

МАТЕМАТИЧНІ МОДЕЛІ СУСПІЛЬНИХ ПРОЦЕСІВ ДЛЯ АНАЛІЗУ ВПЛИВУ ВІЙНИ НА ДИНАМІКУ РОЗВИТКУ ЕКОНОМІКИ ТА ЕКОНОМІЧНИХ ПОКАЗНИКІВ

Діхтяр А.В.¹, Лопатін О.К.²

Національний технічний університет України «Київський політехнічний інститут
ім. Ігоря Сікорського»

¹ wdich12@gmail.com, ² lopatinalexey142@gmail.com

На жаль, у сучасному світі військові конфлікти стають невід'ємною складовою геополітичного ландшафту, породжуючи виклики та наслідки, що відчутні у багатьох аспектах глобальної економіки та життя суспільства. Війни призводять до жахливих людських жертв, трагедій, економічного занепаду та руйнування. Одним із суттєвих завдань, яке стоїть перед науковим співтовариством, є розуміння впливу війни на розвиток економіки та ключових економічних показників. Важкі випробування, що виникли внаслідок світових війн минулого сторіччя, неабияк прискорили розвиток математичних методів та моделей для аналізу та прогнозування динаміки цих процесів, включаючи лінійні, нелінійні та перехідні явища, в тому числі економічні процеси під впливом зовнішніх факторів. Метою роботи є розробка інструментарію для прогнозування та проведення порівняльного аналізу макроекономічних показників під впливом війни. Основним результатом дослідження є побудова системи підтримки прийняття рішень (СППР) для прогнозування та подальшого аналізу економічних процесів.

Ключові слова: макроекономічні процеси, регресійні моделі, рекурентні нейронні мережі, прогнозування, війна, внутрішній валовий продукт.

1. ВСТУП

Війна – це жахливе явище, що приносить смерть, руйнування та страждання мільйонам. Люди втрачають свої життя, країни зазнають великих економічних збитків. Війна порушує мирне співіснування та залишає разючий слід у психології людей, особливо в дітей. Вона спричиняє непоправної шкоди природному середовищу і сприяє зростанню глобальної напруги. Глобальні військові конфлікти мають надзвичайно вагомий вплив на світову економіку, створюючи серію складних викликів та наслідків для різних країн та галузей. Цей вплив стосується різних аспектів економіки та суспільства. Війни спричиняють економічний спад через руйнування інфраструктури, збитки виробництву та зниження споживчого попиту. Нестабільність та невизначеність під час військового конфлікту є важливими факторами, які гальмують економічний розвиток.

Населення найбільше страждає від гуманітарних катастроф і соціальних втрат, які виникають внаслідок війни. Руйнування виробництва та інфраструктури призводять до безробіття та погіршення рівня життя, що може мати довгострокові соціальні наслідки. Війна в Україні значно вплинула на економічну активність в країні, призводячи до прогнозованого спаду ВВП та інших економічних показників через воєнні обставини. Різні сектори економіки,

такі як сільське господарство, промисловість та торгівля, відчули серйозні труднощі через перерви в постачанні та зниження попиту. Регіони, які стали ареною активних бойових дій, виявилися особливо вразливими, зазнавши руйнувань і великих втрат. Національна бюджетна система також відчула складність ситуації через значний зріст витрат на потреби армії, соціальні програми та відновлення пошкоджених регіонів. Збільшення державного боргу та використання нових інструментів фінансування, таких як введення військових облігацій, свідчать про необхідність ефективного фінансового управління в умовах воєнного конфлікту.

Соціальний аспект війни неминуче впливає на зайнятість, рівень життя та еміграцію населення, ставлячи під загрозу соціальну стабільність країни. Незважаючи на це, важливо визначити стратегії для подальшого відновлення та розвитку. Заходи, такі як надання податкових пільг, приваблення інвестицій, розвиток важливих секторів економіки, а також соціальна підтримка населення, стають ключовими для подолання викликів воєнного періоду. Українське суспільство, проявляючи солідарність та патріотизм, потребує ефективного управління, стратегічного планування та співпраці з міжнародними партнерами для подальшого економічного та соціального відновлення після нашої перемоги.

2. ВИБІР ТА ОПИС МЕТОДІВ ДЛЯ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ

Регресійні моделі відіграють ключову роль для аналізу взаємозв'язків між залежними та незалежними змінними. Ці моделі визначаються як фундаментальний інструмент для ретельного аналізу, що відбувається у сфері дослідження зазначених залежностей, розширюючи наше розуміння обсягу даних і надаючи можливість ефективного здійснення прогнозів на підставі цих виявлених залежностей. Важливо пам'ятати, що вибір конкретної моделі повинен бути чітко налаштованим під характеристики використовуваних даних у конкретному контексті роботи [6].

Оцінка якості прогнозів, безперечно, є ключовим етапом, і вона виконується за допомогою різноманітних критеріїв, таких як середньоквадратична похибка (MSE), середня абсолютна похибка (MAE), коефіцієнт детермінації (R-squared) та інші. Вибір конкретного критерію визначається специфікою завдання та конкретними аспектами прогнозу, які привертають увагу дослідника. Наприклад, MSE ефективно реагує на великі похибки, в той час як MAE краще враховує середні значення похибок [3].

Нейронні мережі, зокрема рекурентні та згорткові, мають здатність виявляти складні залежності в наборах даних та застосовувати їх для точного прогнозування. Ці мережі показують вражаючу ефективність в різних областях, таких як обробка тексту, зображень, аудіо та часові послідовності, забезпечуючи результативність у розв'язанні складних завдань, що вимагають аналізу великих обсягів даних.

Оцінка якості моделі включає використання критеріїв адекватності, таких як Критерій Акайке, Байєсівський інформаційний критерій (BIC) та Критерій Дарбіна-Ватсона (DW). Важливо розуміти, що вибір конкретної моделі повинен точно відповідати завданням та специфікаціям використовуваних даних. Деякі критерії можуть виявитися кращими, для обробки обширних наборів даних, інші можуть проявити високу ефективність для аналізу складних моделей [4].

Вищезазначені аспекти є ключовими для успішного аналізу та прогнозування даних, і правильне їх використання має потенціал істотно покращити якість отриманих результатів. Обґрунтованість та уважність у проведенні аналізу моделей є вирішальними для досягнення максимально ефективних результатів. Застосування різноманітних методів та критеріїв для оцінювання та порівняння моделей є важливим завданням, і їх вибір обумовлений конкретною ситуацією. Критерії якості, а також мережі та інструменти для аналізу та прогнозування

постійно еволюціонують, тому важливо слідкувати за новими тенденціями та використовувати їх для постійного удосконалення результатів. Загалом важливо підкреслити вагомість ретельного аналізу, вибору відповідних підходів та моделей для проведення аналізу та прогнозування даних [7].

3. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Основними даними для моделювання та подальшого прогнозування є щомісячні дані внутрішнього валового продукту України із січня 2002 по грудень 2021 року [1]. Допоміжними даними для побудови прогнозів виступали також й інші макроекономічні показники, такі як індекс споживчих цін, індекс промислового виробництва, ціни на газ, податкові надходження, неподаткові надходження та офіційний курс долара.

Розглянемо графік вхідних даних ВВП (рис. 1).

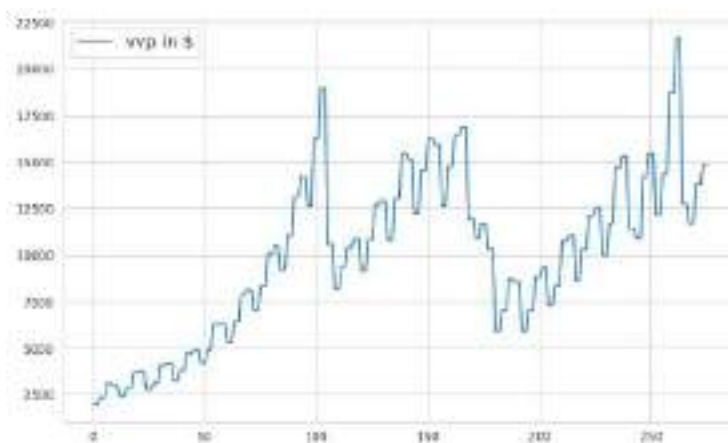


Рисунок 1. Графічне зображення даних

За допомогою системи EViews для вхідних даних були побудовано ряд авторегресійних моделей, в тому числі й вищих порядків [3]. По ходу покращення моделей, вони ускладнювалися та набували складових ендогенних регресорів і нелінійних складових, таких як тренди та періодичні функції. По ходу вдосконалення моделей, їхні параметри та статистики ставали все кращими й кращими. Найкраща модель по всім статистикам це авторегресія 25 порядку із додаванням усіх ендогенних регресорів також із лагом 25 порядку, із додаванням тренду, періодичної функції та квадратичних складових ендогенних регресорів. Графік оцінених найкращою моделлю даних (рис. 2).

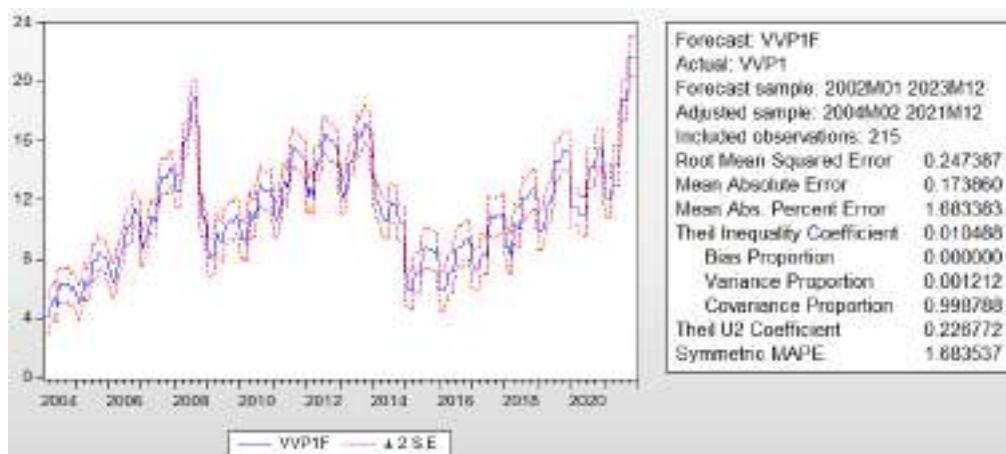


Рисунок 2. Графік оцінених найкращою моделлю даних

Як можемо бачити, характеристики однокрокового прогнозу дуже обнадіюючі.

Також, за допомогою штучних рекурентних нейронних мереж (використовуючи інструмент Google Colab), було побудовано ряд експериментальних одношарових та багатошарових нейронних мереж, використовуючи декілька видів рекурентних нейронів, таких як LSTM та GRU [5]. Не зважаючи на відносно невелику кількість прикладів для навчання, згорткові нейронні мережі показали себе не досить ефективно. Що можна легко пояснити, адже згорткові нейронні мережі мають ефективне прикладне застосування в розпізнаванні зображень та відео, рекомендаційних системах тощо. Процес навчання та статистики якості прогнозів найкращої нейронної мережі подано нижче (рис. 3).

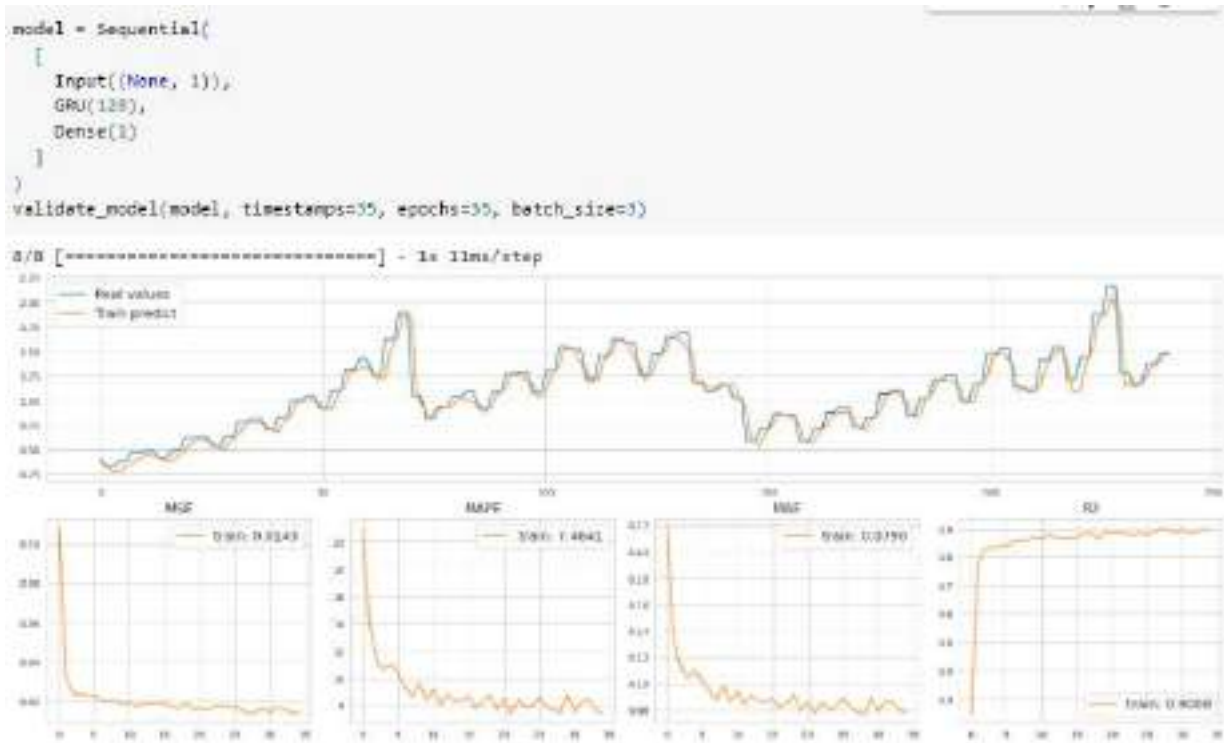


Рисунок 3. Процес навчання та статистики якості прогнозів найкращої нейронної мережі

Порівняємо статистичні характеристики найкращих моделей авторесії та рекурентних нейронних мереж. Будемо оцінювати за наступними критеріями: коефіцієнт детермінації, критерій Дарбіна-Уотсона, сума квадратів похибок, MSE, MAE, MAPE та коефіцієнт Тейла [4] (Табл. 1).

Таблиця 1. Порівняння статистичних параметрів найкращих моделей

Тип моделі	Характеристики моделі			Характеристики прогнозу			
	R^2	$\sum e^2(k)$	DW	MSE	MAE	MAPE	U
AR(25)+REG	0,995	13	2	0,061	0,174	1,683	0,228
RNN	0,972	278		0,401	0,462	4,561	

Як можемо бачити, авторегресійна модель по всіх розглянутих показниках є кращою за модель штучної рекурентної нейронної мережі. Це легко пояснюється даними для навчання: авторегресійна модель приймає на вхід не лише власні часові данні, а ще й ендогенні данні, в

той час як нейронна мережа навчається на своїх попередніх даних, яких, до того ж, недостатньо для гарного навчання останньої. Проте, не зважаючи на недоліки, характеристики адекватності та однокрокових прогнозів обох моделей є дуже хорошими, і це дає нам підставу вважати, що прогнози цих моделей на майбутнє будуть близькими до реальності [2].

За результатами прогнозування авторегресійною моделлю можна припускати, яким було б ВВП України, якби не повномасштабне вторгнення. За 2022 рік ВВП України склав 159,124 мільярдів доларів. За спрогнозованими моделлю даними, отримуємо, що ВВП України за 2022 рік міг би бути 257,85 мільярдів доларів. Падіння ВВП за цей рік в складає 99 мільярдів доларів, або 38.3% (рис. 4).

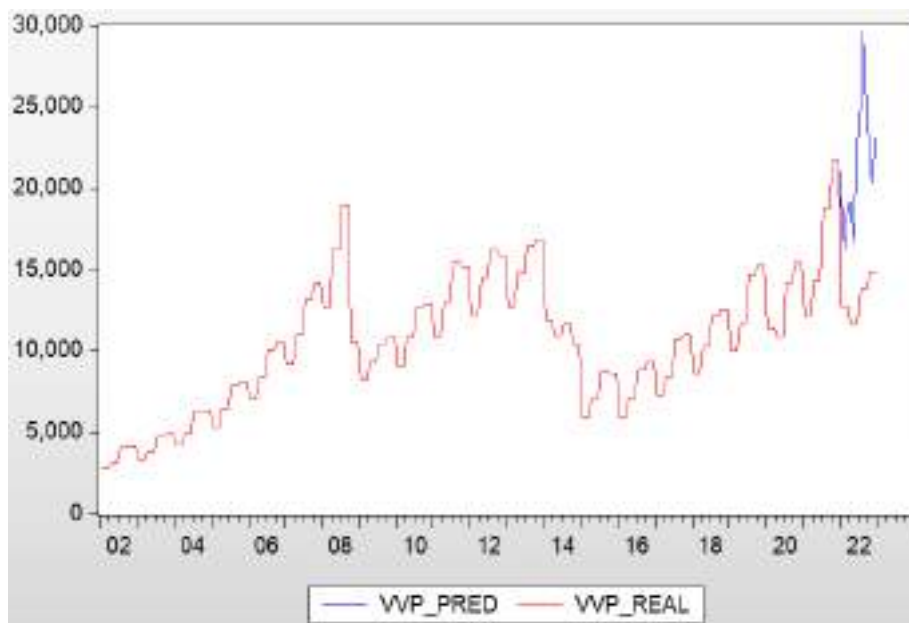


Рисунок 4. Графік розриву прогнозованого регресійною моделлю та реального ВВП

За результатами прогнозування нейронною мережею отримуємо наступне: за 2022 рік ВВП України склав 159,124 мільярдів доларів; по спрогнозованим моделлю даним, маємо, що ВВП України за 2022 рік міг скласти 225,169 мільярдів доларів. Падіння ВВП за останній квартал в складає 20,5 мільярдів доларів, або 31,7% (рис. 5).

