критичними. Відповідно, не враховуючи останній, інвестор ігнорує чинники та показники, які нерідко відіграють ключову роль в прийнятті інвестиційних рішень.

Також не завжди фактори, що мають вплив на формування ціни акції, можуть бути отримані з виключно технічної аналітики або з фінансової звітності компанії. Натомість, значна кількість важливої інформації міститься в новинах, публікаціях, анонсах, звітах державних органів тощо. Відсутність інтеграції даних з текстових джерел для глибшого аналізу ринкового сентименту, що є поширеною проблемою для наявних рішень, призводить до упущення опрацювання вкрай важливих аспектів, від яких залежатиме вартість біржових активів. Крім того, більшість моделей не враховують зашумленість характеристик фінансових даних, що може суттєво впливати на якість прогнозу.

Аналіз існуючих досліджень

Прогнозування цін акцій було предметом цілого ряду досліджень протягом останніх років.

Lu et al. [1] запропонували нову модель для прогнозування цін на акції за допомогою комбінації згорткової нейронної мережі (*Convolutional neural network, CNN*) та мережі з довгою короткочасною пам'яттю (*Long short-term memory, LSTM*). *CNN* використовується для ефективного вилучення ознак з даних, а *LSTM* – для прогнозування ціни на акції з отриманими ознаками даних. Модель використовує щоденні ціни на акції з 1 липня 1991 року по 31 серпня 2020 року, що охоплює 7127 торгових днів. Для цього аналізу було обрано вісім характеристик: ціна відкриття (*Open*) та закриття (*Close*), найвища (*High*) і найнижча (*Low*) ціна, обсяг (*Volume*), оборот (*Turnover*), підйоми і падіння (*Ups and Downs*) та зміни (*Change*).

Структура моделі *CNN-LSTM* була побудована з тривимірного вектора даних, згорткового шару (*convolution layer*), шару об'єднання (*pooling layer*) та шару *LSTM*

для навчання даних та отримання вихідного (цільового) значення.

Дослідження порівнювало ефективність *CNN-LSTM* з іншими моделями, такими як багатошаровий перцептрон (*Multilayer perceptron, MLP*), *CNN*, рекурентна нейронна мережа (*Recurrent neural network, RNN*), *LSTM* та *CNN-RNN*. Модель *CNN-LSTM* продемонструвала найвищу точність прогнозування серед всіх, із найменшою середньою абсолютною похибкою (*MAE*) та середньоквадратичною похибкою (*RMSE*). Зокрема, *MAE* і *RMSE* для *CNN-LSTM* були 27,564 і 39,688 відповідно. Це свідчить про значне покращення порівняно з іншими моделями, демонструючи перевагу моделі *CNN-LSTM* як щодо ступеня відповідності, так і значення помилки.

Jarrah et al. [2] мали на меті спрогнозувати індекси фондового ринку Саудівської Аравії за допомогою підходу глибокого навчання з використанням багатовимірних даних часових рядів, які включають різні змінні, такі як початкова, найнижча, найвища ціна та ціна акцій на момент закриття ринку.

Застосований метод включає кілька етапів, починаючи з використання експоненціального згладжування (*ES*) для усунення шуму з даних. Після цього застосовувався метод рухомого вікна з п'ятьма кроками, щоб перетворити задачу прогнозування часових рядів у контрольовану навчальну задачу. Фінальним кроком було використання мультиваріативного *LSTM* алгоритму глибокого навчання для прогнозування цін на фондовому ринку.

Запропонована мультиваріативна модель глибокого навчання *LSTM* досягла значення точності прогнозування 97,49% і 92,19% для однофакторної моделі. Такий результат підкреслює доцільність використання багатьох джерел інформації для прогнозування цін на фондовому ринку.

Akita et al. [3] досліджували можливість створення моделі глибокого навчання для покращення прогнозування цін акцій на фондовому ринку шляхом використання як числових, так і текстових даних.

Запропонована модель застосовує вектор абзацу (*Paragraph Vector*), щоб отримати розподілене представлення кожної з доступних новин про компанію. Даний процес було проведено з поєднанням водночас обидвох категорій вектору абзацу, а саме *Distributed Memory Model of Paragraph Vectors* (*PV-DM*) і *Distributed Bag of Words of Paragraph Vector* (*PV-DBOW*).

Отримані розподілені представлення та щоденні ціни відкриття 50 компаній Токійської фондової біржі використовуються для передбачення ціни закриття за допомогою регресійного аналізу. Для врегулювання впливу часових рядів була використана модель довготривалої короткочасної пам'яті (*LSTM*).

Для оцінювання точності отриманих результатів використовувалась симуляція ринку, за якою імітувалась або купівля акції при передбаченому рості ціни, або ж її продаж у випадку передбаченого зниження ціни. Мірою точності прогнозування виступав сукупний прибуток, отриманий у результаті проведення симуляції на вказаному тестувальному проміжку. За результатами оцінювання досліджувана модель показала кращий результат у 4 з 5 секторів розглянутих компаній та отримала кращий сумарний результат (прибуток) у розмірі 12,1 млн єн. Даний показник порівнювався з результатами, отриманими методом опорних векторів, *MLP* і *RNN*, що дорівнювали -0,47, -5,6 та 2,6 мільйонів єн відповідно, що показує підвищення рівня точності прогнозування при залученні текстових даних для навчання моделі.

Запропоновані у розглянутих дослідженнях рішення демонструють високі показники точності прогнозування, однак вони містять ряд важливих недоліків та недопрацювань.

По-перше, вони являють собою досить складні моделі, однак суттєвим недоліком більшості з них є використання лише даних про історичну зміну ціни акцій, тобто задіюється лише технічний

аналіз активів. При цьому упускається той факт, що ціна тієї чи іншої акції у довільний момент часу є наслідком переліку факторів, які не обмежуються лише історичними даними, а включають також як показники фундаментального аналізу, дані із фінансової звітності компанії, розподіл її активів, так і макроекономічну ситуацію. Таким чином ігнорується значна частина факторів, які мають прямий та безпосередній вплив на величину, що підлягає прогнозуванню.

По-друге, у дослідженнях інколи використовують технічні показники як додаткові вхідні дані для моделей, однак підхід до вибору цих показників часто не є обґрунтованим. Сучасні підходи до роботи з характеристиками даних (*feature engineering*) включають в себе дослідження залежностей між показниками та цільовою величиною. Крім того, важливим є відбір найбільш відповідних показників на основі визначених критеріїв, що дозволяє підвищити ефективність та точність моделі. Відсутність такого системного підходу може призвести до погіршення якості прогнозування.

По-третє, фінансові дані за своєю природою є дуже «зашумленими». Це означає, що вони можуть містити багато випадкових відхилень, які не мають відношення до основних тенденцій ринку. Недостатнє або ж відсутнє використання механізмів для попереднього очищення даних може призвести до введення моделі в оману та зменшення її загальної ефективності. Сучасні підходи до роботи з характеристиками даних також передбачають методи видалення шуму.

Також потрібно враховувати той факт, що сучасний ринок акцій часто реагує не тільки на числові показники, але і на новини, звіти, соціальні мережі та інші текстові джерела. Ігнорування цієї інформації може призвести до неповного розуміння ринкових процесів, і, як наслідок – зниженої якості моделювання цих процесів і точності отриманих результатів прогнозування. Використання обробки природної мови (*Natural Language Processing, NLP*) та аналізу настрою (*Sentiment Analysis*) може допомогти збагатити моделі цією додатковою інформацією, яка збільшить їхню точність.