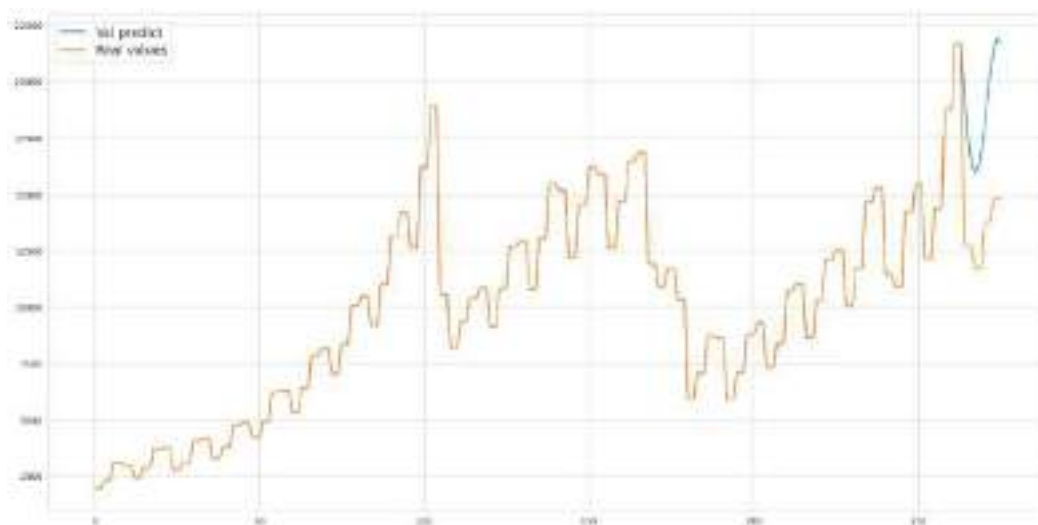


Рисунок 5. Графік розриву прогнозованого RNN та реального ВВП за 2022 рік

Аналізу окремих показників, таких як бюджетні надходження України, виявив цікаві висновки: до початку повномасштабного вторгнення, приблизно 80% податкових надходжень до державного бюджету надходило від внутрішніх джерел, в той час як зовнішні інвестиції та трансї становили лише 1,5%. Однак після першого року повномасштабного вторгнення ця динаміка змінилася: частка внутрішніх податкових надходжень знизилася до 53%, а надходження від іноземних інвестицій та трансїв зростали до 27,5%. За даними на 2023 рік, частка внутрішніх податкових надходжень в бюджет складає всього 45%, в той час як надходження від іноземних джерел становлять практично 21%.

4. ВИСНОВКИ

Це дослідження спрямоване на ретельний аналіз та вивчення впливу війни на динаміку розвитку економіки та макроекономічних показників України, у межах якого було проведено обширне вивчення різних аспектів, включаючи огляд існуючих методів моделювання та прогнозування, розгляд критеріїв якості для оцінки моделей, методи оцінювання параметрів моделі, та обробку важливих макроекономічних показників.

Були проаналізовані окремі макроекономічні показники, що дозволило виявити руйнівний вплив військового конфлікту в Україні на економічні процеси, характер наповнення бюджету та рівень міжнародної підтримки.

Дослідження також включало ряд експериментів з розробки прогнозів для різних макроекономічних показників у випадку, якби вторгнення не відбулося. Отримані результати вражаючі: розрив між реальним і прогнозованим внутрішнім валовим продуктом України становить від 66 до 99 мільярдів доларів, що відповідає 41,5% до 62% від реального ВВП за 2022 рік. У 2014 році ситуація була подібною: виявлено розрив внутрішнього валового продукту України між фактичними та прогнозованими показниками на рівні 77,5 мільярдів доларів. Це становило близько 57,6% у порівнянні з реальним ВВП за 2014 рік. Ці цифри є надзвичайно високими!

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Валовий внутрішній продукт (ВВП) України [Електронний ресурс]: Режим доступу до ресурсу: <https://index.minfin.com.ua/ua/economy/gdp/>
2. Бідюк П.І., Половцев О.В. Аналіз та моделювання економічних процесів перехідного періоду. К.: НТУ КПП, 1999. 230 с.
3. Бідюк П. І., Романенко В. Д., Тимошук О. Л. Аналіз часових рядів: навч. посіб. / ННК «Інститут прикладного системного аналізу» Національний технічний університет України «Київський політехнічний інститут», 2010. 317 с.
4. Бідюк П. І. Економетричний аналіз часових рядів. Київ: Політехніка, 2007. 250 с.
5. Бідюк П. І. Проектування комп'ютерних інформаційних систем підтримки прийняття рішень: навч. посіб. / ННК «Інститут прикладного системного аналізу» Національний технічний університет України «Київський політехнічний інститут», 2010. 340 с.
6. Ставицький А. В. Навчально-методичний комплекс з курсів «Прогнозування» та «Фінансове прогнозування». Київ: Центр учб. літ., 2006. 107 с.
7. Половцев О. В. Системний підхід до моделювання, прогнозування та управління фінансово-економічними процесами. Донецьк: Східний видавничий дім, 2009. 286 с.

СИСТЕМА ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ ОПТИМІЗАЦІЇ РЕКЛАМНИХ КАМПАНІЙ ПІДПРИЄМСТВА НА ОСНОВІ МЕТОДУ МОДЕЛЮВАННЯ ВПЛИВУ З ЗАЛЕЖНИМ ПРЕДСТАВЛЕННЯМ ДАНИХ

Заїка Б.Ю.¹, Терент'єв О.М.²

Національний технічний університет України «Київський політехнічний інститут
імені Ігоря Сікорського», Київ, Україна

¹ zaikabohdan3@gmail.com, ² o.terentiev@gmail.com

У сучасному світі динамічного розвитку технологій та конкурентної боротьби у сфері бізнесу, прогресивні методи аналізу даних стають крайньою необхідністю. Рекламні кампанії є невід'ємною складовою будь якого підприємства та потребують правильного підходу для раціонального використання ресурсів та уникнення негативного впливу на клієнтів. Проблемою традиційних підходів моделювання в даному контексті є фокусування на прогнозі ймовірності виконання цільової дії користувачем після комунікації з ним. Моделювання впливу, на відміну від традиційних моделей, прогнозує вплив взаємодії з користувачем на ймовірність виконання ним цільової дії. Використання такого підходу дозволяє раціональніше використовувати ресурси компанії, фокусуючи комунікацію на користувачів, взаємодія з якими матиме найбільший позитивний вплив на виконання ними цільової дії.

Ключові слова: моделювання впливу, система підтримки прийняття рішень, рекламні кампанії, кероване навчання.

1. ВСТУП

У сучасному світі великої конкуренції та постійних змін, використання даних для розуміння поведінки споживачів та ефективного впливу на них стали критичними завданнями для бізнесу. Рекламні кампанії мають величезне значення для підприємства, оскільки вони забезпечують залучення уваги клієнтів, збільшення продажів, підтримку конкурентоспроможності та комунікацію з аудиторією. Реклама дозволяє підтримувати зв'язок зі споживачами, повідомляти їх про новинки, акції, зміни та відповідати на їхні запитання. Проте нераціональне використання цього інструменту бізнесом може нести втрати у вигляді ресурсів та клієнтів, через зайві комунікації або невдалу персоналізацію реклами.

Традиційні методи класифікації прогнозують ймовірність належності користувача до класу користувачів, які зробили цільову дію після взаємодії з ними. На основі цих прогнозів часто приймаються рішення щодо взаємодії з класифікованими особами. Однак справжньою метою рекламних кампаній має бути визначення різниці у виконання цільової дії користувачем з та без взаємодії з ним. Традиційні методи класифікації не використовують інформацію стосовно контрольних груп і тому мають обмежене застосування в цьому контексті.

На відміну від них, моделювання впливу дозволяє включати контрольну групу та спрямовується на явне моделювання різниці в ймовірності результату між двома групами, тому воно набагато краще підходить для аналізу потенційних отримувачів реклами. Крім того,

моделі впливу дозволяють безпосередньо ідентифікувати користувачів, взаємодія з якими є найефективнішою [1]. Такий підхід до комунікації з клієнтами дозволяє раціонально використовувати ресурси та уникати контакту з клієнтами, яких реклама може відштовхнути від виконання цільової дії. Саме в цьому і полягає актуальність використання моделювання впливу для прийняття рішень стосовно рекламних кампаній підприємства: можливість спрогнозувати як взаємодія вплине на користувача, щоб уникнути небажаної комунікації та розуміти з якими користувачами вигідніше взаємодіяти для досягнення цілі.

Метою роботи є розробка системи підтримки прийняття рішень (СППР) [2] на основі методу моделювання впливу з залежним представленням даних (DDR – Dependent Data Representation) [3], що стане невід’ємним інструментом аналітиків та бізнес користувачів підприємства для швидкого аналізу та створення моделей впливу для проведення оперативних та персоналізованих рекламних кампаній.

2. МЕТОДИКА ПОРІВНЯЛЬНОГО АНАЛІЗУ

Для порівняння методу залежного представлення даних з іншими в першу чергу треба визначити методику порівняння, що дозволить об’єктивно оцінити кожен модель та визначити їх слабкі та сильні сторони. Методика включає наступні кроки:

1. Вибір методів для порівняння
2. Визначення метрик ефективності
3. Збір та підготовка даних
4. Навчання, тестування та порівняння моделей
5. Висновки

Для порівняльного аналізу використано методи моделювання впливу одного учня (S-Learner) [4, 5] та трансформації змінної класів (Class Variable Transformation approach або Z-Learner) [1, 5]. Оскільки всі методи, що порівнюються, є методами мета-навчання (Meta-learners), які використовують звичайні моделі класифікації з певним принципом подачі на вхід даних, то для справедливості порівняння всі методи будуть порівнюватись на одному виді моделей класифікації - XGBoost класифікаторі.

В рамках аналізу зосереджено увагу на метриках оцінки ефективності моделі таких як стовпчикова діаграма впливу за перцентильними рангами (uplift by percentile bar chart), вплив на топ k% (uplift at top k%, для поточного аналізу розраховано на топ 30%), середньозважений вплив (weighted average uplift), крива Квіні (Qini curve) та коефіцієнт Квіні (Qini Coefficient) [5].

Стовпчикова діаграма впливу за перцентильними рангами будується за наступним алгоритмом:

1. Користувачі сортуються за спаданням спрогнозованого значення впливу
2. Відсортовані дані діляться на перцентилі
3. В кожному перцентилі окремо оцінюється вплив як різниця між середнім значенням цільової змінної в цільовій та контрольній групах.

Для побудови кривої Квіні необхідно відсортувати дані за спаданням спрогнозованого значення впливу та побудувати графік за наступною формулою [6]:

$$Qini\ curve(t) = Y_t^T - \frac{Y_t^C N_t^T}{N_t^C}, \text{ де}$$

- t – кількість включених в комунікацію користувачів;
- Y_t^T, Y_t^C – кількість виконаних цільових дій в цільовій (T – Target) та контрольній (C – Control) групі відповідно;
- N_t^T, N_t^C – кількість користувачів в цільовій та контрольній групі відповідно.

Для розрахунку впливу на топ k% необхідно скористатись наступною формулою на топ k% користувачах за спрогнозованим значенням впливу [7]:

$$Uplift\ at\ top\ k\% = \frac{Y_{top\ k\%}^T}{N_{top\ k\%}^T} - \frac{Y_{top\ k\%}^C}{N_{top\ k\%}^C}, \text{ де}$$

- $Y_{top\ k\%}^T, Y_{top\ k\%}^C$ - кількість виконаних цільових дій в цільовій та контрольній групі відповідно серед топ k% користувачів за спрогнозованим значенням впливу;
- $N_{top\ k\%}^T, N_{top\ k\%}^C$ - кількість користувачів в цільовій та контрольній групі відповідно серед топ k% користувачів за спрогнозованим значенням впливу;

Середньозважений вплив є числовим відображенням інформації зі стовпчикової діаграми впливу за перцентильними рангами та розраховується за наступною формулою:

$$Weighted\ average\ uplift = \frac{1}{\sum_{i=1}^{10} N_i^T} \sum_{i=1}^{10} N_i^T uplift_i, \text{ де}$$

- N_i^T – розмір цільової групи в i-ому перцентилі;
- $uplift_i$ – вплив в i-ому перцентилі.

Коефіцієнт Квіні є числовим відображенням інформації з кривої Квіні розраховується за наступною формулою:

$$Qini\ coefficient = \frac{S_{model}}{S_{ideal}}, \text{ де}$$

- S_{model} – площа між кривими Квіні побудованої та випадкової моделей;
- S_{ideal} – площа між кривими Квіні ідеальної та випадкової моделей;

Для тренування та порівняння моделей взято анонімізований набір даних телекомунікаційної компанії [8], який містить в собі наступні ознаки:

- id – ідентифікатор користувача;
- X_1, \dots, X_{50} – 50 анонімізованих ознак, які описують користувача;
- $treatment_group$ – до якої групи відноситься користувач (цільова чи контрольна);
- $conversion$ – чи зробив користувач цільову дію.

Дані розбито випадковим чином на тренувальну (розміром 480 тис. рядків) та тестову (розміром 60 тис. рядків) вибірки. Розподіл на контрольну та цільову групи складає 50% та 20% користувачів виконали цільову дію. При цьому в цільовій групі 22.9% користувачів виконали цільову дію, а в контрольній 17.94%.

Для аналізу даних використано коефіцієнт рангової кореляції Спірмена, який на відміну від коефіцієнта кореляції Пірсона шукає монотонну залежність між даними, а не тільки лінійну [9]. В тренувальних даних було помічено 16 пар змінних, які мали високу кореляції між собою (більше 0.7). Для кожної пари було визначено змінну, що має меншу за модулем кореляцію з цільовою змінною, та видалено її.

Після внесення аналогічних змін у тестові дані можна переходити до тренування та порівняння методів.

3. РЕЗУЛЬТАТИ ПОРІВНЯЛЬНОГО АНАЛІЗУ

З результатами порівняльного аналізу методів моделювання впливу можна ознайомитись в таблиці 1 та на рисунках 1, 2.

Таблиця 1. Табличні результати порівняльного аналізу

Назва методу	Вплив на топ 30%	Середньозважений вплив	Коефіцієнт Квіні
S-Learner	0,18825	0,04996	0,21522
Z-Learner	0,18951	0,04989	0,20253
DDR(feature='control')	0,20224	0,04926	0,22559
DDR(feature='treatment')	0,20367	0,05023	0,22279

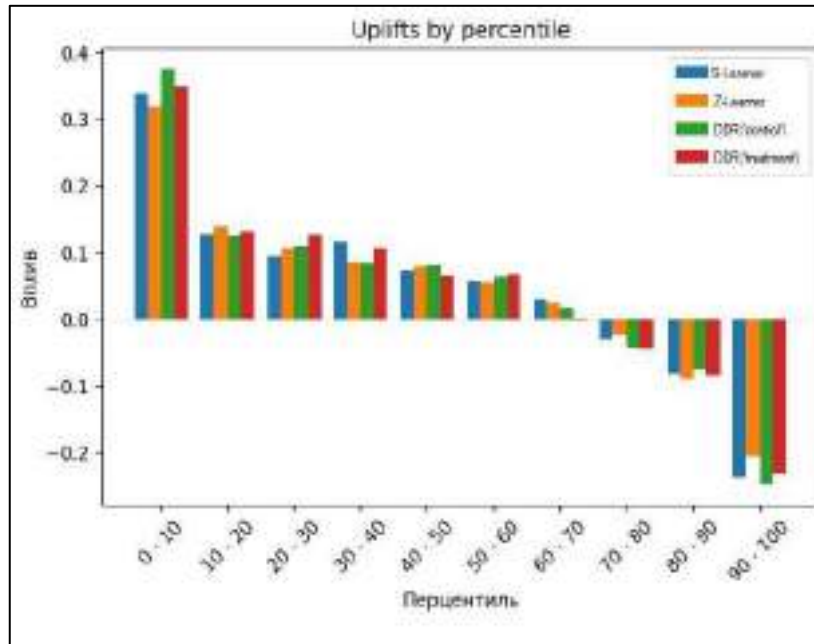


Рисунок 1. Графік впливів побудованих моделей за перцентильними рангами

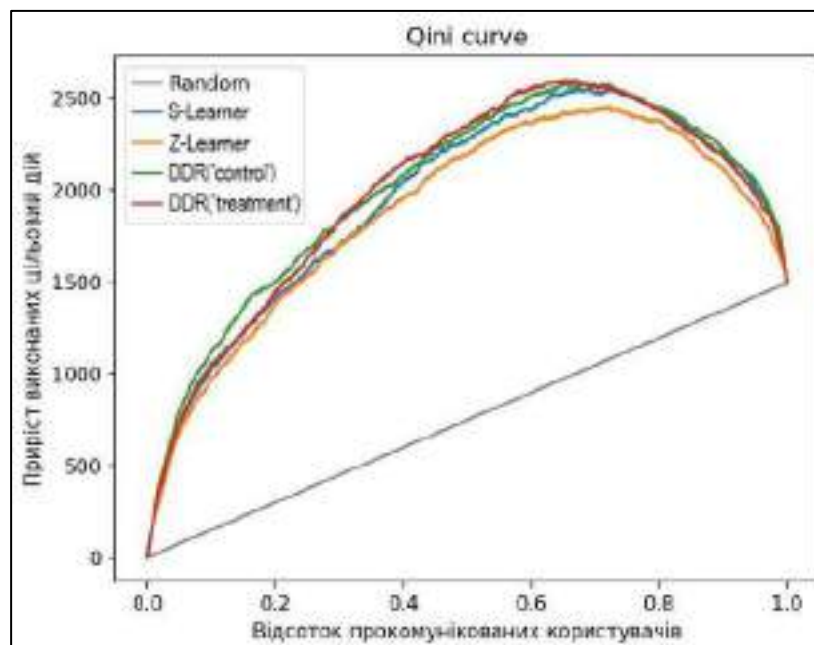


Рисунок 2. Графік кривої Квіні для побудованих моделей

Отримано непогані результати для методу S-Learner, проте при аналізі стовпчикової діаграми впливу за перцентильними рангами можна помітити, що значення впливу 30-го перцентилля нижче 40го. Тобто модель має проблеми з пріоритизацією користувачів за впливом в рамках 30-го та 40-го перцентилів.