Постановка задачі дослідження

Для ефективного управління інвестиційним портфелем перспективним є підхід, що поєднує переваги *LSTM* мереж з ретельно підібраними вхідними даними та техніками обробки. Необхідно розробити модель, яка не тільки включає технічні показники, але й охоплює дані фундаментального аналізу та макроекономічні фактори. Доцільно також використати методи обробки природної мови (*NLP*) та аналіз сентименту для інтеграції інформації з текстових джерел, таких як новини та фінансові звіти, щоб забезпечити всебічний аналіз впливу різних факторів на ціни акцій.

Для перевірки ефективності моделі машинного навчання необхідно провести порівняльний аналіз з іншими методами прогнозування.

Мета

Метою дослідження є розробка моделі, яка не тільки включає технічні показники, але й охоплює дані фундаментального аналізу та макроекономічні фактори для підвищення ефективності процесу управління інвестиційним портфелем та прийняття інвестиційних рішень за рахунок поєднання методів технічного та фундаментального аналізу засобами штучного інтелекту (ШІ) для прогнозування цін активів на фондовому ринку та відповідне наближення фінансових показників портфеля до таких, які інвестор вважає цільовими.

Принципи запропонованого підходу

У світлі аналізу існуючих підходів до прогнозування цін на акції, пропонується новий підхід, який ставить за ціль виправити ключові недоліки попередніх методик. Модель базується на наступних принципах:

- системний підхід до використання показників за рахунок впровадження сучасних методик *feature*

engineering та *feature selection*. Це передбачає детальне дослідження залежностей між різними числовими показниками та цільовою величиною, а також застосування автоматизованих методів відбору ознак;

- знешумлення фінансових даних. З урахуванням природної «зашумленості» фінансових даних, використовуються методики видалення шуму задля зосередження на основних тенденціях ринку;
- врахування текстової інформації. Застосування технік *NLP* та аналізу настрою дозволяє інтегрувати новини, звіти та інші текстові джерела інформації, які можуть впливати на ринкові тенденції та відображати актуальну повістку, яка впливає на цільову величину;
- інтеграція фундаментального та макроекономічного аналізу з технічним. Розроблена модель, окрім технічних показників, також включає в себе ключові показники фундаментального аналізу, такі як чистий прибуток компанії, а також макроекономічні показники, які мають прямий вплив на фондовий ринок.

Ці засади у поєднанні з глибокими нейронними мережами, здатними враховувати часові послідовності даних, створюють базу для розробки ефективної та більш точної моделі прогнозування, яка відображає сучасні тенденції зміни акцій компанії на фондовому ринку.

Відбір факторів для включення в прогнозну модель

Реалізація задачі прогнозування цін активів на практиці є досить складною, адже вимагає врахування великої кількості факторів, які впливають на динаміку цін конкретних активів та ринку загалом, а також визначення відповідного способу включення представлень цих факторів до фінансових моделей.

Ці фактори та їх взаємозв'язки ретельно вивчаються інвесторами, вони використовують різні підходи для аналізу їх впливу на ринок для прийняття обгрунтованих інвестиційних рішень. Вони характеризуються різним характером впливу та його природою – у той час як деякі фактори можуть викликати негайні реакції в цінах акцій, вплив інших може бути видимим здебільшого у довгостроковій перспективі, взаємозв’язки між ними можуть мати як лінійний, так і нелінійний характер. Основні фактори, які впливають на ціну акцій, можна розділити на три категорії: фундаментальні, технічні і макроекономічні.

Фундаментальні фактори відображать фінансовий стан компаній-емітентів активів, вони включать доходи, прибуток, рівень заборгованості компанії, тощо. Аналіз цих факторів полягає у дослідженні фінансової звітності, наприклад звітів про прибутки і збитки (*Income statements*) і баланс компанії (*Balance sheet*) та проведеній на основі цього оцінці активів. Даний підхід лежить в основі методу фундаментального аналізу. Вони включають зокрема:

1. *P*/*E* (*Price-to-Earnings*) *Ratio*, або кошторис ціни до прибутку, це один із ключових показників, які використовують інвестори для оцінки фінансової ефективності та оцінки акцій компанії. *P*/*E Ratio* обчислюється шляхом ділення поточної ринкової ціни акції на прибуток на акцію (*EPS*). Цей показник може використовуватися для оцінки того, наскільки ринкова ціна акції вище або нижче прибутку, який компанія генерує. Високий *P*/*E Ratio* може вказувати на те, що інвестори очікують високого зростання прибутку у майбутньому або просто переплачують за акцію на даний момент. З іншого боку, низький *P*/*E Ratio* може свідчити про недооцінку акції або очікування низького зростання прибутку [4].
2. *P*/*B* (*Price-to-Book*) *Ratio*, або кошторис ціни до вартості активів обчислюється шляхом ділення поточної ринкової ціни акції на вартість активів компанії, враховуючи їх балансову вартість. Якщо *P*/*B Ratio* менше 1, це може свідчити про те, що акції продаються нижче їх балансової вартості, що може вказувати на потенційно недооцінені акції. Якщо *P*/*B Ratio* більше 1, це може вказувати на переоцінку акцій.
3. *PEG* (*Price*/*Earnings-to-Growth*) *Ratio* – це фінансовий показник, який

поєднує в собі два ключові фактори: коефіцієнт ціни до прибутку (*P/E Ratio*) і очікуваний ріст прибутку компанії [4]. Він допомагає інвесторам оцінити вартість акції відносно очікуваного росту прибутку компанії: якщо він дорівнює 1, то це може свідчити про те, що ринкова ціна акції достатньо відображає очікуваний ріст прибутку, якщо ж воно менше або більше 1, то акція вважається недооціненою або переоціненою відповідно. *PEG Ratio* корисний, оскільки він враховує не тільки поточну прибутковість компанії (*P/E Ratio*), але і її потенціал для зростання в майбутньому.

4. Дивідендна дохідність (*Dividend Yield*) – це фінансовий показник, який вказує, яку частину поточної ціни акції складають виплачені дивіденди. Висока дивідендна дохідність може бути привабливою для інвесторів, які шукають стабільний дохід від своїх інвестицій. Однак висока дивідендна дохідність також може вказувати на проблеми всередині компанії, такі як низька прибутковість або нездатність інвестувати у зріст.

5. *ROE* (*Return on Equity*), або прибуток на власний капітал відображає ефективність використання власних коштів акціонерів компанією для генерації прибутку. *ROE* виражається у відсотках і вказує на те, яку частину власного капіталу компанія здатна заробити як чистий прибуток. Вищий *ROE* зазвичай свідчить про більшу ефективність використання капіталу та вищу прибутковість для акціонерів.

6. *Current Ratio* (поточний коефіцієнт) – це фінансовий показник, який вимірює здатність компанії погасити свої короткострокові зобов'язання (зобов'язання, які мають бути сплачені протягом одного року) за допомогою своїх поточних активів. Цей показник важливий для оцінки фінансової стійкості компанії та її здатності впоратися з короткостроковими фінансовими зобов'язаннями, такими як платежі по кредитам чи зобов'язання перед постачальниками.

7. *Quick Ratio*, також відомий як *Acid-Test Ratio* або *Liquidity Ratio,* – це фінансовий показник, який використовується для вимірювання готівкової ліквідності компанії, тобто її здатності виконати короткострокові зобов'язання, враховуючи лише найбільш ліквідні активи (готівку, еквіваленти готівки та цінні папери, які можна легко перетворити на готівку). Зазвичай більший *Quick Ratio* вважається більш безпечним, оскільки це означає, що компанія має більше ліквідних активів для покриття своїх зобов'язань.

Технічні фактори (індикатори) є предметом технічного аналізу – іншого методу аналізу активів. Він базується на використанні історичної інформації про ціни акцій і обсяги торгів, на обчисленні фінансових змінних та показників, і подальшому їх використанні для передбачення майбутньої динаміки цін активів та розробки вдповідних інвестиційних стратегій. Прикладами таких індикаторів є:

1. Рухомі середні (*Moving Averages*) – це один із ключових інструментів у технічному аналізі фінансових ринків. Вони використовуються для визначення загальних тенденцій ціни акцій, що допомагає інвесторам і трейдерам приймати рішення про їх купівлю або продаж.

Види рухомих середніх включають просте (*SMA*), зважене (*WMA*) та експоненціальне (*EMA*). Вибір періоду і типу рухомого середнього залежить від інвестиційних уподобань та характеру аналізованих даних. Короткі періоди відображають поточну динаміку ціни акцій, тоді як довгі періоди роблять його більш згладженим і придатним для визначення та аналізу довгострокових трендів.

2. Індикатор сходження / розходження рухомих середніх (*MACD*) – це технічний показник осциляторного типу, який показує силу тренду та дозволяє відстежувати його зміну. *MACD* використовує рухомі середні для визначення імпульсу (моментуму) цін акції або іншого торгового активу. Являючи собою індикатор моментуму, *MACD* є корисним показником для інвесторів, оскільки він дозволяє визначити швидкість руху ціни та імовірність збереження чи зміни її тренду.

3. Індекс відносної сили (*RSI*) – це технічний показник, який використовується в аналізі фінансових ринків, особливо на ринку акцій, для визначення ступеня перекупленості чи перепроданості цінних паперів. Дивергенція *RSI* вказує на можливість зміни тренду [5]. Наприклад, якщо *RSI* формує новий високий пік, а ціни не роблять цього, це може свідчити про втрату міцності поточного тренду.

Макроекономічні фактори відображають економічну ситуацію в цілому та характеризують економічне середовище в якому оперують компанії-емітенти. До них відносяться:

1. ВВП (Валовий внутрішній продукт – *Gross Domestic Product* або *GDP*) – це ключовий макроекономічний показник, який вимірює загальну вартість всіх товарів і послуг, що виробляються в економіці країни протягом певного періоду часу. Зростаючий ВВП вказує на здорову економіку, що може призвести до збільшення прибутків для компаній та потенційного підвищення цін акцій. І навпаки, зниження ВВП вказує на падіння економіки та, як результат – потенційного зниження прибутків компаній-емітентів та цін акцій.

2. *CPI* (*Consumer Price Index*), або Індекс споживчих цін, є економічним показником, який вимірює зміни середнього рівня цін на споживчі товари та послуги в країні протягом певного періоду часу. Зростання *CPI* вказує на інфляцію, а зменшення – на дефляцію. Наслідком інфляції є зменшення споживчих витрат і зниження прибутків компаній, що в свою чергу має потенційно негативний вплив на ціни акцій. З іншого ж боку, дефляція може сигналізувати про уповільнення економіки, що також може мати негативний вплив на ціни акцій.

3. Рівень безробіття (*Unemployment Rate*) – це ключовий економічний показник, який вимірює відсоток робочої сили, яка на даний момент є безробітною та активно шукає роботу в економіці. В контексті впливу на фондовий ринок вищий рівень безробіття призводить до зниження споживчих витрат, потенційно знижуючи прибутковість компаній та відповідно ціни акцій.

4. Процентні ставки (*Interest Rates*) – економічний показник, який визначає вартість (відсоток) користування позиковими грошима, які комерційні банки короткостроково займають один в одного у процесі своєї діяльності. Цей показник є корисним для розгляду у інвестиційному аналізі, оскільки при підвищенні процентних ставок вартість позичання грошей для бізнесу збільшується, що може знизити прибутковість компаній та як наслідок викликати зниження цін акції компаній на фондових ринках. І навпаки – нижчі процентні ставки знижують вартості запозичень для компаній, потенційно збільшуючи прибутковість і ціни на акції, водночас роблячи акції більш привабливими для інвестування порівняно з облігаціями.

Також ринкові умови та тенденції значним чином формуються під впливом інформаційних потоків які включають текстові дані, такі як новини про компанії, економічні показники, політичні події та глобальні події, новини про релізи нових продуктів компанії тощо. Тому важливим для врахування у розроблюваній нами моделі прогнозування є результати аналізу текстових даних з соціальних медіа, новин та інших альтернативних джерел.

На основі наведеного вище переліку факторів, які мають вплив на ціни акцій було сформовано набір даних для використання розроблюваній прогнозній моделі. Він включає відповідно історичні дані компанії *Apple* (*AAPL*) періодом 10 років (з кінця вересня 2013 до початку жовтня 2023 року) про саму ціну акцій (*Open, High, Low, Closed*), обсяги торгів (*Volume*), ряд технічних індикаторів, фундаментальних та макроекономічних показників, для яких аналітично було виявлено найбільш сильні залежності з ціною акцій. Більшість цих даних було зібрано з відкритих джерел за допомогою модуля автоматизованого веб-скрапінгу, решту – обраховано програмно згідно відповідних формул.

Крім цього датасет включає результати обробки текстової інформації (новин

про компанію) – показник оцінки тональності тексту для кожної новини, обрахований за допомогою моделі *BERT*.

BERT (*Bidirectional Encoder Representations from Transformers*) – це революційна модель у сфері *NLP*, здатна виконувати завдання класифікації тексту, включаючи аналіз тональності тексту. Вона відома тим, що вловлює складні текстові шаблони, що є корисним для аналізу новинних статей. У нашому підході було використано заздалегідь натреновану на фінансових текстових даних *BERT* модель – *FinBERT*.

Попередня обробка даних

Фінансові дані для відібраних факторів мають різну дискретність. Дані про ціни акцій та технічні індикатори наявні на кожен робочий день. Фундаментальні дані, взяті із квартальної звітності компаній, наявні лише за квартальні проміжки часу. Макроекономічні показники мають періодичність у квартал, а деякі з них – у місяць. Тому необхідним етапом попередньої обробки є заповнення тих данних, яких не вистачає, адже для включення до моделі дані повинні мати однакову дискретність у часі. Для вирішення цієї задачі запропоновано використати метод інтерполяції – поліноміальний другого порядку, оскільки він найефективніше враховує природу наявних даних у датасеті.

Також враховуючи природну «зашумленість» фінансових даних, та потенційний ризик виникнення перенавчання моделі до шумів у вхідних даних, наступним кроком попередньої обробки даних було виконано знешумлення даних за допомогою методу експоненційного згладжування (*Exponential Smoothing*). Експоненційне згладжування є широко використовуваним методом зменшення шуму в даних на етапі попередньої обробки. Доцільним буде використання цього методу для задачі прогнозування в запропонованій моделі, оскільки надає більші вагові коефіцієнти останнім спостереженням у наборі даних, враховуючи при цьому всі історичні дані, так як у випадку нашої задачі останні тенденції більше вказують на майбутні значення.

Застосування техніки експоненційного згладжування дозволить отримати більш плавний тренд, який відповідає вхідних даним, але при цьому зменшити наявні в них короткочасні коливання, як показано на рис. 1 на прикладі колонки *Close price* з вхідного датасету.