Z-Learner трохи краще визначає топ 30% найкращих користувачів для комунікації і немає проблеми з перцентильями, які були у S-Learner. Проте гірші значення середньозваженого впливу та коефіцієнту Квіні свідчать про те, що на всьому обсязі тестових даних Z-Learner гірше справляється з оцінкою впливу комунікації на користувачів.

Метод DDR з використанням результатів моделі контрольної групи в моделі цільової групи має найкращі значення коефіцієнту Квіні та впливу на топ 30% користувачів серед розглянутих методів. Також на стовпчиковій діаграмі можна помітити що впливи 10-го та 30-го перцентилів збільшилися, а 20-го трохи зменшився в порівнянні з минулими методами. Можна прийти до висновку, що загалом модель показала кращі результати, особливо на топ 30% користувачів.

Метод DDR з використанням результатів моделі цільової групи в моделі контрольної групи отримав найкращі значення впливу на топ 30% користувачів та середньозваженого впливу, а також друге найвище значення коефіцієнту Квіні. Це робить його найкращим методом серед усіх розглянутих для поточного набору даних.

Порівняльний аналіз показав, що обидва варіанти обраного методу показали кращі результати за інші розглянуті методи. Опираючись на результати, можна прийти до висновку, що якщо бюджет рекламної кампанії розрахований до 30% користувачів, то для розглянутого набору даних краще використовувати метод DDR з використанням результатів моделі контрольної групи в моделі цільової групи. У всіх інших випадках, краще використовувати метод DDR з використанням результатів моделі цільової групи в моделі контрольної групи для розглянутого набору даних.

4. ОГЛЯД РЕАЛІЗОВАНОЇ СППР

При створенні до СППР було висунуто наступні умови: інтерактивність - система повинна мати інтуїтивно зрозумілий графічний інтерфейс користувача (GUI), який дозволяє користувачам взаємодіяти з системою за допомогою кнопок, меню, графічних елементів тощо; доступність – система повинна бути легкодоступною для користувачів, чим менше користувачу потрібно зробити для початку користування продуктом, тим більше ймовірності привернути його увагу; повнота – в системі має бути можливість провести повноцінний аналіз впливу: починаючи від завантаження даних для тренування моделі, закінчуючи використанням нових даних на створеній моделі та вивантаженням отриманих результатів.

Блок-схема алгоритму роботи з СППР зображена на рисунку 3.

Реалізацію вищезгаданих етапів роботи з СППР зображено на рисунках 4, 5, 6 та 7. На рисунку 4 зображено завантаження тренувальної та тестової вибірок. На рисунку 5 зображено таблиці для аналізу кореляцій Спірмена на тренувальних даних. Рисунок 6 відображає інтерфейс для тренування та порівняння моделей впливу на тренувальних та тестових даних. На рисунку 7 зображено використання обраної натренованої моделі на нових даних з можливістю вивантаження отриманих результатів для подальшого планування рекламної кампанії на їх основі. За наведеними рисунками можна побачити, що інтерфейс є інтуїтивно зрозумілим і СППР дозволяє провести повноцінний аналіз впливу, тобто вона відповідає висунутим до неї вимогам.

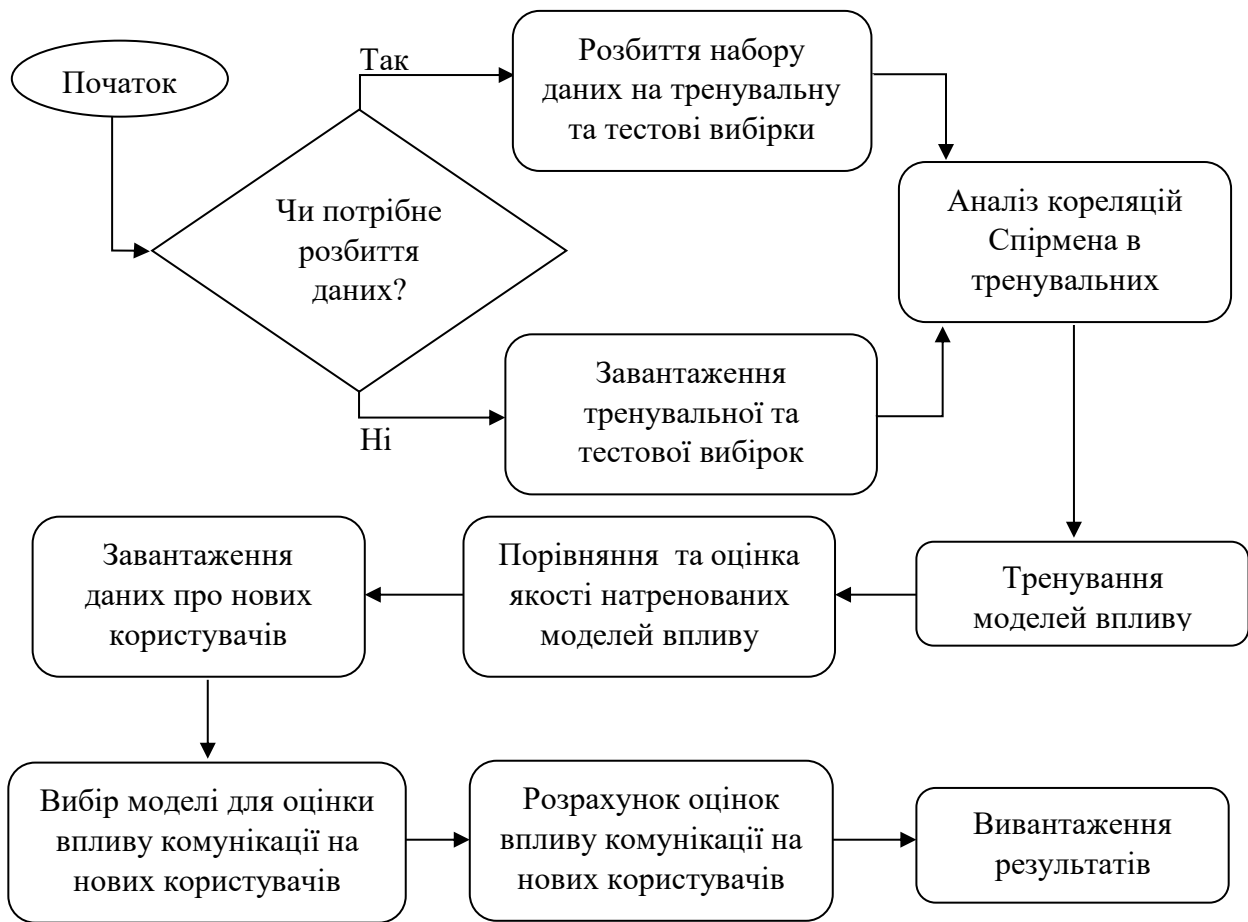


Рисунок 3. Блок-схема алгоритму роботи з СППР

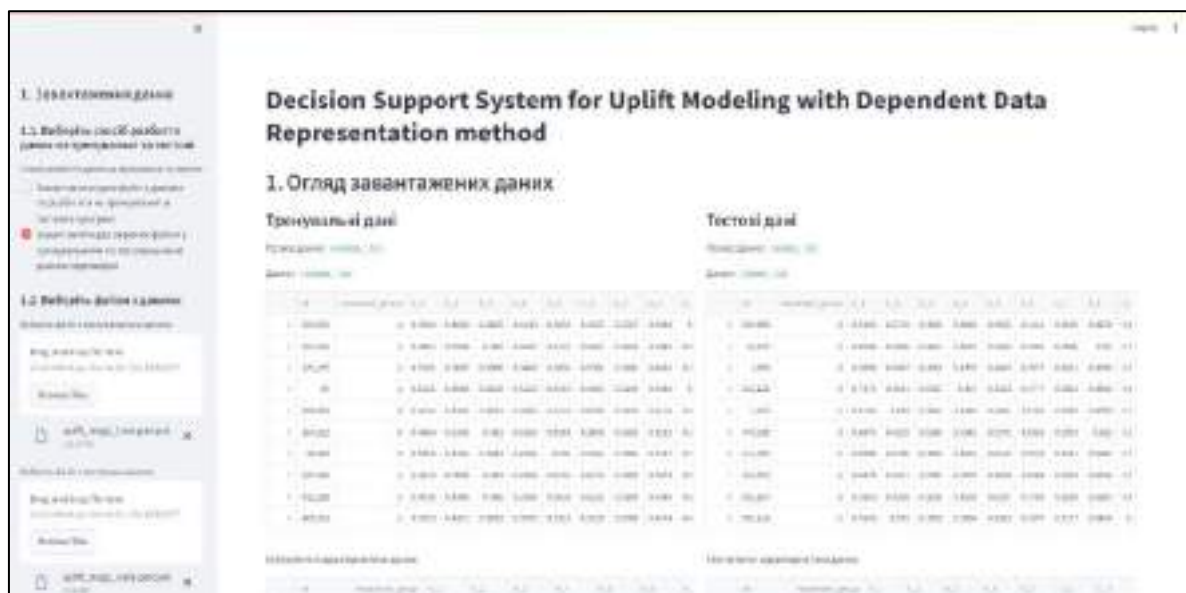


Рисунок 4. Завантаження тренувальної та тестових вибірок в СППР

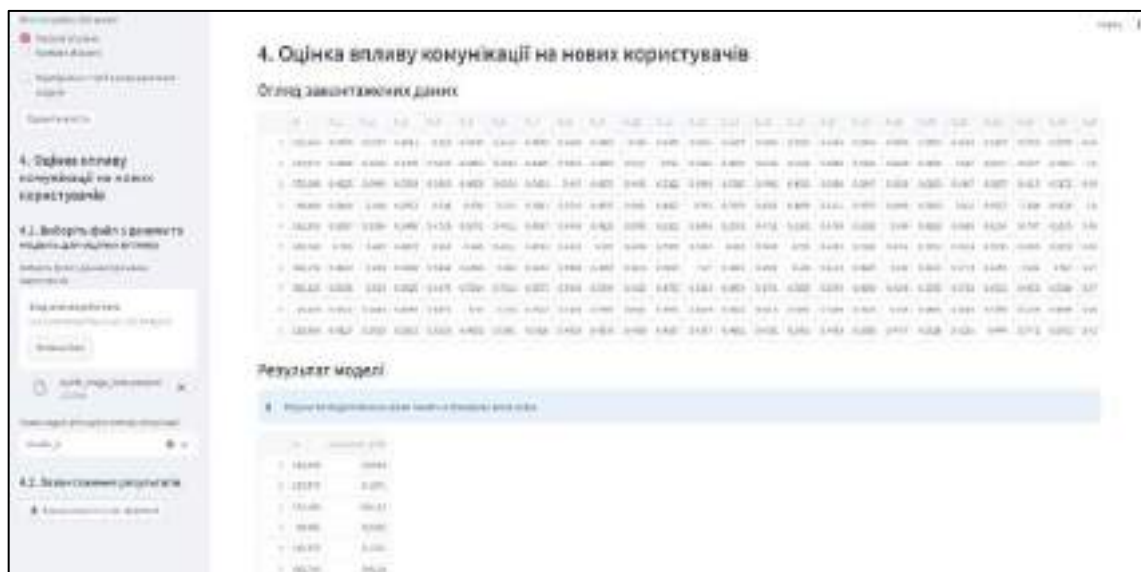


Рисунок 7. Використання побудованої моделі на новому наборі даних та можливість вивантаження отриманих результатів

5. ВИСНОВКИ

В рамках роботи було проведено порівняльний аналіз методу моделювання впливу з залежним представленням даних з іншими та розроблено СППР на його основі.

Порівняльний аналіз показав, що для розглянутого набору даних обидва варіанти методу з залежним представленням даних виявились найкращими, в залежності від потреб підприємства.

Під час розробки СППР було висунуто вимоги до неї та обрано інструменти для створення. Приклад роботи з СППР та огляд інтерфейсу показують, що СППР відповідає висунутим до неї вимогам.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Jaskowski, Maciej; Jaroszewicz, Szymon. Uplift modeling for clinical trial data. In: ICML Workshop on Clinical Data Analysis. 2012. p. 79-95. URL:https://people.cs.pitt.edu/~milos/icml_clinicaldata_2012/Papers/Oral_Jaroszewicz_ICML_Clinical_2012.pdf

2. Бідюк П.І. Проектування комп'ютерних інформаційних систем підтримки прийняття рішень: навч. посіб. / Бідюк П.І., Коршевнюк Л.О. – К.: ННК "ПСА" НТУУ "КПІ", 2010. – 340 с.

3. Betlei, Artem; Diemert, Eustache; Amini, Massih-Reza. Uplift prediction with dependent feature representation in imbalanced treatment and control conditions. In: Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13–16, 2018, Proceedings, Part V 25. Springer International Publishing, 2018. p. 47-57. URL:https://bitlater.github.io/files/iconip_paper.pdf

4. Lo, Victor SY. The true lift model: a novel data mining approach to response modeling in database marketing. ACM SIGKDD Explorations Newsletter, 2002, 4.2: 78-86. URL:https://www.kdd.org/exploration_files/lo.pdf

5. Bon, Michaël; Feutry, Clément; Meftah, Sara. An in-depth benchmark study of the CATE estimation problem: experimental framework, metrics and models Version. URL:https://www.sjscience.org/files/papers/809/CoScience_809.pdf

6. Radcliffe, Nicholas. Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal*, 2007, 14-21.
7. Radcliffe, Nicholas J.; Surry, Patrick D. Real-world uplift modelling with significance-based uplift trees. White Paper TR-2011-1, Stochastic Solutions, 2011, 1-33.
8. Відкритий датасет Megafon: веб-сайт. URL: https://www.uplift-modeling.com/en/latest/api/datasets/fetch_megafon.html
9. De Winter, Joost CF; Gosling, Samuel D.; Potter, Jeff. Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological methods*, 2016, 21.3: 273. URL:https://www.researchgate.net/profile/Joost-De-Winter/publication/307902372_Comparing_the_pearson_and_spearman_correlation_coefficients_a_cross_distributions_and_sample_sizes_A_tutorial_using_simulations_and_empirical_data/links/63b-aa5d5c3c99660ebdc3f60/Comparing-the-pearson-and-spearman-correlation-coefficients-across-distributions-and-sample-sizes-A-tutorial-using-simulations-and-empirical-data.pdf

МОДЕЛЬ УПРАВЛІННЯ РЕСУРСАМИ ГЕТЕРОГЕННИХ БАЗ ДАНИХ В ХМАРНОМУ СЕРЕДОВИЩІ

Зайцев О.В.¹, Мухін В.Є.

Національний технічний університет України «Київський політехнічний інститут
ім. Ігоря Сікорського»

¹ zaytsev961st@gmail.com

Гетерогенні бази даних представляють складну систему з великою кількістю параметрів. Для ефективного управління процесами обробки інформації важливо оцінювати та передбачати витрати ресурсів на ці процеси, що вимагає урахування багатьох зазначених параметрів. Була розроблена модифікація моделі для оцінювання витрат часу запиту у розподілених гетерогенних базах даних, що дозволяє враховувати поточну кількість запитів у системі, а також метод пошуку оптимальної конфігурації системи для мінімізації цих витрат.

Ключові слова: гетерогенні бази даних, оцінювання часу запиту, конфігурація вузлів, хмарне середовище.

1. ВСТУП

Обсяг інформації, що вимагає постійного та швидкого доступу зростає. Це призвело до використання розподілених баз даних для забезпечення ефективного доступу. Однак для ефективної роботи таких баз даних необхідно постійно оцінювати використані ресурси, зокрема час, необхідний для обробки запитів і даних [1]. Таким чином, виникає потреба у розробці різних моделей для оцінки витрат цих ресурсів, зокрема часу, який необхідний для обробки запитів та даних. Прогнозування часу обробки запитів є ключовою проблемою, оскільки від надійних прогнозів залежить ефективне планування робочих процесів та розподіл ресурсів системи [1].

У даній роботі розглядається модель, яка чисельно оцінює витрати часу на обробку запитів в розподілених гетерогенних базах даних у хмарних середовищах, забезпечуючи ефективне управління цими системами. Для покращення результатів дослідження розглядається модифікована модель, що враховує додаткові параметри системи – навантаження системи та ефективність кожного вузла розподіленої бази даних.

2. ОСНОВНИЙ ЗМІСТ

Архітектура системи керування гетерогенними розподіленими базами даних містить 2 основні компоненти, які є необхідними для роботи системи.

Модуль керування відповідає за надання доступу до системи запитами верифікує запити, створює план виконання, надсилає запити до баз даних та обробляє відповіді.

Модуль сховища містить бази даних, файлові системи та інші джерела.

Користувач створює запити, які авторизуються модулем керування, який далі розподіляє запити до джерел даних, обробляє відповіді, та вертає результат користувачу.

Пропонується підключати інші модулі у якості сторонніх сервісів з різних областей: федеративні системи, які дозволяють користувачам подавати запити через єдиний інтерфейс, не потребуючи детального розуміння реалізації та розташування компонентів; централізований інтерфейс управління для моніторингу, управління та оптимізації ресурсів

гетерогенних баз даних; оптимізатор запитів, що аналізує запити до баз даних та визначає оптимальний спосіб їх виконання, враховуючи розташування даних і поточне навантаження на систему; інструменти моніторингу і аналітики для відстеження стану ресурсів, продуктивності системи та виявлення можливих проблем; а також механізми автоматичного масштабування, які автоматично регулюють ресурси залежно від поточного навантаження.

У [2, 3] розглядається система обробки запитів в розподіленій системі з використанням гетерогенних розподілених баз даних, що задається такою функцією:

$$QPS(t, n) = f(SN, SC, SP, SPDB, SPCN), \text{ де}$$

SN – множина вузлів розподіленої системи обробки даних;

SC – множина зв'язків між вузлами розподіленої системи обробки даних;

SP – множина параметрів розподіленої системи обробки даних;

SPDB – множина параметрів бази даних;

SPCN – множина параметрів вузлів керування розподіленою системою обробки даних;

t – одиниця часу;

n – кількість вузлів системи в даний момент часу.