Сформований після попередньої обробки даних датасет представлений на рис. 2.

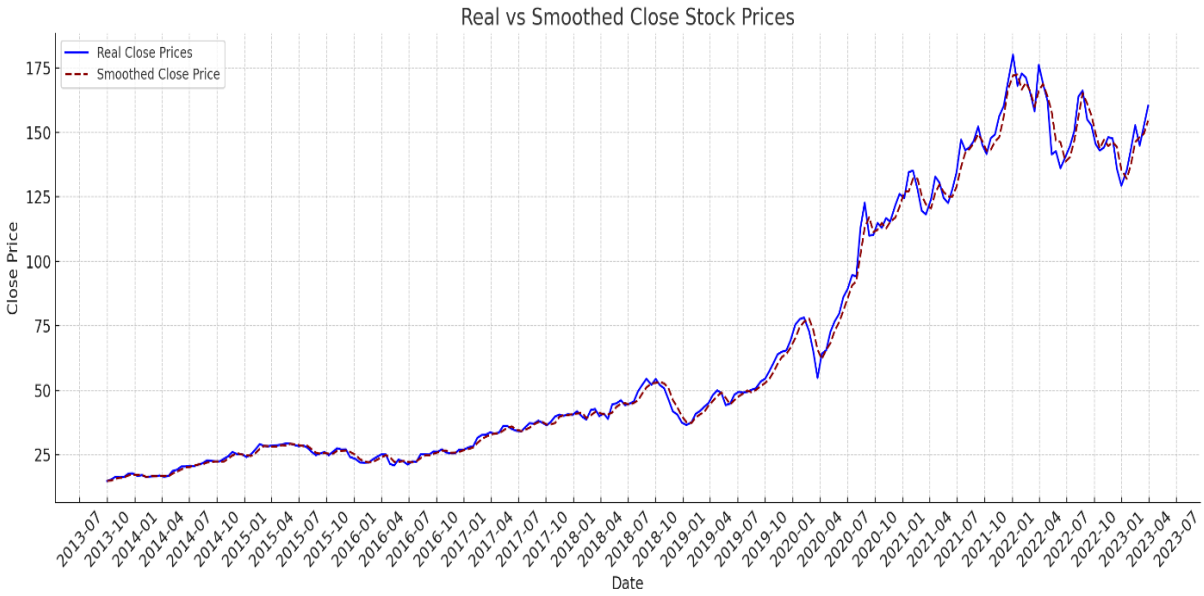

Рис. 1. Зменшення шуму у вхідних даних

	Date	2013-09-30	2013-10-01	2013-10-02	2013-10-03	2013-10-04
Технічні індикатори	Open	14.876	14.913	15.137	15.289	15.082
	High	15.013	15.246	15.329	15.346	15.105
	Low	14.787	14.911	15.078	14.985	14.918
	Close	14.86	15.21	15.259	15.068	15.056
	Volume	260.156M	353.884M	289.184M	322.753M	258.868M
	WMA50	15.009	15.024	15.041	15.048	15.053
	WMA200	14.052	14.063	14.074	14.084	14.093
	SMA50	14.805	14.845	14.891	14.92	14.949
	SMA200	14.12	14.115	14.114	14.11	14.103
	EMA50	14.709	14.728	14.749	14.762	14.773
	EMA200	14.625	14.631	14.637	14.642	14.646
	RSI7	46.91	57.728	59.116	51.532	51.06
	RSI14	48.999	54.397	55.127	51.701	51.488
	RSI21	50.713	54.402	54.908	52.555	52.409
Макроекономочні показники	Unemployment	7.205	7.2	7.195	7.189	7.183
	GDP	4587.009	4586.321	4585.575	4584.771	4583.91
	CPI	233.567	233.546	233.525	233.505	233.484
	Interest Rate	0.09	0.09	0.09	0.09	0.09
Фундаментальні показники	Quick Ratio	1.64	1.637	1.635	1.632	1.63
	Current Ratio	1.68	1.677	1.675	1.672	1.669
	P/E Ratio	10.5	10.525	10.55	10.575	10.599
	P/B Ratio	3.03	3.036	3.041	3.047	3.052
	PEG Ratio	-1.993	-2.046	-2.097	-2.146	-2.195
	EPS	0.3	0.302	0.304	0.305	0.307
	RoE	29.06	29.054	29.048	29.042	29.036
	Net Income	7.740B	7.853B	7.964B	8.074B	8.183B
	Dividend Yield	0	0	0	0	0
	Dividends	0	0	0	0	0
	Stock Splits	0	0	0	0	0
Текстові дані	Sentiment score	2.5	2	2	1.5	2.5

Рис. 2. Структура набору даних, сформованого аналітичних чином, для використання в прогнозній моделі

Feature engineering та Feature selection

Наступним кроком відбору факторів (ознак) для включення в модель прогнозування є застосування методу градієнтного бустингу для обрахунку значимості ознак та виділення найбільш вдалих ознак для включення до моделі зі сформованого аналітично датасету.

Градієнтний бустинг – це потужний метод, який базується на використанні ансамблів дерев рішень для обрахунку важливості ознак для прогнозної моделі. Метод градієнтного бустингу полягає у формуванні ансамблю дерев рішень в поетапній манері – де кожна нова модель (дерево рішень) навчається виправляти помилки, зроблені попередніми. Під час процесу навчання обчислюється важливість ознак з наданого датасету – ознаки, які більш часто використовуються у формуванні ключових поділів у цих деревах рішень вважаються більш значими та мають вищий показник відносної важливості. Цей показник розраховується явно для кожного атрибута в наборі даних, що дозволяє ранжувати атрибути (ознаки) та порівнювати їх базуючись на його значеннях. Важливість обчислюється для окремого дерева рішень на основі значення того, наскільки кожна точка розділення атрибутів покращує показник продуктивності, зваженого за кількістю спостережень, за які відповідає вузол. Потім важливість ознак усереднюється по всіх деревах рішень у моделі.

В розроблюваному в рамках даного дослідження рішенні було використано бібліотеку *XGBoost* для реалізації методу градієнтного бустингу для відбору предикторів для нашої моделі. Отримані показники відносної важливості ознак для підготованого датасету приведено на рис. 3.

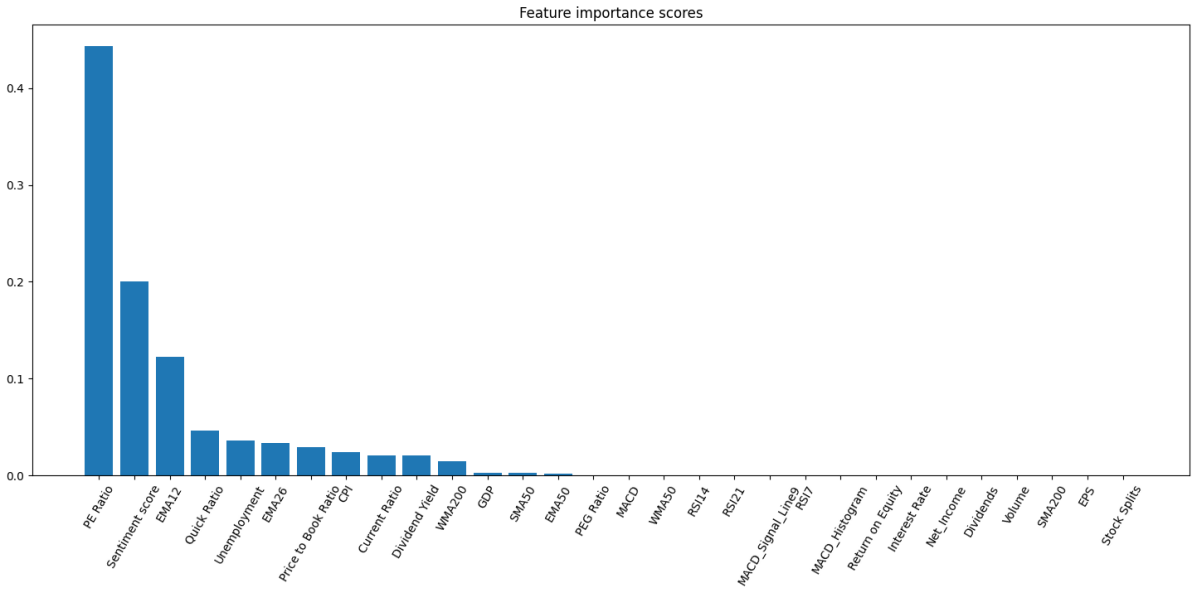

Рис. 3. Показники важливості ознак

На основі отриманих значень було відібрано 10 ознак з найбільшими значеннями відносного показника важливості для включення до нашої моделі.

Побудова нейронної мережі для прогнозування

Для реалізації рішення задачі прогнозування було обрано модель на базі *LSTM*.

Long Short-Term Memory (*LSTM*) мережі є спеціалізованим класом рекурентних нейронних мереж (*RNN*), що здатні вчитися на довготривалих залежностях. Вони були запропоновані Хохрайтером та Шмідгубером у 1997 році як рішення проблеми зникнення градієнта, яка характерна для стандартних *RNN* при роботі з довготривалими залежностями. *LSTM* мережі впроваджують концепцію «вентилів» (*gates*), що контролюють поток інформації [6]. Зв'язки між компонентами мережі *LSTM* приведені на рис. 4.

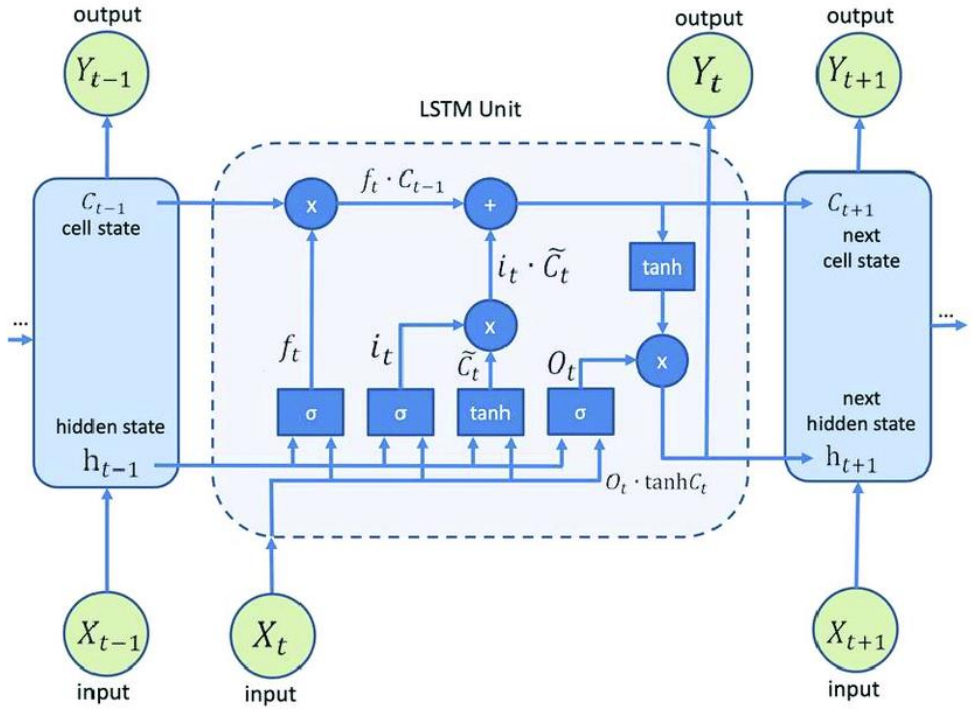

Рис. 4. Комірка *LSTM* у її загальній мережі

Структура *LSTM* складається з таких основних компонентів:

1. Комірка Пам'яті (*Cell State*)

Комірка пам'яті, C_t, є центральною частиною *LSTM* і працює як «носій» інформації через послідовні кроки часу. Вона має здатність додавати або видаляти інформацію через вентилі. Значення комірки вираховується за формулою (1):

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (1)$$

де:

- C_t – оновлений стан комірки на кроці часу *t*;
- C_{t-1} – попередній стан комірки;
- f_t – забувальний вентиль на кроці часу *t*;
- i_t – вхідний вентиль на кроці часу *t*;
- $\tilde{C}_t$ – кандидат на новий стан комірки на кроці часу *t*, отриманий через гіперболічний тангенс, що дає значення між -1 та 1.

2. Вентилі (*Gates*)

Існують три основні типи вентилів у *LSTM*:

Вентиль забуття (*Forget Gate*) – визначає, яка частина інформації з попереднього стану комірки повинна бути забута. Він використовує сигмоїдну функцію активації для генерації значень між 0 та 1, де 0 означає повне «забуття», а 1 означає повне «зберігання» інформації, і обчислюється наступним чином (2):

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

де:

- σ – сигмоїдна функція активації, що повертає значення між 0 та 1;
- W_f – вагова матриця забувального вентиля;
- $[h_{t-1}, x_t]$ – конкатенація попереднього прихованого стану h_{t-1} та поточного вводу x_t;
- b_f – вектор зсуву (*bias*) для забувального вентиля.

Вхідний вентиль (*Input Gate*) – визначає нову інформацію, яку слід додати до стану комірки. Вхідний вентиль також складається з сигмоїдної частини, яка вирішує, які значення оновлювати, та тангенсальної частини (3), яка створює новий кандидат на оновлення значень (4).

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3)$$

$$\tilde{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (4)$$

де:

- W_i, W_C – вагові матриці вхідного вентиля та кандидата стану комірки;
- b_i, b_C – вектори зсуву для вхідного вентиля та кандидата стану комірки.

Вихідний вентиль (*Output Gate*) – встановлює, яка частина інформації з комірки пам'яті буде використана як вихід на наступному кроці (5):

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

де:

- o_t – вихідний вентиль на кроці часу *t*;
- h_{t-1} – прихований стан на кроці часу *t-1*;
- W_o – вагова матриця вихідного вентиля;
- b_o – вектор зсуву для вихідного вентиля.

3. Прихований Стан (*Hidden State*)

Прихований стан, h_t (6), є виходом *LSTM* на кожному кроці часу, який може бути переданий до наступного кроку в послідовності або використаний для прогнозування.

$$h_t = o_t \cdot tanh\,(C_t) \quad (6)$$

де:

- h_t – прихований стан на кроці часу *t*, який може служити як вихід мережі для цього кроку часу або як ввод для наступного кроку;
- *tanh* – гіперболічний тангенс, функція активації, яка нормалізує значення стану комірки до діапазону між -1 та 1.

Робота *LSTM* на кожному кроці часу включає наступні етапи:

1. Визначення стану забувального вентиля f_t, що вирішує, яка інформація з попереднього стану комірки пам'яті C_{t-1} має бути забута.