Множина вузлів розподіленої системи обробки даних визначається як функція кількох параметрів [2, 3]:

$$SN(t, n) = fN(SCN, SDBN, SEIN, SRN), \text{ де}$$

SCN – множина вузлів, в яких встановлено програмне забезпечення для керування розподіленою системою обробки даних;

SDBN – множина вузлів, в яких встановлено програмне забезпечення для гетерогенних розподілених баз даних;

SEIN – множина проміжних вузлів, які виконують функцію маршрутизації пакетів;

SRN – множина вузлів, з яких надходять запити до розподіленої системи обробки даних.

Модель, в якій множина параметрів розподіленої системи обробки даних описана функцією [2, 3]:

$$SP(t, n) = fp(P, V, R, DB, T_{db}, T_c), \text{ де}$$

P – продуктивність вузла обробки даних в розподіленій системі обробки даних;

V – швидкість передачі даних по каналу зв'язку в розподіленій системі обробки даних;

R – надійність вузла розподіленої системи обробки даних;

DB – індикатор присутності в вузлі гетерогенної розподіленої бази даних; T_{db} — час обробки пакету – запиту в вузлі розподіленій системі обробки даних в якому знаходиться система управління базою даних;

T_c – час обробки запиту вузлом керування розподіленої системи обробки даних.

Запропоновано модифікацію моделі:

Вводимо додатковий параметр L – поточну кількість запитів користувача в системі за одиницю часу.

$$SP(t, n) = fp(P, V, R, DB, T_{db}, T_c, L)$$

Множина доступних машин, що можуть бути використані при побудові у системі складається з:

$$U_{db} = (T_{req_cn}, T_{req_dbn}, R_{max}, C),$$

де T_{req_cn} – час обробки запиту машиною, якщо виконує роль вузла керування.

T_{req_dbn} – час обробки запиту машиною, якщо виконує роль вузла сховища даних

R_{max} – пропускна здатність вузла (кількість запитів, що вузол може обробити за одиницю часу)

C – ціна оренди вузла в хмарного провайдера.

Якщо вузол є перевантаженим, тобто $L_{node} > L_{jmax}$, то запит чекає на обробку в черзі. Вважаємо що запити розподіляються між вузлами одного типу циклічно.

Час обробки запиту користувача у системі:

$$T(L) = T_c(L) + T_{db}(L) + T_{transport}, \text{ де}$$

$$T_c(L_{node}, L_{max}, T_{req}) = T_{c_i} * \max\{1, L_{node} - L_{imax}\}$$

$$T_{db}(L_{node}, L_{max}, T_{req}) = T_{db_j} * \max\{1, L_{node} - L_{jmax}\}$$

$$i = \min\left(k \mid \sum_{l=1}^k R_l \geq (i \bmod L_{total})\right), R_{total} = \sum_c n_i L_i$$

$$j = \min\left(k \mid \sum_{l=1}^k R_l \geq (j \bmod L_{total})\right), R_{total} = \sum_{db} m_j L_j$$

$$T_{transport} = \frac{S * N}{V},$$

де S – середній розмір пакету, V – швидкість передачі даних в мережі, N – кількість пересилань даних.

Ціна оренди за період $C_{total} = \sum_c n_i C_i + \sum_{db} m_j C_j$

Задача полягає у знаходженні кількості вузлів кожного типу для мінімізації середніх витрат часу на запит T_r , за обмежень бюджету $C_{total} \leq C_{max}$ та заданої навантаженості системи L запитами користувачів.

3. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Було проведено низку експериментів. Доступні для вибору вузли наведено в таблиці 1.

Таблиця 1. Вузли для формування конфігурацій

Доступні машини	Час обробки запиту керуючим вузлом, с	Час обробки запиту вузлом бази даних, с	Максимальна кількість зв'язків	Вартість, \$
Тип 1	0,005	0,02	50	200
Тип 2	0,003	0,015	40	250
Тип 3	0,009	0,025	70	180

Системні параметри:

- Кількість проміжних вузлів системи: 10;
- Середня швидкість мережі, що з'єднує вузли: 10 мб/с;
- Середній розмір пакету даних: 500 КВ;

Знайдені конфігурації та часові витрати на запит можна побачити у таблиці 2.

Таблиця 2. Результати експериментів

Кількість запитів користувачів	Бюджет, \$	Знайдена конфігурація системи – кількість машин типу 1, 2, 3				Очікуваний час виконання, мс	Грошові витрати, \$
		CN	5	1	2		
2000	1000	CN	5	1	2	32,87	860
		DBN	4	1	2		
10000	1000	CN	1	3	4	35,51	990
		DBN	0	3	5		
10000	1500	CN	1	2	3	35,39	1100
		DBN	2	2	5		
500	1500	CN	1	1	0	32,12	180
		DBN	1	1	0		

Бачимо, що при збільшенні навантаження на систему з 2000 до 10000 запитів у вузлах утворюються черги, витрати часу на запит зростають, також підійшли до межі бюджету та більше не можемо додавати нові вузли. Конфігурація змінилась, за попередньої конфігурації середній час на запит склав 37,11 мс, отже зміна конфігурації пішла на користь.

За рахунок збільшення бюджету до 1500\$ змогли додати нові вузли, за рахунок цього витрати часу трохи зменшились.

Після зменшення навантаження на систему, нехай в системі 500 користувачів надіслали запити, мережа сильно спростилась, оскільки навантаження впало, більше нема необхідності в великій кількості серверів.

4. ВИСНОВКИ

Гетерогенні розподілені бази даних в хмарному середовищі забезпечують ряд переваг, що роблять їх привабливими для великої кількості застосувань. По-перше, такий підхід дозволяє об'єднувати різноманітні джерела даних, що може бути надзвичайно корисним для підтримки різноманітних бізнес-процесів та додатків. Він сприяє ефективному використанню різноманітних технологій та систем у хмарному середовищі, таких як обчислення, зберігання та обробка даних. Крім того, гетерогенність дозволяє оптимізувати розподілені ресурси для різних видів завдань та додатків, забезпечуючи гнучкість та масштабованість.

Було запропоновано архітектуру системи управління ресурсами гетерогенних баз даних в хмарному середовищі, а також модифікацію моделі для оцінювання витрат часу на запит для цієї системи та пошуку оптимальної конфігурації системи для оптимізації цих витрат, проведено експерименти та знайдено оптимальну конфігурацію за заданих параметрів.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Özsu, M. T. and Valduriez, P. (2020). Principles of distributed database systems. Springer, 4th edition. DOI: 10.1007/978-3-030-26253-2.

2. Корнага Я. І. Моделі та методи організації та управління гетерогенними розподіленими базами даних з динамічною структурою на основі мережецентричного підходу : дис. докт. техн. наук : 05.13.06 / Корнага Ярослав Ігорович – Київ, 2020. – 328 с. 11.

3. V. Mukhin, Y. Kornaga, V. Bondarenko, V. Zavgorodnii, O. Herasymenko and O. Sholokhov, "Mathematical Model for Heterogeneous Databases Parameters Estimation in Distributed Systems with Dynamic Structure," 2020 IEEE 2nd International Conference on Advanced Trends in Information Theory (ATIT), 2020, pp. 158-161, doi: 10.1109/ATIT50783.2020.9349331.

СИСТЕМА ПРОГНОЗУВАННЯ МЕТЕОРОЛОГІЧНИХ УМОВ НА ОСНОВІ МЕТОДІВ АНАЛІЗУ ДАНИХ ТА ШТУЧНОГО ІНТЕЛЕКТУ

Іванійчук А.П.¹, Гуськова В.Г.

Національний технічний університет України «Київський політехнічний інститут
ім. Ігоря Сікорського»

¹ arsen.ivaniichuk@gmail.com

Штучний інтелект відкриває перед нами широкі можливості у покращенні прогнозування погодних умов, роблячи цей процес більш точним і надійним. Метою роботи є аналіз існуючих підходів до прогнозування метеорологічних умов за допомогою методів аналізу даних та штучного інтелекту. У результаті дослідження була розроблена система прогнозування метеорологічних умов, яка ґрунтується на використанні моделі штучного інтелекту і може бути використана для забезпечення точних та актуальних даних про погоду, що важливо для різних галузей, включаючи сільське господарство, транспорт та безпеку громадян. У роботі було використано теоретичні та емпіричні методи дослідження.

Ключові слова: прогнозування погоди, аналіз метеорологічних даних, рекурентні нейронні мережі, LSTM.

1. ВСТУП

З використанням штучного інтелекту для прогнозування метеорологічних умов відкриваються перспективи покращення існуючих підходів до розв'язання даної проблеми. Ця нова парадигма дозволяє підвищити точність і надійність погодних прогнозів, що має вагомий науковий та практичний цінність у численних галузях, включаючи аграрну сферу, транспорт та інші.

Використання штучного інтелекту дозволяє здійснювати обробку великого обсягу метеорологічних даних та проводити складний аналіз, урахувавши різноманітні фактори, що впливають на погодні явища. Цей підхід дозволяє отримувати результати, які є точнішими та передбачуванішими.

Для досягнення максимальної ефективності використання можливостей штучного інтелекту в погодному прогнозуванні, є доцільним створення системи, заснованої на ефективній моделі. У рамках наших досліджень було розроблено систему прогнозування метеорологічних умов з використанням моделі Long Short-Term Memory або LSTM, що дозволяє враховувати динаміку погодних явищ з високою точністю та прогнозувати їх на майбутнє з великою достовірністю.

Крім того, наша система використовує технологію REST API, що забезпечує її зручність у використанні та інтеграції з іншими додатками і системами.

2. ОГЛЯД ІСНУЮЧИХ МЕТОДІВ ПРОГНОЗУВАННЯ МЕТЕОРОЛОГІЧНИХ ПОКАЗНИКІВ

Прогнозування погоди – це завдання, яке вимагає великого обсягу даних та аналізу для отримання точних результатів. У цьому контексті існують різні методи аналізу, які допомагають прогнозувати метеорологічні умови.

Один із ключових підходів до прогнозування погоди – це використання математичних моделей, які моделюють атмосферні процеси. Ці моделі базуються на рівняннях, що описують фізичні закони руху повітря, теплообміну та інші атмосферні явища. Вони розділяють атмосферу на велику кількість областей та обчислюють зміни параметрів атмосфери в кожній з них з плином часу. Моделі дозволяють прогнозувати погоду на основі початкових умов, які визначаються спостереженнями та даними з метеорологічних станцій та супутників.

Інший підхід до прогнозування погоди полягає в аналізі статистичних зв'язків між різними погодними явищами на основі історичних даних. Наприклад, метод регресії може бути використаний для встановлення зв'язків між температурою, атмосферним тиском та іншими параметрами. За допомогою цих зв'язків можна прогнозувати майбутні значення погодних параметрів.

Сучасні технології штучного інтелекту, зокрема нейронні мережі, відіграють важливу роль у прогнозуванні погоди. Рекурентні нейронні мережі, такі як LSTM, можуть аналізувати складні закономірності в часових рядах метеорологічних даних. Вони здатні враховувати динаміку погодних явищ і прогнозувати їх на короткий та середній терміни. Цей підхід дозволяє отримувати точні та передбачувані результати в погодному прогнозуванні.

3. АРХІТЕКТУРА МОДЕЛІ LSTM

Архітектура моделі LSTM є різновидом рекурентних нейронних мереж (RNN) і була запропонована з метою вирішення проблем з затуханням та вибуховим зростанням градієнтів у RNN.

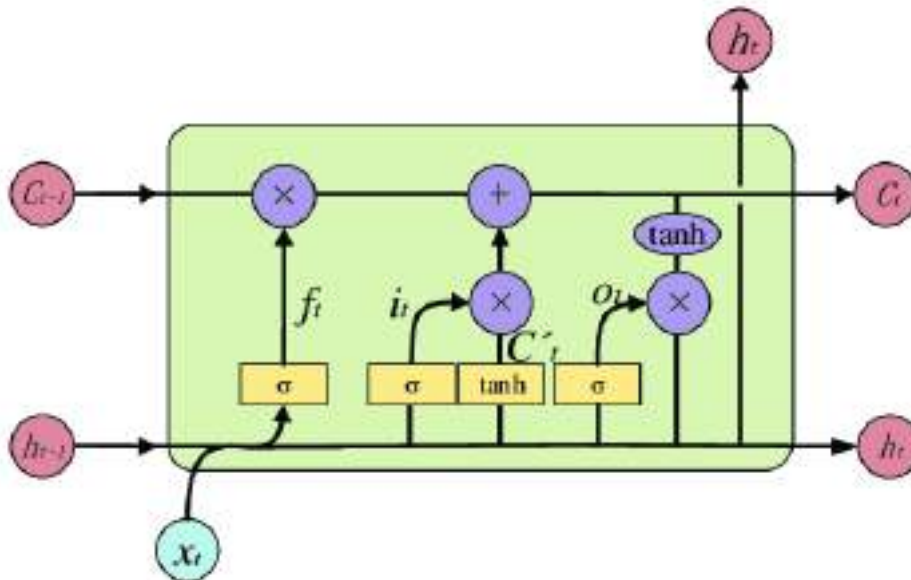


Рисунок 1. Архітектура блоку моделі LSTM

Відмінною особливістю LSTM є використання так званих "комірок пам'яті", які контролюють інформаційний потік за допомогою трьох воріт: вхідного, вихідного та забування. Розглянемо детальніше кожен з воріт:

- Вхідні ворота визначають, яка частина нової інформації буде збережена в комірці пам'яті.
- Ворота забування вирішують, яка частина існуючої інформації у комірці буде втрачена.
- Вихідні ворота визначають, яка частина інформації в комірці пам'яті буде передана до наступного рівня.

Кожні ворота складаються з сігмоїдальної активаційної функції, яка виводить значення між 0 та 1, та покомпонентного множення, що дозволяє контролювати потік інформації.

Ці механізми дозволяють LSTM зберігати, модифікувати або втрачати інформацію з комірки пам'яті в залежності від завдань, і це робить їх ефективними при роботі з послідовностями даних довгого терміну. Тому LSTM широко використовуються в задачах обробки природної мови, прогнозування часових рядів та інших завданнях, де потрібно моделювати завдання з великими відстанями між пов'язаними подіями.

4. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

У результаті проведеної роботи було розроблено систему прогнозування погодних умов, що має архітектуру, зображену на рисунку 2.

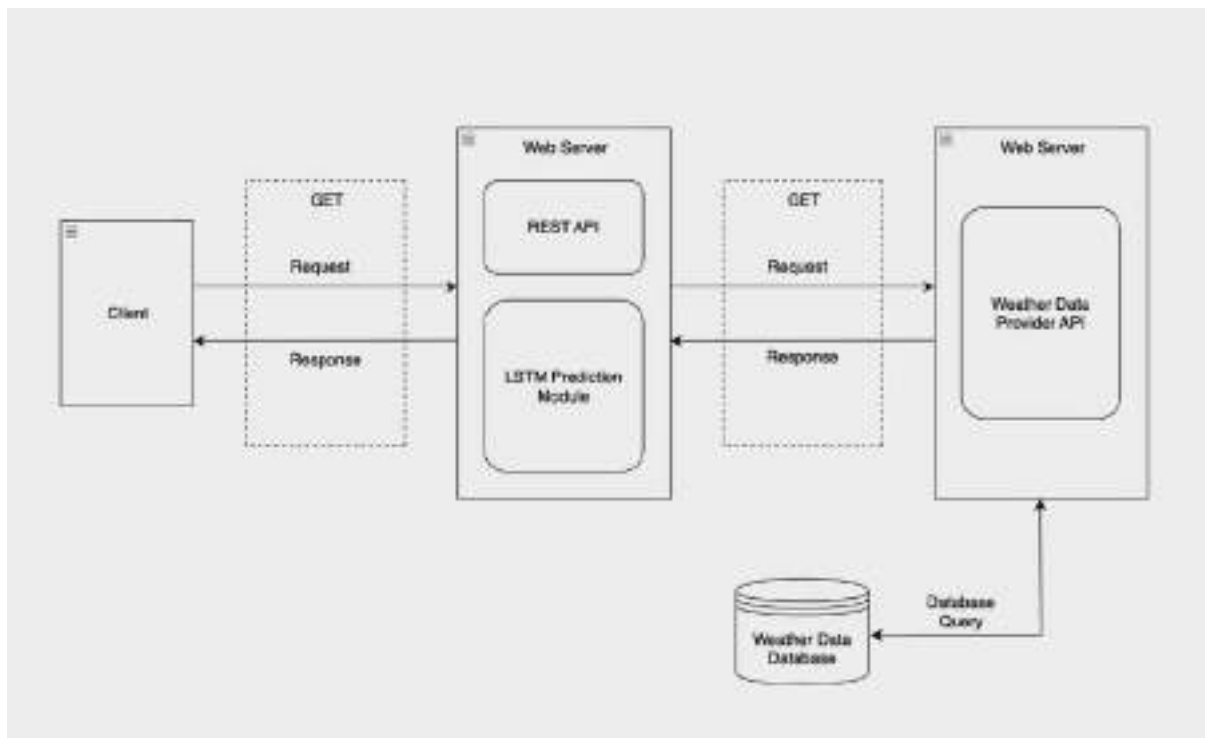


Рисунок 2. Архітектура системи

Запропонована архітектура системи забезпечує ефективний обмін даними між клієнтом, сервером прогнозування та постачальником історичних метеорологічних даних. Використання LSTM моделі в комбінації із технологією REST API створює гнучку, масштабовану та високоефективну систему прогнозування погоди.

Було проведено тренування та тестування моделі для отримання 48-годинного прогнозу на основі даних про 18 попередніх діб.

Для оцінки якості використовуються метрики Mean Absolute Error (MAE), Mean Squared Error (MSE), і Mean Absolute Percentage Error (MAPE) - це метрики, які використовуються для оцінки точності прогнозування або моделювання в наукових дослідженнях і аналізі даних.

Ми розглядаємо датасет, що представляє докладну інформацію про погодні умови в Києві, столиці України, з 1 січня 1881 року. Незважаючи на різноманіття джерел метеорологічної інформації, що доступні сьогодні, такий обширний історичний масив інформації може надавати в основному лише гідрометеорологічна служба. Це зумовлено історичними, організаційними та науковими особливостями збору та зберігання даних.

Нижче наведено графіки передбачень моделі для показника середньої температури на тренувальній та тестовій вибірках, а також значення метрик для кожного випадку.

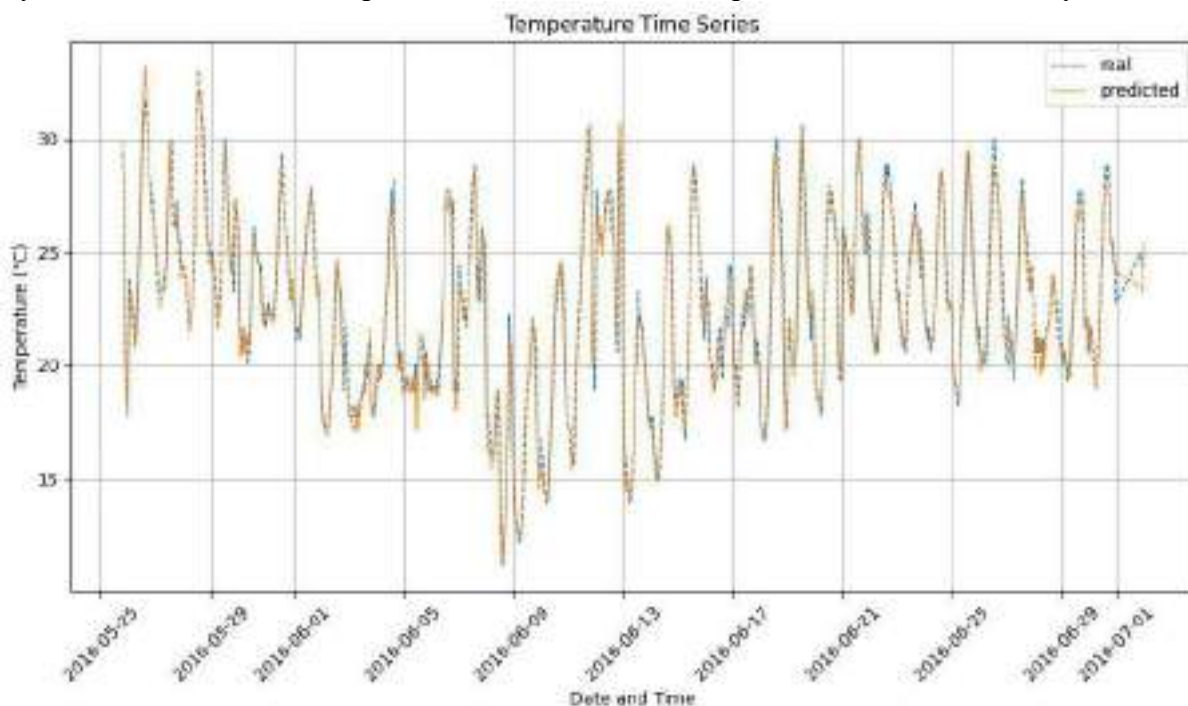


Рисунок 3. Порівняння графіків середньої температури для тренувальної вибірки

На рисунку 3 зображено порівняння графіків середньої температури для тренувальної вибірки (80% дата сету).

Показники метрик:

- MAE (Середня абсолютна похибка): 0,53;
- MSE (Середня квадратична похибка): 0,60;
- MAPE (Середня абсолютна похибка у відсотках): 2,9%.

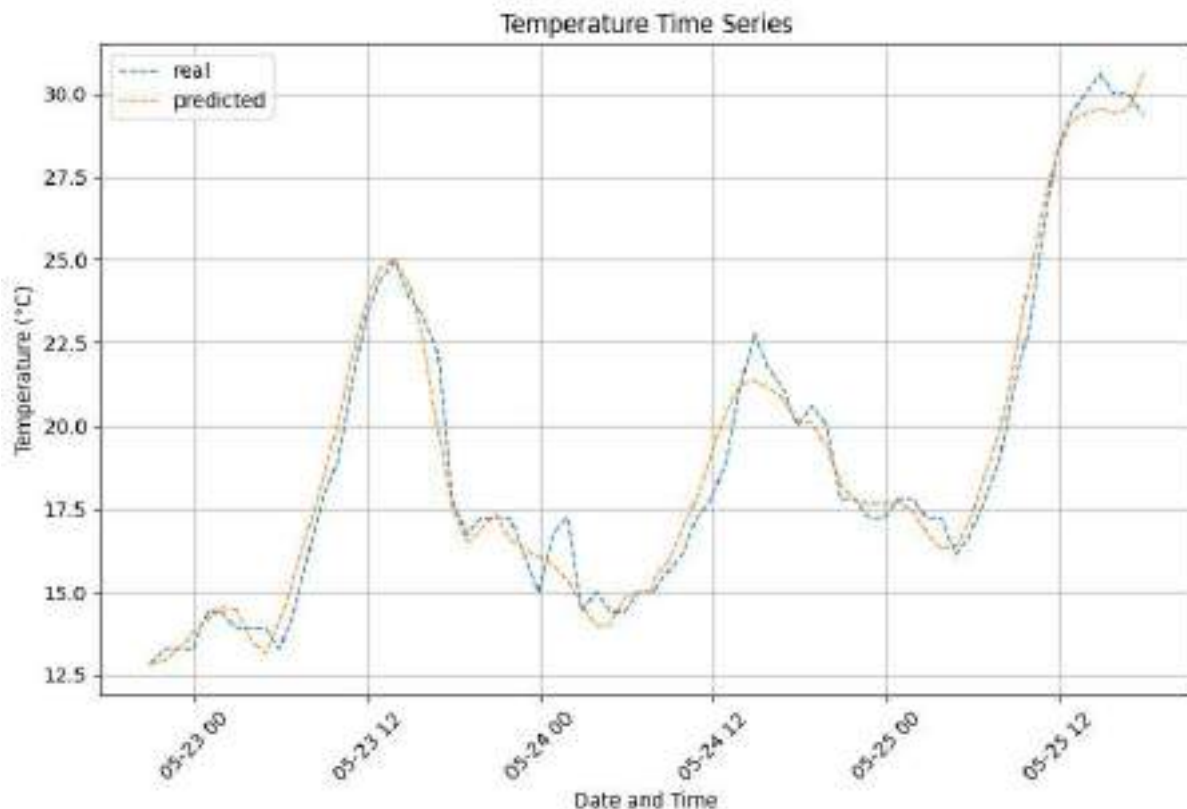


Рисунок 4. Порівняння графіків середньої температури для тестової вибірки

На рисунку 4 зображено порівняння графіків середньої температури для тестової вибірки (20% дата сету).

Показники метрик:

- MAE (Середня абсолютна похибка): 0,59
- MSE (Середня квадратична похибка): 0,82
- MAPE (Середня абсолютна похибка у відсотках): 3,5 %

5. ВИСНОВКИ

Розглядаючи реалізацію системи прогнозування метеорологічних умов на основі REST API та LSTM, можна відзначити ключову роль комбінації цих технологій. Інтеграція REST API забезпечує гнучкість та масштабованість системи, дозволяючи ефективно обмінюватися даними між її компонентами. Ця гнучкість стає особливо актуальною, оскільки система автоматично звертається до постачальника метеорологічної інформації через його власний API для отримання актуальних даних.

LSTM, в свою чергу, демонструє високу ефективність при аналізі часових рядів, зокрема в метеорологічному прогнозуванні. Ця модель враховує довгострокові залежності в даних, що підтверджується точністю прогнозів системи щодо погодних умов.

Архітектура системи розроблена таким чином, що вона легко адаптується до різних джерел даних, дозволяючи масштабувати систему. Крім того, завдяки її відкритості для інновацій, можливе впровадження нових методів аналізу даних або алгоритмів штучного інтелекту в майбутньому, що робить систему довгостроково стійкою до технологічних змін.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Han, J. M., Ang, Y. Q., Malkawi, A., & Samuelson, H. W. (2021). Using recurrent neural networks for localized weather prediction with combined use of public airport data and on-site measurements. *Building and Environment*, 192, 107601.
2. Sabzipour, B., Arsenault, R., Troin, M., Martel, J.-L., Brissette, F., Brunet, F., & Mai, J. (2023). Comparing a long short-term memory (LSTM) neural network with a physically-based hydrological model for streamflow forecasting over a Canadian catchment. *Journal of Hydrology*, 130380
3. Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (November 15, 1997), 1735–1780.
4. Chen, J., & Huang, C. (2020). A Comprehensive Review of Weather Forecasting Using Machine Learning. *IEEE Access*, 8, 65288-65311.
5. LSTM: Long Short-Term Memory. (n.d.). Retrieved from https://www.tensorflow.org/guide/keras/rnn#lstm_long_short-term_memory
6. REST API Documentation. (n.d.). Retrieved from <https://developer.mozilla.org/en-US/docs/Glossary/REST>
7. Roy, A. (2019). Building a RESTful Web Service. Retrieved from <https://spring.io/guides/gs/rest-service/>
8. OpenWeather. (n.d.). Official Website. Retrieved from <https://openweathermap.org/>
9. Paramasivan, S. K. (2021). Deep learning based recurrent neural networks to enhance the performance of wind energy forecasting: A review. *Revue d'Intelligence Artificielle*, 35(1).

КЛАСТЕРИЗАЦІЯ ЗА ДОПОМОГОЮ OPTICS: АНАЛІЗ ТА ОПТИМІЗАЦІЯ З ВИКОРИСТАННЯМ ГРАФІКІВ ТА МЕТРИК

Іванюта О.О.¹, Недашківська Н.І.²

Національний технічний університет України «Київський політехнічний інститут
імені Ігоря Сікорського», Київ, Україна

¹ ivanyuta.olexandr@lil.kpi.ua, ² nedashkovskaya.nadezhda@lil.kpi.ua [0000-0002-8277-3095]

Розглядається задача кластеризації за допомогою OPTICS. Результати виведені на графіках досяжності. Проведено аналіз метрик якості, таких як Estimated number of clusters, Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI) та Silhouette Coefficient, для кількісної оцінки точності та роздільності кластерів. На основі цих метрик визначено оптимальні моделі для кожного набору даних. Проведено експерименти з різними параметрами для визначення оптимальних значень. Розроблено програмне забезпечення для побудови кластеризації.

Ключові слова: OPTICS, DBSCAN, графік досяжності (reachability plot)

1. ВСТУП

В останні роки активно досліджується область алгоритмів кластеризації на основі щільності, які знаходять широке застосування у визначенні груп схожих об'єктів у великих обсягах даних для розпізнавання патернів та машинного навчання. Алгоритми кластеризації на основі щільності можуть обробляти зашумлені точки і виявляти кластери різної складної форми, але також зазвичай потребують значних часових витрат [1]. Представниками щільнісних алгоритмів є DBSCAN [2] та OPTICS [3]. Важливою властивістю багатьох наборів реальних даних є те, що їх внутрішня кластерна структура не може бути охарактеризована параметрами глобальної щільності. Для виявлення кластерів у різних регіонах простору даних можуть знадобитися різні локальні щільності. Основна ідея щільнісних алгоритмів кластеризації полягає в тому, що для кожної точки в кластері має існувати окіл заданого радіусу ϵ , який містить хоча б мінімальну задану кількість точок MinPts.

Останні публікації у цій області пропонують ряд інноваційних підходів та вдосконалень до алгоритму OPTICS [4–9]. Наприклад, в одному з досліджень вводиться новий метод кластеризації на основі щільності під назвою Fast Principal Component Analysis Pruning (FPCAP) [5]. Цей метод дозволяє ефективно прискорити виконання алгоритмів кластеризації, зокрема, вдосконалює алгоритми DBSCAN та BLOCK-DBSCAN [5].

З іншого боку, деякі дослідження акцентують увагу на важливості візуалізації та інтерактивного аналізу результатів роботи OPTICS. Зокрема, розроблено візуалізаційний інструмент VizOPTICS [6], який глибоко інтегрує людський та машинний інтелект для полегшення розуміння та використання OPTICS у виділенні значущих кластерів.

2. МЕТОДИ ТА МАТЕРІАЛИ

2.1 Основні означення

OPTICS (Ordering Points To Identify the Clustering Structure) – розширення DBSCAN. В OPTICS додатково будується графік досяжності: для кожної точки даних зберігається відстань до ядра та відстань досяжності, а також місце точки у відсортованій множині точок даних, необхідні для визначення належності до кластерів.

OPTICS базується на наступних означеннях [3].

ϵ -околом точки $p \in D$ називається множина $D_\epsilon(p) = \{q \in D \mid \text{distance}(p, q) \leq \epsilon\}$.

Параметр ϵ , строго кажучи, не є обов'язковим. Можна задати діапазон зміни значень ϵ , встановлюючи максимально можливе значення \max_eps цього параметру.

Точка $p \in D$ називається *основною точкою (core point) або ядром* якщо її ϵ -окіл містить принаймні MinPts точок, де ϵ і MinPts – задані.

MinPts – мінімальна кількість точок, потрібна для утворення кластера. Обирають $\text{MinPts} \geq m + 1$, де m – розмірність набору даних.

Точка називається *граничною (border point)*, якщо у її ϵ -околі точок менше ніж MinPts , але в ϵ -околі потрапляє основна точка. Точка називається *шумовою (noise)*, якщо у її ϵ -околі точок менше ніж MinPts і жодної основної точки туди не потрапляє.

Точка q *безпосередньо досяжна за щільністю* з точки p (позначимо $p \rightarrow q$) в множині точок D , якщо: 1) $q \in D_\epsilon(p)$ і 2) p – ядро, тобто $|D_\epsilon(p)| \geq \text{MinPts}$.

Умова 2) показує, що лише з основних точок інші точки можуть бути безпосередньо досяжними за щільністю.

Точка q *досяжна за щільністю (density-reachable)* з точки p в множині точок D , якщо існує послідовність точок q_1, \dots, q_n , $q_i \in D$, таких що $q_1 = p$, $q_n = q$ та $q_i \rightarrow q_{i+1}$:

$$p = q_1 \rightarrow q_2 \rightarrow \dots \rightarrow q_n = q$$

Відношення досяжності за щільністю не є симетричним в загальному випадку. Тільки основні точки (ядра) можуть бути взаємно досяжними за щільністю [3].

Точка q *зв'язана за щільністю (density-connected)* з точкою p в множині точок D , якщо існує точка $o \in D$, така що обидві точки p і q досяжні за щільністю з точки o .

Відношення зв'язності за щільністю симетричне.

На основі розглянутих означень, кластер – це множина з максимальної кількості зв'язаних за щільністю точок. Шумовими названо точки, які не містяться в жодному кластері.

Нехай D – множина точок даних, параметри ϵ і MinPts – задані.

Кластером C називається непорожня підмножина множини D , яка задовольняє наступним умовам [3]:

1) *Максимальність*: для кожних $p, q \in D$: якщо $p \in C$ і точка q досяжна за щільністю з p , то $q \in C$.

2) *Зв'язність*: для кожних $p, q \in C$: точка q зв'язана за щільністю з p .

Кластер складається не лише з основних точок (ядер), а також з граничних точок, які безпосередньо досяжні за щільністю з принаймні однієї основної точки цього кластеру.

2.2 OPTICS – покращений метод кластеризації

Алгоритм OPTICS вирішує основну слабкість DBSCAN – проблему виявлення кластерів даних з різною щільністю [9]. Результат кластеризації за алгоритмом DBSCAN дуже чутливий до значень початкових параметрів ϵ та MinPts , і різні значення цих параметрів можуть призводити до різних результатів кластеризації. Зокрема, в алгоритмі DBSCAN може бути важко знайти такі значення ϵ та MinPts , щоб правильно визначити всі кластери в даних.

Алгоритм OPTICS не є чутливим до радіуса ϵ , що робить його більш гнучким у виявленні кластерів. На рисунку 1 видно п'ять кластерів (A, B, C₁, C₂ і C₃). За відповідно підібраними значеннями ϵ_{ps} та MinPts, алгоритм DBSCAN може кластеризувати дані і отримати A, B, C, або кластеризувати так, щоб отримати C₁, C₂, C₃, але важко визначити параметри так, щоб отримати всі п'ять кластерів одночасно. У той час як алгоритм OPTICS, завдяки тому, що він не є чутливим до параметрів, може успішно виявити всі п'ять кластерів (A, B, C₁, C₂, C₃) одночасно.

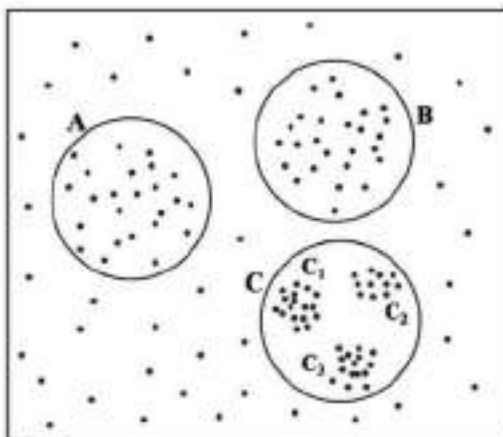


Рисунок 1. Кластери з різними параметрами щільності [3]

3. ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ АЛГОРИТМУ ТА ВИВЕДЕННЯ РЕЗУЛЬТАТІВ

Вхідними даними обрано два набори даних.

Для кожного з цих наборів було створено модель за допомогою алгоритму OPTICS. Параметри моделей налаштовані на оптимальні значення. На рисунку 2 продемонстровані результати роботи алгоритму.

Для першого набору даних, алгоритм визначив 2 набори даних та шум. Для другого набору – 6 кластерів та шум. З них повністю правильно визначено лише 1 кластер.

Кластери на графіках досяжності проявляються як долини на графіку. Глибина долин (тобто відстань досяжності між точками) вказує на щільність кластерів. Глибші долини означають щільніші кластери, і навпаки. На рис. 3.а видно дві глибоких долини A та B. Також на графіку присутній шум. На рисунку 3.б видно шість глибоких долин, які позначені латинськими літерами від A до F. Також на графіку присутній шум.