2. Визначення вхідного вентиля i_t та кандидата на оновлення комірки пам'яті $\tilde{C}_t$, що вирішує, яку нову інформацію додати.

3. Оновлення комірки пам'яті з врахуванням інформації, яку треба забути та нової інформації, що додається.

4. Визначення вихідного вентиля o_t , який вирішує, яка частина оновленого стану комірки пам'яті буде використана у прихованому стані h_t.

5. Розрахунок нового прихованого стану, який базується на оновленому стані комірки пам'яті та вихідному вентилі.

Результатом є h_t, який може бути виходом моделі (наприклад, при прогнозуванні наступного значення в послідовності) або переданий до наступного кроку в послідовності. У випадку прогнозування фінансових часових рядів, h_t зазвичай проходить через додатковий шар (наприклад, повнозв'язний шар) перед тим, як здійснити фінальний прогноз.

Побудована модель визначається та характеризується параметрами, приведеними в табл. 1.

Таблиця 1. Параметри розробленої моделі

Параметри	**Значення**
Кількість входів	10
Кількість *LSTM* шарів	5
Кількість комірок *LSTM* шару	120
Функція активації *LSTM* шару	*tanh*
Крок часу	120
Розмір пакета	64
Оптимізатор	*Adam*
Функція втрат	Середня квадратична помилка
Рівень відкидання	0.3
Епохи	50

1. Кількість входів (*Number of inputs*) – це кількість вхідних змінних, які модель використовує для прогнозування. Наприклад, це можуть бути різні технічні показники на фондовому ринку.

2. Кількість комірок *LSTM* шару (*Number of units in LSTM layer*) – кількість нейронів в кожному з *LSTM* шарів. Більше нейронів може забезпечити більшу здатність моделі до навчання, але також збільшує ризик перенавчання.

3. Функція активації *LSTM* шару (*LSTM layer activation function*) – функція, яка застосовується до обчислення виходів нейронів *LSTM* шару.

4. Крок часу (*Time step*) – кількість часових точок, які модель використовує для прогнозування наступного значення.

5. Розмір пакета (*Batch size*) – кількість зразків даних, які обробляються за один крок навчання.

6. Оптимізатор (*Optimizer*) – алгоритм, який використовується для оновлення ваг моделі під час навчання.

7. Функція втрат (*Loss function*) – критерій, за яким модель оцінює помилки прогнозування та прагне мінімізувати.

8. Рівень відкидання (*Dropout rate*) – відсоток нейронів, які випадково ігноруються під час навчання, щоб запобігти перенавчанню.

9. Епохи (*Epochs*) – кількість повних проходів навчального набору даних, які виконуються під час навчання моделі.

Відповідно до приведених параметрів та архітектури моделі, її робота відбувається наступним чином. Вхідний набір навчальних даних є тривимірним вектором розміром (*None x* 120 *x* 10), де 120 є підібраним кроком часу, а 10 представляє кількість вхідних параметрів (ознак). Спершу ці дані поступають на вхідний шар, з

якого вони поступово передаються на задану кількість шарів *LSTM*, за кожним з яких слідує шар відкидання (*Dropout layer*). Обидва види шарів формують вихідні значення у вигляді вектору розміром (*None x* 120 *x* 120), і результуючий набір даних потрапляє на повнозв'язний шар (*Dense layer*), який і формує скалярне значення, що є результатом роботи моделі.

Навчання моделі

Один з ключових аспектів, що впливає на точність результатів прогнозування, є правильний підбір функції активації та втрат, а також оптимізатора, так як саме ці параметри безпосередньо впливають на формування результуючого значення та швидкість навчання моделі.

Функція активації є математичним перетворенням між входом, що подається на нейрон, та його виходом, що йде на наступний шар. Ця функція визначає, наскільки сильно активується нейрон, відповідно до вхідного сигналу. Вона додає нелінійності до моделі, що є необхідним для навчання складних шаблонів.

Для даної предметної області доречним вибором функції активації є гіперболічний тангенс (*tanh*), що виводить значення в діапазоні від -1 до 1. Він центрований в нулі, що робить навчання ефективнішим, оскільки середнє значення виходів шарів наближається до нуля, сприяючи стабільнішому градієнтному спуску. Перевагою гіперболічного тангенсу є його ефективність у мережах, де потрібно, щоб вихідні значення могли мати знак, що є особливо корисним у рекурентних нейронних мережах, які працюють із послідовностями даних.

Функція втрат визначає, наскільки точно модель працює під час навчання. Для задачі прогнозування ціни акції найкращим чином проявила себе середньоквадратична помилка (*Mean Squared Error, MSE*). *MSE* є мірою середнього квадрату відхилень прогнозованих значень від реальних. Оскільки *MSE* підносить помилки до квадрату, великі помилки «штрафуються» надмірно, що спонукає модель точніше прогнозувати значення.

При навчанні модель калібрує вагові коефіцієнти, маючи на меті зниження значення функції втрат. Спосіб безпосередньо калібрування визначається оптимізатором – алгоритмом, що використовується для зміни вагових коефіцієнтів та підвищення швидкості навчання. Для поставленої задачі було обрано оптимізатор *Adam* (*Adaptive Moment Estimation*), який поєднує ідеї двох інших популярних методів оптимізації: *RMSprop* і *Stochastic Gradient Descent with Momentum*. *Adam* адаптує швидкість навчання для кожного параметра моделі індивідуально, використовуючи оцінки перших та других моментів градієнтів. За рахунок своїх переваг, а саме адаптивності, ефективності на різноманітних даних, а також швидкості збіжності, у контексті прогнозування цін акцій, де модель може зіткнутися з даними, що сильно коливаються, *Adam* може допомогти моделі швидше адаптуватися та зменшити помилки прогнозування.

Важливим явищем, що може значно повпливати на точність та гнучкість роботи моделі в негативну сторону, є перенавчання (*overfitting*), що наступає, коли модель надто точно відтворює навчальний набір даних, втрачаючи при цьому здатність до узагальнення на нових даних. Це означає, що модель «запам'ятовує» навчальні дані, включаючи шум та випадковості, замість того, щоб вивчити загальні закономірності. Результатом перенавчання є висока точність на навчальному наборі даних, але погана продуктивність на тестовому наборі або на нових даних, що не брали участь у тренуванні.

Одним із способів недопущення перенавчання, задіяних у моделі, стало додавання шарів відкидання (*dropout*), що є реалізацією техніки регуляризації. У кожній ітерації навчання, визначений відсоток нейронів ігнорується, що запобігає їхній активації та впливу на вихід моделі. Така техніка ефективно симулює тренування великої кількості мереж та їхнє усереднення, що покращує загальну стійкість моделі до перенавчання. Таким чином, відкидання допомагає моделі підвищити рівень

робастності, змушуючи її не покладатися на будь-який конкретний набір нейронів, а навчатися більш рівномірно по всій мережі.

Результати навчання моделі з обраними параметрами представлені на рис. 5 в розрізі зміни функції втрат протягом етапу навчання з залученням окремого тестувального набору даних.