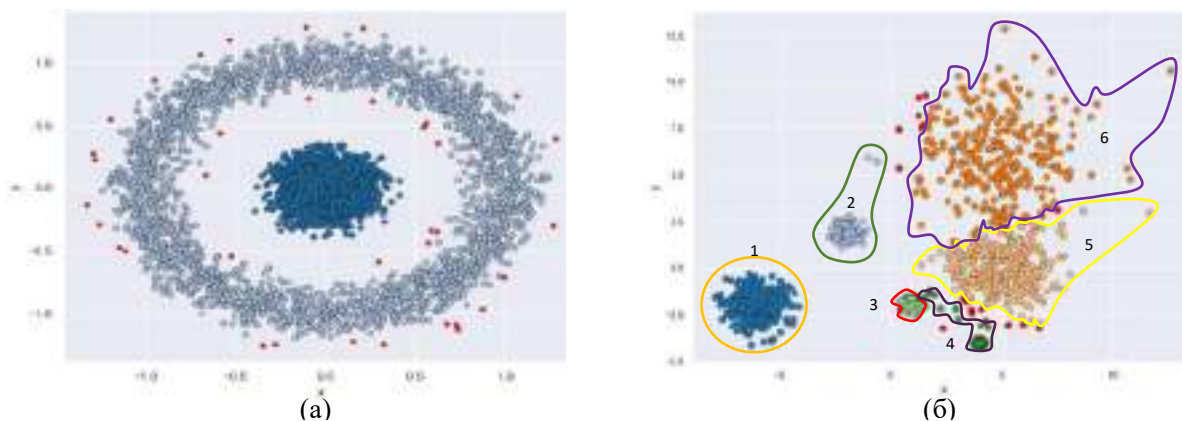


Рисунок 2. Результати кластеризації для першого набору даних (а) та для другого (б)

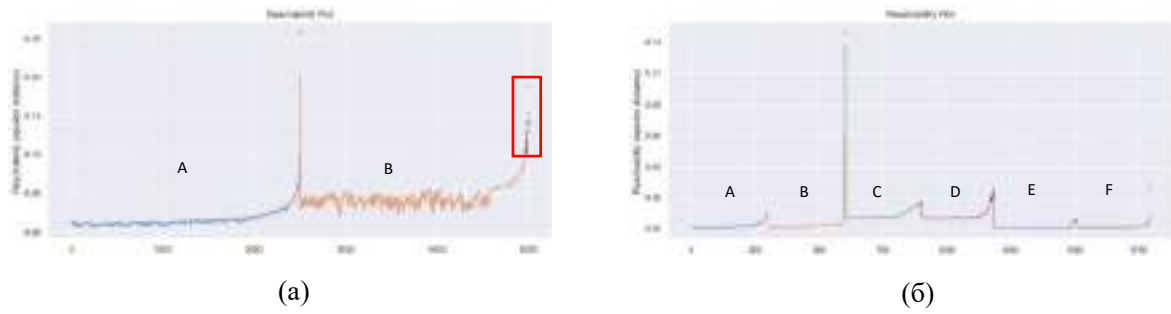


Рисунок 3. Графіки досяжності для першого(а) та другого(б) наборів даних

3.1. Результати перебору параметрів

На рисунку 4 зображені моделі з різними метриками. Чудовий результат показували моделі, які використовували DBSCAN для виділення кластерів.

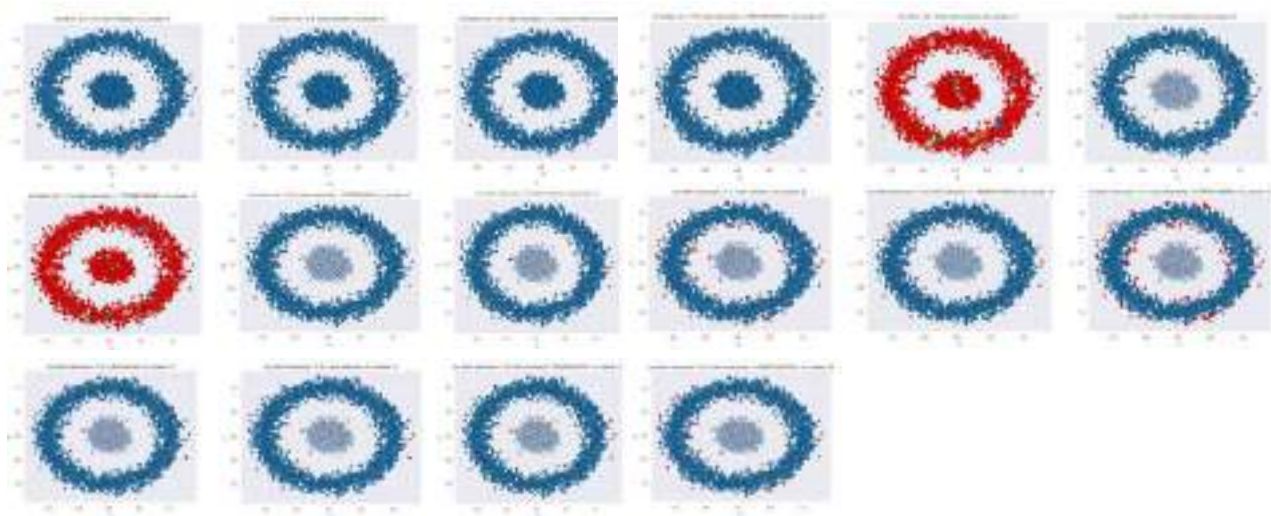


Рисунок 4. Графік альтернативних моделей для першого набору даних

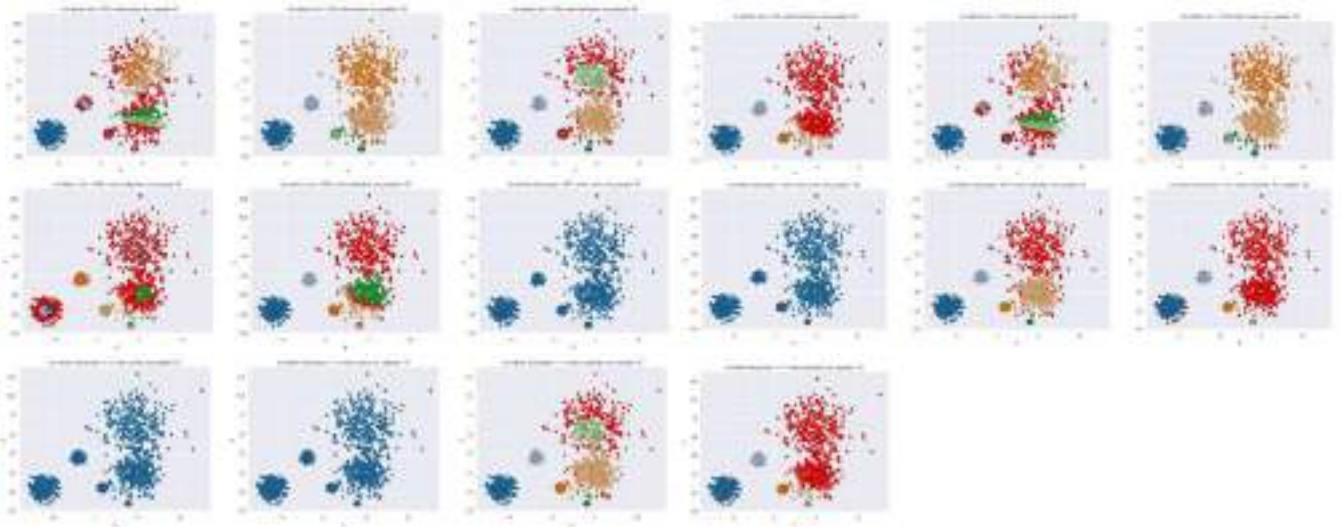


Рисунок 5. Графік альтернативних моделей для другого набору даних

Найгірші результати показали такі моделі, в яких ϵ дорівнює 0.9 (завелике значення), і моделі, в яких не використовувався DBSCAN та мінімальна кількість точок в кластері дорівнює 10.

При малих значеннях $min_samples$, у моделях, що не використовують DBSCAN, виділяють занадто багато кластерів, а в тих, що використовують виділяється максимум 3 кластери. Поєднання метрики cosine та великого значення ϵ призводить до того, що всі данні об'єднуються в один кластер. Інші моделі вийшли доволі хорошими. На цьому наборі даних, особливо при використанні DBSCAN, можна помітити, що алгоритм є дуже чутливим в залежності від поєднання різних параметрів.

3.2. Пошук найкращих параметрів

Було проведено дослідження, в якому було знайдено найкращі моделі за такими метриками:

1. Adjusted Rand Index (ARI) [10] – ARI вимірює подібність між справжніми мітками кластерів та прогнозованими мітками, враховуючи випадкові перестановки. Значення ARI від 0 до 1, де 1 вказує на ідеальне узгодження між справжніми та прогнозованими мітками
2. Adjusted Mutual Information (AMI) [11] – AMI знаходить ступінь взаємодії між справжніми та прогнозованими мітками. Значення AMI також варіюється від 0 до 1, де 1 вказує на повну взаємодію між мітками.
3. Silhouette Score [12] – Коефіцієнт силуету вимірює, наскільки добре відокремлені кластери в порівнянні з іншими кластерами. Значення варіюються від -1 до 1.

Для першого набору даних найкращим значенням мір якості відповідає одне й те ж розбиття. З рисунка 6 можна зробити висновок, що найкращою є модель з такими параметрами: $\epsilon = 0.11$, $metric:chebyshev$, $min_samples: 30$. Дана модель правильно визначила всі кластери для даного набору.

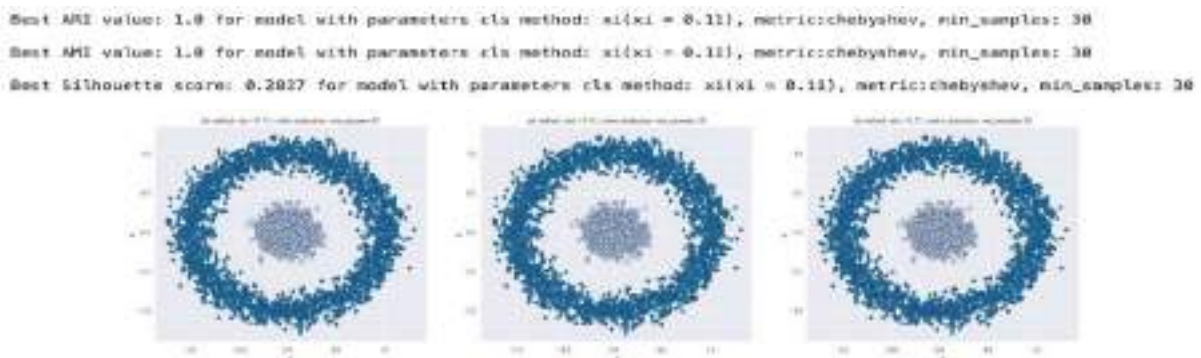


Рисунок 6. Найкращі значення кожної з метрик для першого набору даних та графіки відповідних моделей

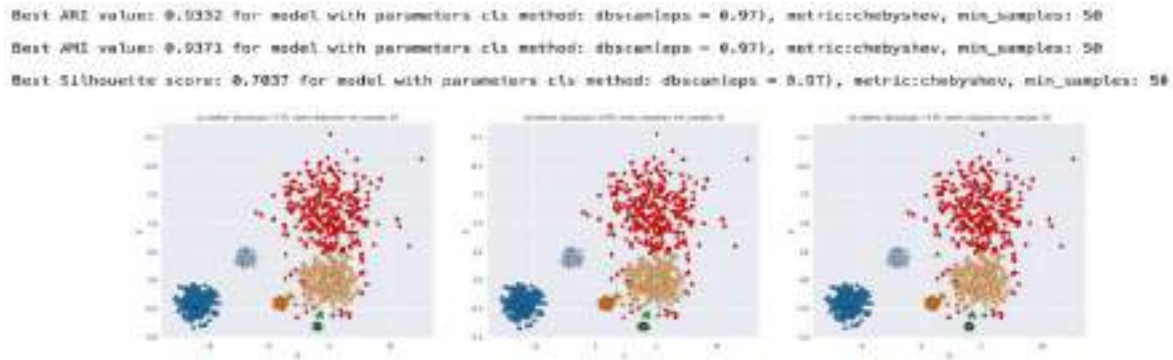


Рисунок 7. Найкращі значення кожної з метрик для другого набору даних та графіки відповідних моделей

З рисунка 7 можна побачити, що для другого набору даних найкращим значенням мір якості також відповідає одне й те ж розбиття. Проте ця модель визначає тільки 5 кластерів правильно, а шостий – як шум. Такі гарні результати метрик пов'язані з тим, що вони не відрізняють лейбл, який відповідає шумовим елементам, і враховують його як звичайний кластер. Тож найкращою є все таки наступна модель (рис. 8), яка має приблизно такі ж значення мір якості:

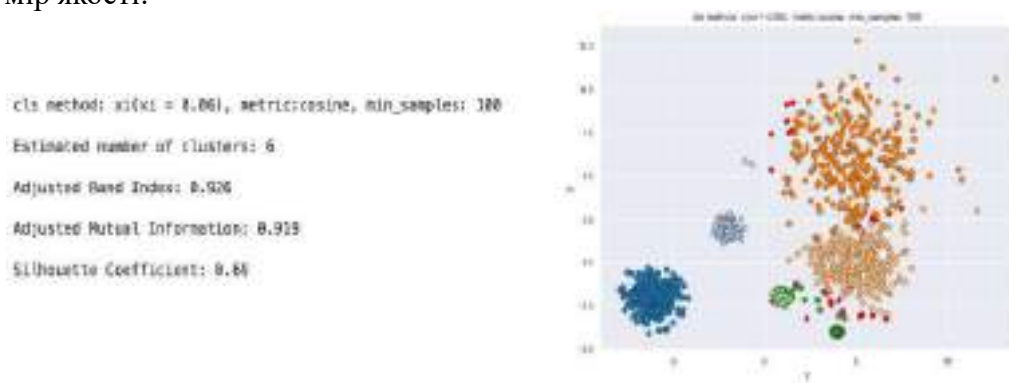


Рисунок 8. Графік моделі з хорошими значеннями метрик та шістьма визначеними кластерами

4. ВИСНОВКИ

Дослідження підтвердило ефективність OPTICS у визначенні кластерів та аналізі структури даних. Граф досяжності визначив межі кластерів та точки переходу. Оптимальні параметри налаштовані для кращого результату на кожному наборі даних. Метрики якості кластеризації оцінили точність та відокремленість кластерів, підтверджуючи успішне застосування OPTICS для аналізу даних.

Важливо врахувати, що деякі точки визначені як шум, що може вимагати додаткового дослідження. Загальний висновок полягає в тому, що OPTICS - потужний інструмент для кластеризації даних, проте його оптимальність залежить від конкретного контексту та завдань.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Feng L., Liu K., Tang F., Meng Q. GO-DBSCAN: improvements of DBSCAN. Algorithm based on grid, *Int. J. Comput. Theory Eng.* 2017. Vol. 9. P. 151–155.
2. Ester M., Kriegel H.P., Sander J., Xu X. A density-based algorithm for discovering. Clusters in large spatial databases with noise. *Second Int. Conf. Knowl. Discov. Data Min.* 1996. P. 226–231.

3. Ankerst M., Breunig M.M., Kriegel H.-P., Sander J. OPTICS: ordering points to identify. The clustering structure. *ACM SIGMOD Record*. 1999. Vol. 28. P. 49–60.
4. Schubert E., Gertz M. Improving the cluster structure extracted from OPTICS plots. *CEUR Workshop Proc.* 2018. Vol. 2191. P. 318–329.
5. Cheng D., Xu R., Zhang Bo, Jin R. Fast density estimation for density-based clustering methods. *Neurocomputing*. 2023. Vol. 532. P. 170–182.
6. Wu C., Chen Y., Dong Y., Zhou F., Zhao Y., Liang C.J. VizOPTICS: Getting insights into OPTICS via interactive visual analysis. *Computers & Electrical Engineering*. 2023. Vol. 107.
7. Wang J., Liu Z., Zhao Y., Xie Y., Xie Y. EAST-NBI experimental data processing method based on improved OPTICS algorithm. *Fusion Engineering and Design*. 2021. Vol. 172. P. 53–68.
8. Kamil I.S., Al-Mamory S.O. Enhancement of OPTICS' time complexity by using fuzzy clusters. *Materials Today: Proceedings*. 2023. Vol. 80. P. 2625–2630.
9. Grover N., A study of various fuzzy clustering algorithms. *International Journal of Engineering Research*. 2014. Vol. 3. P. 177–181.
10. Hoffman_M., Steinley_D., Brusco M. J. A note on using the adjusted Rand index for link prediction in networks. *Social Networks*. 2015. Vol. 42. P. 72–79.
11. Lazarenko D., Bonald T. Pairwise Adjusted Mutual Information. *NeurIPS 2021 Conference*. 2021. P. 53–59.
12. Shahapure K. R., Nicholas C. Cluster Quality Analysis Using Silhouette Score. *IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. 2020. P. 124–127.

МОДЕЛЮВАННЯ СЕЙСМІЧНИХ ХВИЛЬ З ВИКОРИСТАННЯМ МАШИННОГО НАВЧАННЯ

Каніовська І.Ю.¹, Кавара А.О.²

Національний технічний університет України «Київський політехнічний інститут
ім. Ігоря Сікорського»

¹ ikaniovska@gmail.com, ² akavara2000@gmail.com

Відомо, що значна частка населення Землі проживає у сейсмічно активних районах. Приміром, на Тихоокеанському вогняному кільці (найбільш сейсмічно активна частина світу) розташовані такі країни, як Японія (із населенням близько 130 млн осіб), Філіппіни (100 млн), Індонезія (270 млн). Землетруси можуть нести значну небезпеку для життя населення, спричиняти руйнуванню будівель та інфраструктури. Саме тому важливим є покращення методів аналізу сейсмічних хвиль, зокрема застосовуючи методи машинного та глибокого навчання. Результатом дослідження є створення найбільш ефективної моделі, яка може відтворювати основні властивості сейсмічних хвиль.

Ключові слова: сейсмічні хвилі, аналіз часових рядів, машинне навчання, глибокі нейронні мережі.

1. ВСТУП

Землетруси несуть велику небезпеку для населення через масштаби потенційних руйнувань та раптовість настання цієї події. Однією із найбільших проблем є те, що цю подію неможливо передбачити завчасно, що значно зменшує можливості для створення антикризової стратегії запобігання важким наслідкам, таким, як знищення інфраструктури та людські втрати.

Із розвитком машинного та глибокого навчання, збільшенням обчислювальних потужностей, а також зі збільшенням кількості даних, яка накопичується дослідницькими станціями, є змога краще дослідити природу даного явища та більш ефективно оцінювати його основні характеристики. Проведення експериментів із моделюванням сейсмічних хвиль є важливим кроком для того, щоб наблизитися до можливості отримувати більш точні прогнози настання та сили майбутніх землетрусів.

2. МЕТОДИ ТА ПІДХОДИ ДО АНАЛІЗУ ТА МОДЕЛЮВАННЯ СЕЙСМІЧНИХ ХВИЛЬ

Основним джерелом дослідження сейсмічних явищ є збір та аналіз сейсмограм. Сейсмограма — це графік, який виводиться спеціальним приладом – сейсмографом. Даний графік визначає запис руху землі на вимірювальній станції як функції часу. Сейсмограми зазвичай реєструють коливання у просторі R^3 (координат x, y і z), причому вісь z перпендикулярна до поверхні Землі, а осі x і y паралельні поверхні. Енергія, виміряна на сейсмограмі, може бути результатом землетрусу або іншого джерела, наприклад вибуху.

До основних завдань до обробки цифрових сейсмограм відносять:

- дослідження характеристик, що залежать від часу, а саме:
 - виявлення імпульсів, їх фільтрацію, відновлення та симуляцію;
 - обирання фази землетрусу;

- поляризаційний аналіз;
- аналіз веспаграм (velocity spectrum analysis, спектральний аналіз швидкостей хвиль);
- формування векторів напрямку на основі вимірів набору хвиль.
- аналіз частотно-хвильових характеристик (f-k аналіз) та спектральний аналіз.

Основним вхідним параметром для визначення факту настання сейсмічної події є прибуття так званої Р-хвилі (primary or pressure wave, початкова хвиля або хвиля тиску). Особливістю Р-хвилі є те, що вона рухається найшвидше серед інших сейсмічних хвиль, тому під час певної сейсмічної події саме вона є першочерговим індикатором землетрусу.

Щодо виявлення імпульсів та появи Р- та S- (secondary waves) хвиль, то стандартний підхід передбачає обчислення ковзного середнього із різною довжиною вікон: короткострокове (STA, short time average) та довгострокове (LTA, long time average). Після цього обчислюється відношення сигнал/шум (SNR, signal-to-noise-ratio). Тоді сейсмічний сигнал виявляється після того, як SNR перевищує певне визначене значення. Перед підрахунком даної метрики хорошою практикою є використання фільтрів для очищення сейсмограми від шумів. Фільтрування також використовується для стандартизації вигляду сейсмограми та приведення її до одного зі стандартних виглядів [1].

Оскільки сейсмічним хвилям притаманні епістемічна невизначеність (тобто невизначеність, яка виникає через недостатні знання про систему, зокрема фізичні та геологічні особливості даних процесів) та алеаторична невизначеність (що виникає через помилки вимірювання), деякі моделі та алгоритми можуть мати обмеження та занадто сильні вимоги до розподілу даних, через що їхнє використання буде неефективним. Саме тому, при створенні моделі варто надати перевагу методам глибокого навчання перед традиційними методами машинного навчання. Зокрема, найбільш ефективними у даному випадку будуть наступні алгоритми, які можна додати до архітектури моделі навчання:

- на основі байєсовських технік (шар дропауту Монте-Карло, ланцюги Маркова у поєднанні із методом Монте-Карло, варіаційний вивід, варіаційні автокодувальники, навчання із підкріпленням та ін.)
- на основі ансамблів (глибокі ансамблі та їхні варіації) [2].

Ефективними також є згорткові нейронні мережі, генетичні та еволюційні алгоритми для аналізу [3].

3. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

У якості основного датасету було обрано STanford EArthquake Dataset (STEAD) [4]. Масив даних складається із майже 1.2 мільйона часових рядів, які відповідають більш, ніж 19000 годин запису сейсмічних сигналів. Дані складаються із двох частин:

- локальні хвилі землетрусів (тобто такі, які були записані на відстані не більш, ніж 350 км від гіпоцентру);
- шумові хвилі (тобто ті, які відповідають руху автомобілів, вибухам на поверхні, видобутку корисних копалин, та безпосередньо не відносяться до тектонічних процесів Землі).

Загалом дані були отримані із 2613 станцій, які розміщені на всіх континентах. Дані станції розміщені у сейсмічно активних районах, що забезпечує різноманітність вхідних даних по різним характеристикам (тривалість, магнітуда, глибина землетрусу тощо).

Приклад вхідних даних представлено на рис. 1.

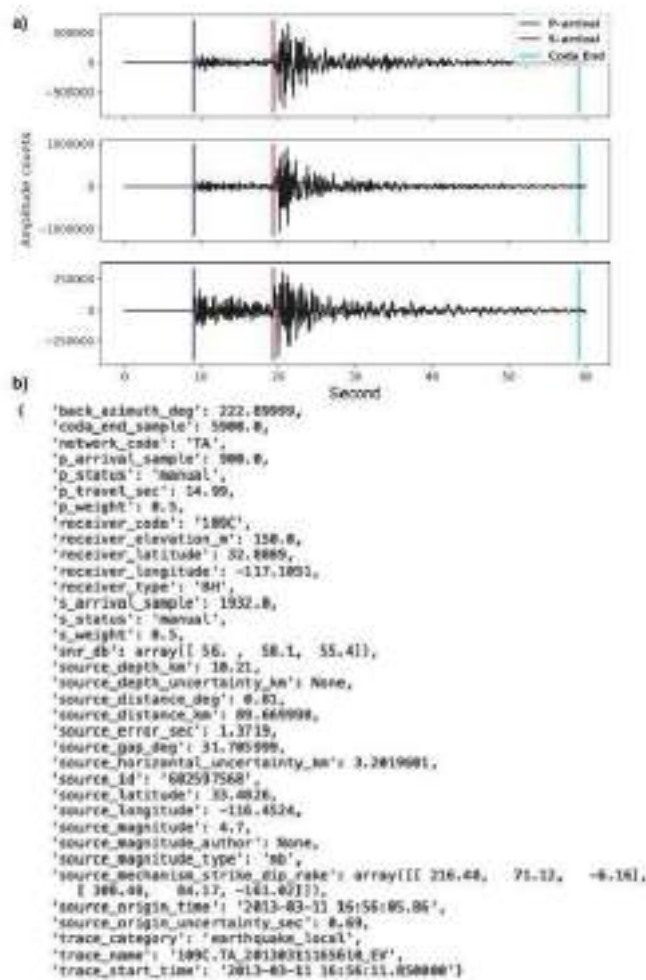


Рисунок 1. Приклад сейсмограми, що відповідає землетрусу. Секція а) часові ряди для поштовхів у напрямках схід-захід, північ-південь та для вертикальні поштовхів. Секція б) метадані сейсмограми

Власна модель прогнозування магнітуди землетрусів ґрунтується на моделі MagNet (Machine-Learning Approach for Earthquake Magnitude Estimation) [5]. Архітектура моделі представлена на рис. 2.

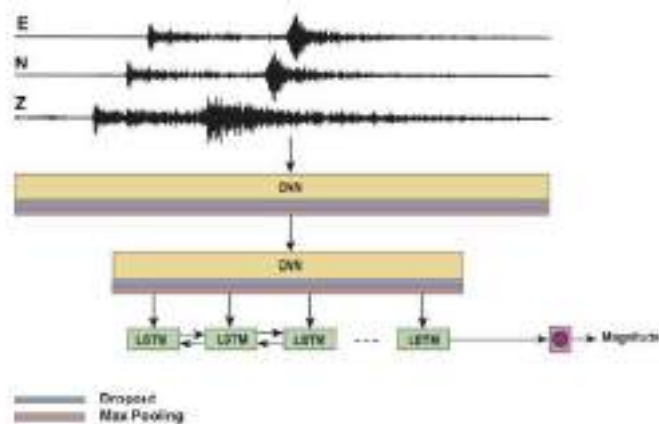


Рисунок 2. Архітектура моделі MagNet

На вхід подаються часові ряди для поштовхів усіх напрямків (схід-захід, північ-південь, ортогональні коливання). Далі дані обробляються за допомогою двох шарів згорткових нейронних мереж із дропаутом та шаром maxpooling 4*4 без функцій активації. Після цього іде шар зв'язних LSTM (long-short term-memory) комірок.

Для визначення часу прибуття Р- та S-хвилі необхідна складніша модель із більшою кількістю шарів. За основу власної моделі була взята попередньо навчена модель EQTransformer [6]. Архітектура моделі представлена на рис. 3.

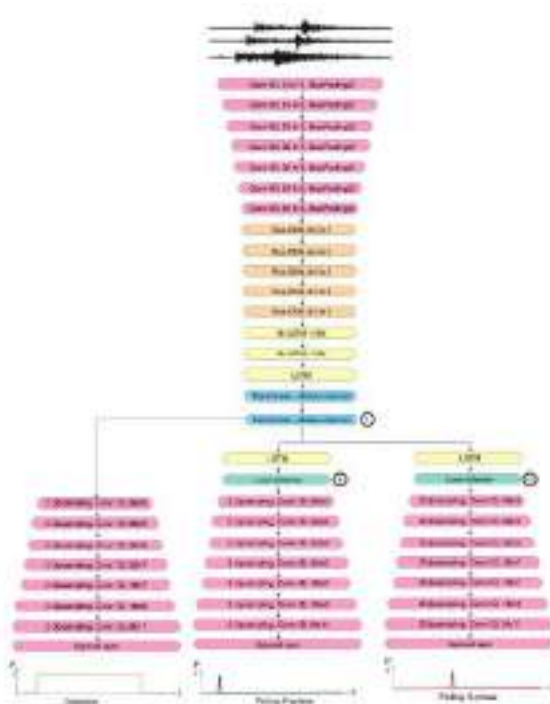


Рисунок 3. Архітектура моделі EQTransformer

Загальна структура мережі містить один глибокий кодувальник і три окремих декодувальники. Кодувальник приймає на вхід часовий ряд сейсмічного сигналу та генерує представлення високого рівня та контекстну інформацію про їх часові залежності. Потім цю інформацію використовують декодувальники для відображення ознак високого рівня для трьох послідовностей імовірностей, які пов'язані з існуванням землетрусу, Р-фазою та S-фазою відповідно.

Значення основних метрик для отриманої моделі для визначення магнітуди землетрусу представлена у таблиці 1.

Таблиця 1. Метрики визначення якості моделі

Метрика	Значення
Співвідношення тренувальних та тестувальних даних	4:1
MAE (Mean Average Error)	0,1987
Стандартне відхилення MAE	0,2245
MSE (Mean Squared Error)	0,0899
Середня алеаторна невизначеність	0,0126
Середня епістемічна невизначеність	0,0679

Таким чином виявлено, що найвдаліша архітектура передбачає комбінацію шарів рекурентних та згорткових нейронних мереж. В подальшому можна використовувати моделі, які основані на архітектурі трансформера; байєсовські методи, що допоможуть визначати землетруси, які настають через одну й ту саму сейсмічну подію (наприклад, виверження вулкану).

4. ВИСНОВКИ

Землетруси – це небезпечне природне явище, яке впливає на життя населення Землі. Своєчасне повідомлення щодо настання землетрусів у певній місцевості має велику кількість переваг, оскільки допоможе знизити ризики та зберегти людські життя.

Основним носієм інформації про землетруси є сейсмограми. Правильна інтерпретація сейсмічних хвиль допомагає у подальшому вивченні даного природного явища. У ході дослідження було розглянуто основні завдання та принципи побудови моделей для проведення аналізу сейсмограм. У ході дослідження з'ясувалося, що модель, яка ґрунтується на шарах рекурентних та згорткових нейронних мереж, має непогані показники для поставленої задачі. Також було проведено експерименти щодо визначення магнітуди землетрусів на основі часових рядів сейсмічних сигналів, які підтвердили ефективність обраної стратегії.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Bormann P. Chapter 11. Data Analysis and Seismogram Interpretation / P. Bormann, K. Klinge, S. Wendt., 2009. – 102 с.
2. A review of uncertainty quantification in deep learning: Techniques, applications and challenges / [M. Abdar, F. Pourpanah, S. Hussain та ін.] // Information Fusion / [M. Abdar, F. Pourpanah, S. Hussain та ін.], 2021.
3. Research on Seismic Signal Analysis Based on Machine Learning / [Y. Xinxin, L. Feng, C. Run та ін.] // Special Issue Intelligent Computing and Remote Sensing / [Y. Xinxin, L. Feng, C. Run та ін.], 2022.
4. Mousavi, S. M., Sheng, Y., Zhu, W., Beroza G.C., (2019). STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI, IEEE Access, doi:10.1109/ACCESS.2019.2947848
5. Mousavi, S. M., & Beroza, G. C. (2019). A Machine-Learning Approach for Earthquake Magnitude Estimation. Geophysical Research Letters.
6. Mousavi, S.M., Ellsworth, W.L., Zhu, W., Chuang, L, Y., and Beroza, G, C. Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. Nat Commun 11, 3952 (2020). <https://doi.org/10.1038/s41467-020-17591-w>

МОДЕЛІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ ОЦІНЮВАННЯ ФІНАНСОВИХ МОДЕЛЕЙ

Коваленко О.М.¹, Гуськова В.Г.²

Національний технічний університет України «Київський політехнічний інститут
імені Ігоря Сікорського», Київ, Україна

¹ kovalenko.oleksandr@lil.kpi.ua, ² guskovavera2009@gmail.com

Робота присвячена дослідженню і створенню моделей та розробці ефективних методів і алгоритмів інтелектуального аналізу фінансових даних з задачею прогнозування і визначення аномалій

Ключові слова: математична статистика, теорія ймовірностей, нейронні мережі, моделі і методи інтелектуального аналізу даних

1. ВСТУП

Сучасний світ щоденно виробляє 2,5 мільярди гігабайт інформації на день, що накопичуються, зберігаються і потребують аналізу. Те ж стосується і економічних даних. Однак, зростання обсягів даних та їх складнощів створюють виклики для традиційних методів аналізу [1].

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

Однак, обробка та аналіз великого обсягу фінансових даних вимагають потужних обчислювальних ресурсів. В цьому контексті, хмарні технології надають універсальне та масштабоване середовище для зберігання, обробки та аналізу даних, що робить їх надзвичайно привабливими для фінансових аналітиків та дослідників.

У цій роботі міститься дослідження та розробка моделей і методів інтелектуального аналізу фінансових даних з використанням хмарних технологій. Конкретно, робота спрямована на створення ефективних алгоритмів та методів для оцінювання фінансових даних, прогнозування трендів та виявлення аномалій.

2. ОГЛЯД МОДЕЛЕЙ ТА МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

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

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

- він відображає результати дослідження системи, відображаючи об'єктивну реальність.
- він представлений у зрозумілій людині формі, використовуючи загальноприйняті символи, поняття та природну мову.
- він компактний у своєму описі, що дозволяє його легко розуміти, інтерпретувати та використовувати.

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

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

Поділяються на стаціонарні та нестаціонарні.

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

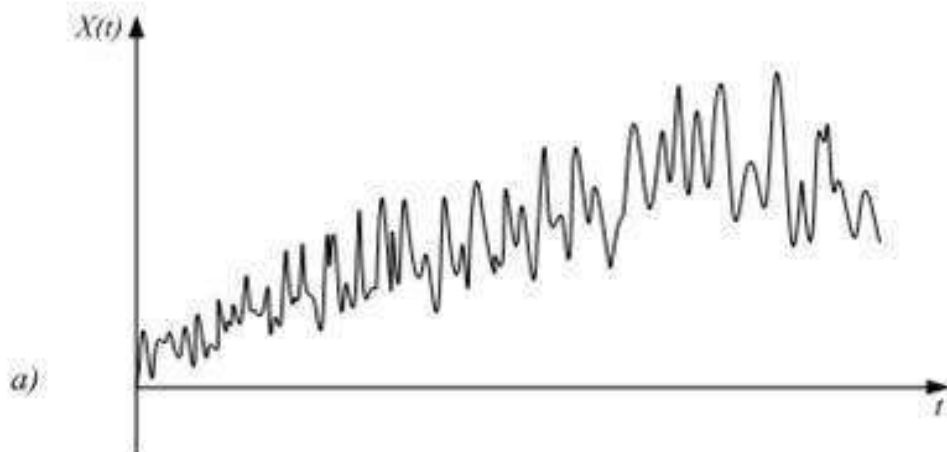


Рисунок 1. Вигляд нестаціонарного часового ряду

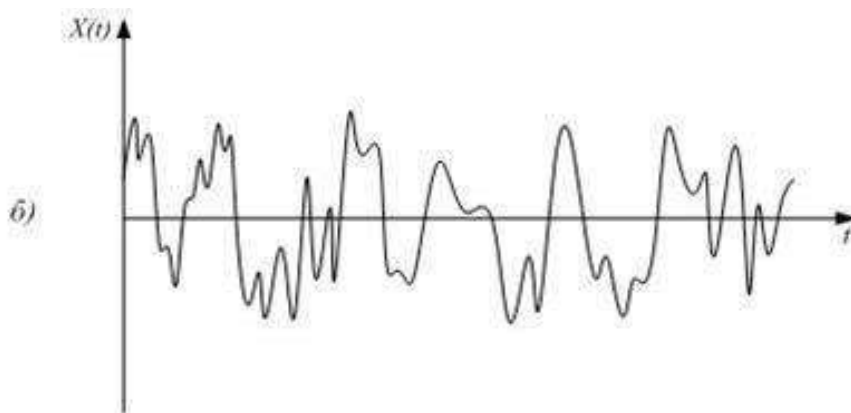


Рисунок 2. Вигляд стаціонарного часового ряду

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

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

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

Інтелектуальний аналіз даних (ІАД) найчастіше розв'язує чотири основних завдання: асоціацію, кластеризацію, класифікацію і регресію.