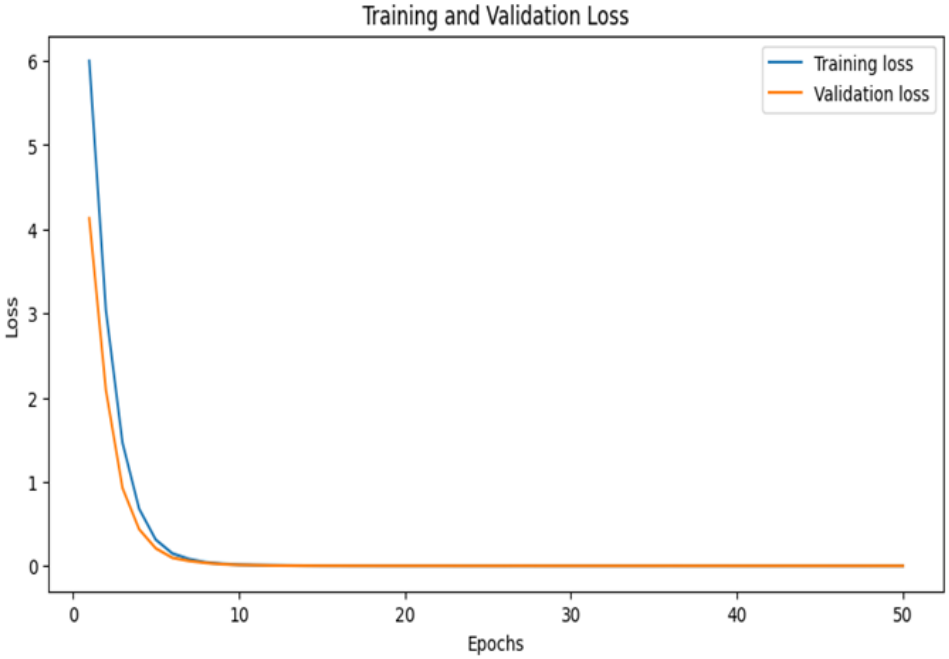

Рис. 5. Зміна значення функції втрат протягом навчання моделі

Оцінювання результатів

Для проведення оцінки ефективності роботи моделі необхідним етапом є вибір базової моделі для порівняння результатів. Ціни акцій по своїй природі близько слідують гіпотезі про «випадкове блукання» (*Radom walk hypothesis*), це означає, що зміни цієї величини є близькими за своїм характером до випадкових. Як відомо, для таких величин найкращим прогнозуванням є так зване «наївне прогнозування» (*Naive forecast*), суттю якого є копіювання останнього відомого значення прогнозованої величини на весь часовий проміжок прогнозування. Тому оцінку результатів роботи розробленої моделі було вирішено проводити у порівнянні з моделлю наївного прогнозування.

Для оцінки якості роботи запропонованої моделі та базової (моделі наївного прогнозування) було вибрано наступні метрики: *MSE*, *RMSE*, *MAPE* та *R*2. *MSE* було вирішено включити до набору обрахованих метрик для аналізу оскільки ця метрика є тою, що мінімізуєтьсян а етапі навчання нашої моделі. Вибір *RMSE* для включення до метрик обумовлений тим, що метрики *MSE* та *RMSE* відображають якість роботи моделі різним способом: *MSE* дає приблизне уявлення про величину помилки, а *RMSE* є квадратним коренем із *MSE*. Цей квадратний корінь є значущим, оскільки він означає, що *RMSE* має той самий масштаб, що й вихідні дані, що робить його більш зручним для інтерпретації, так як відображає помилку в тих самих одиницях, що й вихідна змінна. Крім цього, застосування у оцінці якості моделі комбінації метрик *MSE* і *RMSE* обумовлене тим, що помилки зводяться в квадрат перед усередненням, *RMSE* надає відносно високу вагу великим помилкам, це означає, що *RMSE* є чутливим до викидів.

Однак, оскільки помилки зводяться в квадрат перед усередненням, як *MSE*, так і *RMSE* в цілому більш чутливі до викидів, ніж інші показники, такі як середня абсолютна похибка. Тому для оцінки якості розробленої моделі було вирішено також включити метрики *MAPE* та *R*2.

Результати

Після проведення навчання моделі, на вхід були передані дані тестувального набору, що включає інформацію за 30 робочих днів, і отримано прогнозоване значення ціни закриття для кожної з поданих дат. Результати прогнозування разом з базовим варіантом для порівняння (наївним прогнозом) приведені на рис. 6. На основі отриманих значень, були вираховані попередньо обрані метрики для прогнозу моделі та базового сценарію (табл. 2).

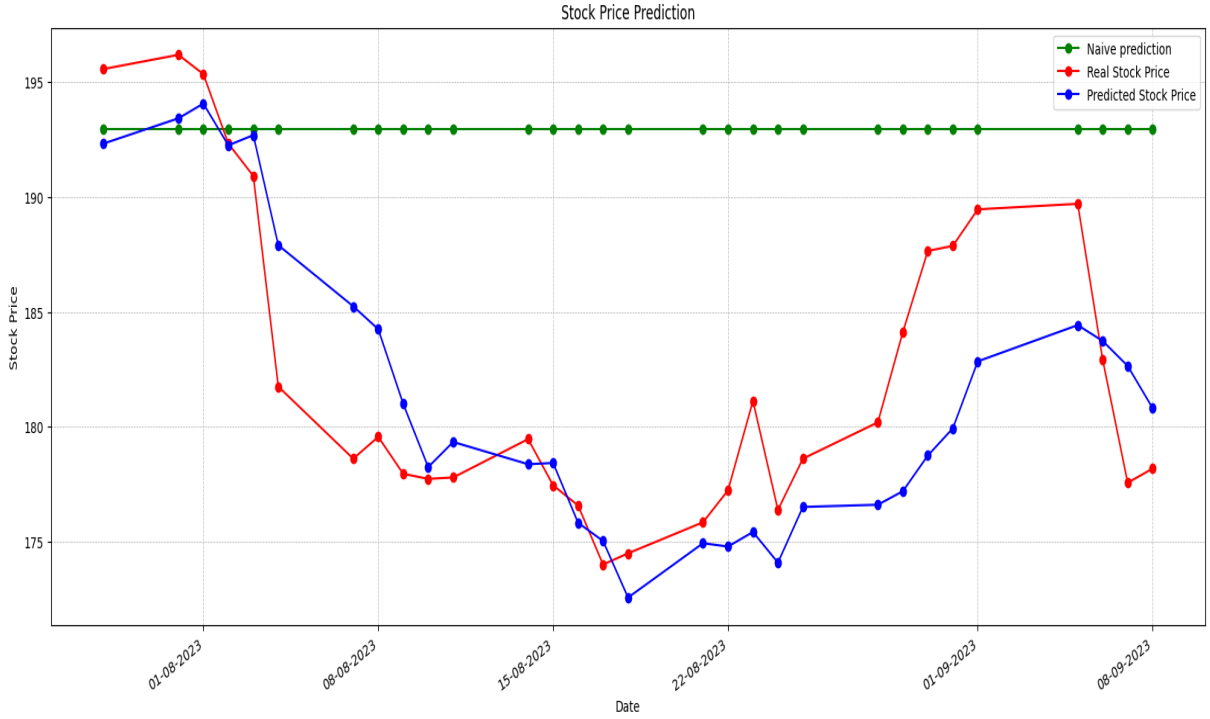

Рис. 6. Результати прогнозування моделі

Таблиця 2. Обчилені метрики прогнозування

Метрика	Розроблена модель	Наївний прогноз
R^2	0,714	-2,527
MAPE	0,015	0,062
MSE	12,584	155,143
RMSE	3,547	12,456

Як бачимо з приведених вище результатів, розроблена модель з використанням *LSTM* змогла перевершити базову модель прогнозування за усіма обраними метриками, довівши свою високу ефективність у порівнянні з наївним прогнозом. Розроблена модель досягає необхідного рівня близькості із дійсними даними, а також дає змогу виділяти тренди і передбачати напрямок зміни ціни.

Висновки

В даній роботі було запропоновано підхід, який ставить за ціль виправити ключові недоліки існуючих методик та полягає у використанні системного підходу до включення різних класів показників для прогнозування цін фінансових інструментів, включаючи технічні, фундаментальні та макроекономічні показники. Такий набір даних також було доповнено за рахунок врахування текстової інформації. Застосування технік *NLP* та аналізу настрою дозволило інтегрувати новини, звіти та інші текстові джерела інформації, які впливають на ринкові тенденції та відображають актуальну повістку, яка впливає на цільову величину.

У поєднанні результатів цього комплексного підходу до формування набору даних, на основі яких здійснюється

прогноз, з методами глибокого навчання було створено нейронну модель на базі *LSTM* для прогнозування цін фінансових інструментів на фондовому ринку. Для оцінки її ефективності було обрано набір метрик, а також проведено порівняльний аналіз результатів з базовою моделлю (наївним прогнозуванням). Отримані результати показують високу точність роботи створеної моделі, як і те, що розроблена модель значним чином перевершує у своїй точності модель наївного прогнозування.

Проведений аналіз метрик якості роботи розробленої моделі на основі глибоких нейронних мереж і моделі наївного прогнозування, представлений у таблиці 2, вказує на значні переваги застосування розробленої нами моделі. Зокрема, коефіцієнт детермінації R^2 для розробленої моделі складає 0,714, що істотно перевищує показник моделі наївного прогнозу, що має негативне значення -2,527. Це свідчить про високу спроможність моделі відтворювати динаміку цін фінансових інструментів і здатність ефективно прогнозувати майбутні значення. Щодо середньої абсолютної відсоткової помилки (*MAPE*), то для розробленої моделі цей показник становить всього 0,015 у порівнянні з 0,062 для наївного прогнозу, що підкреслює значно вищу точність прогнозів. Крім того, значення середньої квадратичної помилки (*MSE*) і кореня середньоквадратичної помилки (*RMSE*) для розробленої моделі є значно нижчими (відповідно 12,584 та 3,547), ніж для наївного прогнозу (155,143 та 12,456), що також свідчить про вищу точність та надійність розробленої моделі. Таким чином, застосування нашої моделі як засобу аналізу фінансових ринків демонструє суттєву перевагу перед традиційними методами прогнозування, що відкриває нові перспективи для розробки та використання подібних технологій у фінансовому аналізі.

Предметом майбутніх досліджень та покращень запропонованого підходу та розробленої моделі є вдосконалення етапу обробки текстової інформації з додаванням класифікації текстових даних за джерелами та авторитетністю авторів та присвоєння їм відповідних вагових коефіцієнтів у доповнення до вже використовуваного обчислення показників тональності тексту. Окрім цього вартим уваги є розширення базового набору залучених показників (предикторів) задля кращого моделювання зв'язків між чинниками, що формують вихідну величину. Також наступним кроком досліджень є врахування кореляцій між цінами акцій різних компаній у рамках одної індустрії, так як ціноутворення фінансових інструментів не відбувається ізольовано, а має наслідки для множинного числа залучених сторін.

Література

1. Lu W., Li J., Li Y., Sun A., Wang J. A CNN-LSTM-based model to forecast stock prices. *Complexity*. 2020. P.1–10.

2. Jarrah M., Derbali M. Predicting Saudi Stock Market Index by Using Multivariate Time Series Based on Deep Learning. *Applied Sciences*. 2023. V. 13(14). 8356.

3. Akita R., Yoshihara A., Matsubara T., Uehara K. Deep learning for stock prediction using numerical and textual information. *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)* / Okayama, Japan, 2016. P. 1–6.

4. Drake P.P., Fabozzi F.J. The basics of finance: An introduction to financial markets, business finance, and portfolio management. John Wiley & Sons, 2010. 672 p.

5. Gravetter F.J., Wallnau L.B., Forzano L.A.B., Witnauer J.E. Essentials of statistics for the behavioral sciences. Cengage Learning, 2020. P. 490–512.

6. Gers F.A., Schraudolph N.N., Schmidhuber J. Learning precise timing with LSTM recurrent networks. *Journal of machine learning research*. 2002. Iss. 3. P. 115–143.

Саттарова М.Л., Тхорик В.Б., Савченко А.С., Моденов Ю.Б.

ПРОГНОЗУВАННЯ ЦІН ФІНАНСОВИХ ІНСТРУМЕНТІВ ЗА ДОПОМОГОЮ ЗАСОБІВ ГЛИБОКОГО НАВЧАННЯ

Робота присвячена проблемі прогнозування цін акцій на фондовому ринку, актуальності якої значно зросла у сучасному світі, будучи важливою складовою процесів ведення фінансової діяльності та прийняття обгрунтованих інвестиційних рішень. Було проведено огляд та порівняльний аналіз методик, запропонованих у існуючих дослідженнях, виділено наявні у них недоліки та недопрацювання. На основі цього було запропоновано новий підхід для вирішення цієї задачі. Запропонований підхід грунтується на врахуванні комплексного набору факторів для прогнозування, включаючи технічні показники, дані фундаментального аналізу та макроекономічні фактори, використанні системного підходу для відбору предикторів (факторів) для прогнозування та включення у модель, впровадженні сучасних метожик feature engineering та feature selection, видалення шуму у вхідних даних, застосування технік NLP та аналізу настрою для інтеграції текстових даних, які впливають на ринкові тенденції, підвищуючи таким чином точність моделювання ринкових процесів. Ці засади було скомбіновано з методиками машинного та глибокого навчання, здатними враховувати часові послідовності даних та складні взаємозв'язки і залежності між ними, та побудовано нейронну модель для прогнозування цін акцій. Результати тестування моделі та отримані значення метрик точності роботи розробленої моделі показують її високу точність у порівнянні з базовою моделлю, обраною для порівняння, а також доводять ефективність використання запропонованого підходу.

Ключові слова: *прогнозування, фінансові показники, глибоке навчання, Long Short-Term Memory, Natural Language Processing.*

Sattarova M.L., Tkhoryk V.B., Savchenko A.S., Modenov Yu.B.

FORECASTING PRICES OF FINANCIAL INSTRUMENTS USING DEEP LEARNING METHODS

This study is devoted to the problem of forecasting financial instruments' prices on the stock market, the relevance of which has increased significantly in the modern world, and which is an important component of the processes of conducting financial activities and making informed investment decisions. A review and comparative analysis of the methods proposed in the existing studies was conducted, their shortcomings and shortcomings were revealed. Based on this, a new approach to solving this problem was proposed. The proposed approach is based on taking into account a complex set of factors for forecasting, including technical indicators, fundamental analysis data and macroeconomic factors, the use of a systematic approach to the selection of predictors (factors) for forecasting and inclusion in the model, the introduction of modern feature engineering and feature selection techniques, noise removal in input data, applying NLP techniques and sentiment analysis to integrate textual data that influence market trends, thus increasing the accuracy of modeling market processes. These principles were incorporated with machine and deep learning techniques capable of taking into account time sequences of data and complex relationships and dependencies between them, and a predictive model was built for forecasting stock prices. The results of model testing and evaluation of obtained performance accuracy metrics' values for the developed model show it's higher accuracy in comparison with the base model chosen for comparison, and also prove the effectiveness of using the proposed approach.

Keywords: *forecasting, financial indicators, deep learning, Long Short-Term Memory, Natural Language Processing.*