Також існують ситуації нерівномірного розподілення даних у датасеті, коли з'являється необхідність застосовувати гібридний тип навчання, як з вчителем, так і без вчителя. Наприклад, кількість маркованих даних більша. В такому разі відбувається навчання з вчителем, будується аналітична модель, після цього відбувається навчання без вчителя з підкріпленням і побудовою аналітичної моделі та просто використовуючи навчання з підкріпленням [2].

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

1. Перевірити значущість різниці двох середніх значень підмножин вибірки за критерієм перевірки гіпотези про рівність середніх двох нормально розподілених вибірок (z-критерій).
2. Перевірити сталість дисперсії. Наприклад, використовуючи F-критерій (критерій Фішера про відношення вибірових дисперсій).

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

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

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

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

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

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

Велика кількість майже некорельованих моделей (дерев) як комітет буде працювати краще, ніж будь-яка з окремих моделей-компонентів. [5]

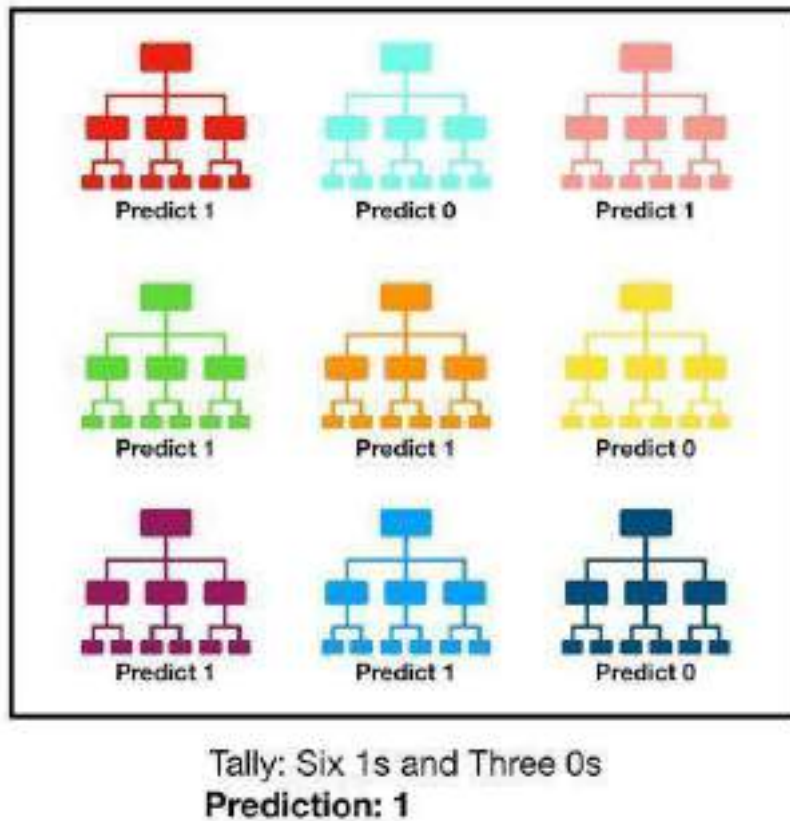


Рисунок 3. Візуалізація методу випадкового лісу, що робить передбачення

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

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

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

3. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

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

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

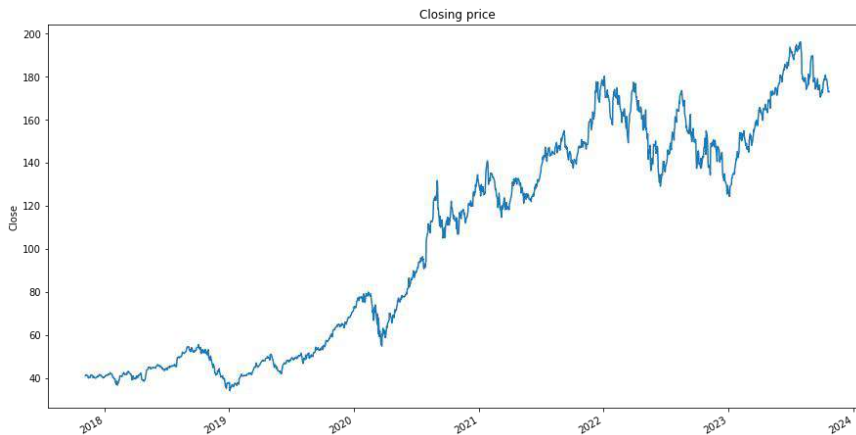


Рисунок 4. Close параметр відносно часу

Для прогнозування даного набору даних було створено 7 моделей.

1. Vanilla LSTM. Це стандартна реалізація мережі Long Short-Term Memory, що є варіацією рекурентної нейронної мережі і призначена для роботи послідовностями даних, такими як звук, текст, часові ряди. Основна ідея – відтворення LSTM без модифікацій або розширень. Має два шари, центральний та прихований, які оновлюються і передають інформацію в кожному кроці.

2. Stacked LSTM. Є модифікацією стандартної LSTM моделі, де декілька прихованих шарів розташовані один за одним (або один на одному – звідси і назва Stacked). Вони краще розуміють інгібиторні та зворотні залежності, мають більшу глибину, краще адаптуються до складних завдань та підходять для задач, де часові залежності є важливими.

3. Bi-directional LSTM. Чергова варіація LSTM, що дозволяє інформації переміщуватися в обох напрямках вздовж послідовності даних, на відміну від стандартного відтворення. Головна перевага – здатність враховувати контекст як попередньої так і наступної частини даних. Хоч дана модель і потребує більших обчислювальних потужностей, вона є досить універсальною та може використовуватися для різних задач.

4. ARIMA – модель часового ряду, яка використовується для аналізу та прогнозування часових рядів. Має такі характеристики:

(a) AR – авторегресія, залежність поточного значення часового ряду від попередніх.

(b) I – інтегрування, процедура диференціації, що призначена для перетворення ряду до стаціонарного, тобто позбавленого тренду та сезонності.

(c) MA – ковзне середнє, середнє значення змінних за певний період часу

5. SARIMA – модифікація ARIMA з додаванням компоненту сезонності. Така модель призначена для моделювання з вираженими сезонними паттернами.

6. Decision tree, дерево рішень, один з найбільш популярних алгоритмів Інтелектуального Аналізу Даних для вирішення проблем регресії та класифікації. Має структуру дерева, де внутрішні вузли представлені прийняттям рішень на основі конкретної ознаки, а кожен листок представляє прогнозоване значення. Ця модель легко інтерпритується та візуалізується для розуміння, як модель приймає рішення, корисна в визначенні ознак, проте попри все, може бути легко перенавченою

7. Random Forest – це ансамбльний алгоритм машинного навчання, що базується на деревах рішень. Така модель робить багато прогнозів для кожного прикладу даних і, в результаті, дані об'єднуються для отримання точного, зазвичай, результату. Однією з переваг можна назвати можливість працювати з великою кількістю ознак і даних. Random Forest є

популярним через високу точність, стійкість до перевантаження та можливість працювати з різними типами даних.

Результати якостей моделей

Таблиця 1. Результати навчання моделей на даних компанії_1

	R2	MSE	RMSE	MAE	MAPE
Vanilla LSTM	0,828	60,889	7,803	6,041	0,037
Stacked LSTM	0,889	39,221	6,263	5,454	0,035
Bi-directional LSTM	0,968	11,324	3,365	2,627	0,017
ARIMA	-2,507	0,053	0,229	0,197	0,039
SARIMA	-1,434	0,037	0,191	0,152	0,03
Decision Tree	-1,158	763,652	27,634	21,764	0,131
Random Forest	-1,131	754,069	27,46	21,572	0,13

Для порівняння візьмемо дані іншої компанії, що має суттєвий обвал акцій на початку 2022 року після різкого зростання. Надалі вартість то зростала, то спадала. Нижче на рис. 5 наведений графік вартості акції наприкінці торговельного періоду.

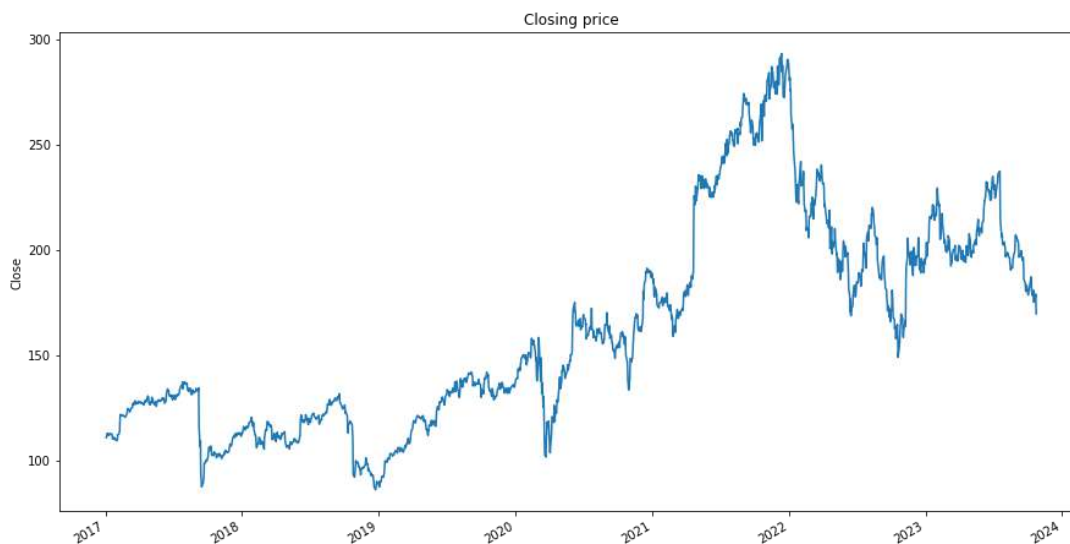


Рисунок 5. Вартість акцій компанії_2

Тепер побудуємо аналогічні моделі та поспостерігаємо за їх метриками.

Таблиця 2. Результат якості моделей компанії_2

	R2	MSE	RMSE	MAE	MAPE
Vanilla LSTM	0,926	22,065	4,697	3,689	0,019
Stacked LSTM	0,929	21,225	4,607	3,572	0,018
Bi-directional LSTM	0,939	18,154	4,261	3,167	0,016
ARIMA	-4,131	0,041	0,203	0,18	0,034
SARIMA	-2,704	0,03	0,172	0,154	0,029
Decision Tree	0,962	11,646	3,413	2,634	0,014
Random Forest	0,972	8,477	2,912	2,311	0,012

4. ВИСНОВКИ

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

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

Таким чином, можна дійти до деяких висновків аналізуючи наведені факти. При розробці моделей для Інтелектуального Аналізу Даних потрібно завжди брати до уваги декілька варіантів для його здійснення, зациклення лиш на одному, чи двох може мати серйозні негативні наслідки у вигляді поганого прогнозу чи оцінці ситуації. Компетентний та структурований підхід має неабияку роль для досягнення високого результату. Лише в стабільних умовах за відсутності зовнішніх чинників історичне моделювання дає точні результати. Для реального світу з сучасними проблемами необхідно завжди обирати комплекс заходів.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Методи інтелектуального аналізу даних URL: <https://buklib.net/books/24506/>
2. Ланде Д.В., Субач І.Ю., Бояринова Ю.Є. Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки: навчальний посібник. — К.: ІСЗЗІ КПІ ім. Ігоря Сікорського», 2018. — 297 с.
3. Гороховатський В. О. Методи інтелектуального аналізу та оброблення даних : навч. посіб. / В. О. Гороховатський, І. С. Творошенко ; М-во освіти і науки України, Харків. нац. ун-т радіоелектроніки. – Харків : ХНУРЕ, 2021. – 92 с.
4. Kuznietsova N. V. Identification and dealing with uncertainties in the form of incomplete data by data mining methods. System research and information technologies. 2016. No. 2. P. 104. URL: <https://doi.org/10.20535/srit.23088893.2016.2.10> (date of access: 08.06.2023).
5. Understanding Random Forest URL: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

СУЧАСНІ МОДЕЛІ ОЦІНЮВАННЯ ФІНАНСОВИХ РИЗИКІВ

Костенко М.О.¹, Кузнєцова Н.В.

Національний технічний університет України «Київський політехнічний інститут
ім. Ігоря Сікорського»

¹ kostenko.max@iitl.kpi.ua

Сучасні методи математичного моделювання нелінійних процесів охоплюють різноманіття підходів, а стандартним підходом вважається використання моделі ARIMA та її варіацій. Класичні регресійні підходи прогнозують цільову змінну лінійною комбінацією минулих значень цієї змінної. Тому доволі просто використовуються як з теоретичної, так і з обчислювальної точки зору завдяки простій структурі. Даний підхід обмежується складністю врахування великої кількості зовнішніх факторів через проблему мультиколінеарності, а також їх можливий нелінійний вплив. Нейронні мережі навчаються на досвіді і адаптуються до змін середовища, яке моделюється. Нейронні технології застосовуються для нелінійного моделювання, стійкі до інформаційних завад і здатні до узагальнення на основі історичних даних, а їх використання може покращити результат прогнозування. Для роботи з послідовностями представлені нейронні мережі, а також їх комбінації з моделями ARIMA. Це дозволяє вирішити поставлену задачу моделювання з урахуванням нелінійного або комбінованого впливу зовнішніх факторів. Для поліпшення точності прогнозів також представлений метод аугментації часових рядів за допомогою лінійної інтерполяції.

Ключові слова: фінансові ризики, системний підхід, часові ряди, ARIMA, NARX, DNN, LSTM.

1. ВСТУП

Моделювання і прогнозування нелінійних процесів є складним завданням, і ефективні методи варіюються в залежності від конкретного контексту та даних. Для досягнення точних та надійних результатів використовуються різні моделі. Зокрема, модель ARIMA (авторегресійна інтегрована змінна середня) використовується для моделювання та прогнозування стаціонарних часових рядів, зокрема враховуючи автокореляцію та інтегруючи диференціацію. Модель NARX (Nonlinear AutoRegressive with eXogenous inputs) дозволяє враховувати нелінійні взаємодії та вплив зовнішніх факторів на динаміку цін акцій. Ця модель дозволяє включати екзогенні змінні, такі як новини або інші фінансові показники. Нейронні мережі, такі як LSTM (Long Short-Term Memory), виявляються ефективними для моделювання складних та нестаціонарних залежностей в часових рядах цін акцій Tesla. LSTM може автоматично виявляти та враховувати довгострокові зв'язки, а також адаптуватися до змінних умов ринку.

Крім того, для підвищення точності прогнозів, доцільно також використовувати техніки аугментації часових рядів, що забезпечить більш якісне оцінювання моделі на невеликих об'ємах даних.

2. МОДЕЛІ ПРОГНОЗУВАННЯ НЕЛІНІЙНИХ ПРОЦЕСІВ

Прогнозування поведінки валют на фінансових ринках є завданням високої складності. Фінансові ринки визначаються надто багатофакторною природою, і побудова надійної математичної моделі для точного передбачення їхньої динаміки залишається викликом. Це діяльність, де завжди присутні ризики, і щоденно тисячі трейдерів та інвесторів стикаються з несприятливими обставинами. Для подолання цих проблем гравці фінансових ринків створюють різноманітні математичні моделі для прогнозування та аналізу нелінійної динаміки цих процесів.

Для розробки прогнозних моделей у нашому дослідженні ми використовуємо три різновиди: класичну модель ARIMA (Autoregressive Integrated Moving Average), NARX (Nonlinear AutoRegressive with eXogenous inputs) та LSTM (Long Short-Term Memory).

ARIMA (Autoregressive Integrated Moving Average): Модель ARIMA є ефективним методом для прогнозування часових рядів і була запропонована Боксом і Дженкінсом на початку 1970-х років [1]. Ця модель базується на регресії залежної змінної лише за її власними значеннями, поточним значенням та значеннями запізнення випадкової помилки. ARIMA використовується для перетворення нестационарних часових рядів в стаціонарні та ефективно моделює їхню динаміку.

NARX (Nonlinear AutoRegressive with eXogenous inputs): Модель NARX дозволяє враховувати нелінійні взаємодії та вплив зовнішніх факторів на динаміку часових рядів [2, 3]. Цей підхід використовує авторегресійну структуру, але також допускає включення екзогенних вхідних змінних, що може покращити прогнозування в умовах складних та змінних середовищ.

Модель NARX зазвичай приймає форму множини нелінійних різницевих рівнянь:

$$y(t) = f(y(t-1), \dots, y(t-n_y), x(t-d), \dots, x(t-n_x), \varepsilon(t-1), \dots, \varepsilon(t-n_\varepsilon)) + \varepsilon(t),$$

де $x(t)$, $t \geq 1$ – вхідні змінні системи,

n_x, n_y – максимальні вхідні та вихідні лаги відповідно,

n_ε – максимальний лаг білого шуму ε ,

$y(t)$ – досліджуваний вихідний процес, або вихідні змінні,

f – невідома нелінійна функція,

$\varepsilon(t)$ – білий шум, змінна, що враховує наслідки вимірювання шуму, похибок моделювання та невимірних порушень, вважається обмеженою ($|\varepsilon(t)| < \delta$) та некорельованою із вхідними даними та n_ε , та обчислюється залежно від процесу оцінки.

LSTM (Long Short-Term Memory): Модель LSTM є варіацією рекурентних нейронних мереж і здатна ефективно моделювати складні та довгострокові залежності в часових рядах [5]. Вона особливо ефективна для задач прогнозування, де важливі динамічні та нестандартні взаємодії між даними. LSTM може автоматично визначати та враховувати важливі патерни у великому обсязі даних.

Мережі з довго короткостроковою пам'яттю (Long Short Term Memory) – особливий вид рекурентних нейронних мереж (РНМ), здатних до навчання довгостроковим залежностям. Вони дають можливість отримати високоякісні результати на великій різноманітності проблем і в даний момент широко застосовуються для моделювання нелінійних процесів [6, 7].

LSTM спеціально спроектовані таким чином, щоб уникнути проблеми довгострокових залежностей. Запам'ятовувати інформацію на тривалий період часу – це практично їх типова поведінка. LSTM мають ланцюгову структуру як і класичні РНМ, але повторюваний модуль

має іншу структуру. Замість одного нейронного шару тут присутні чотири шари, причому вони взаємодіють особливим чином (Рис. 1).

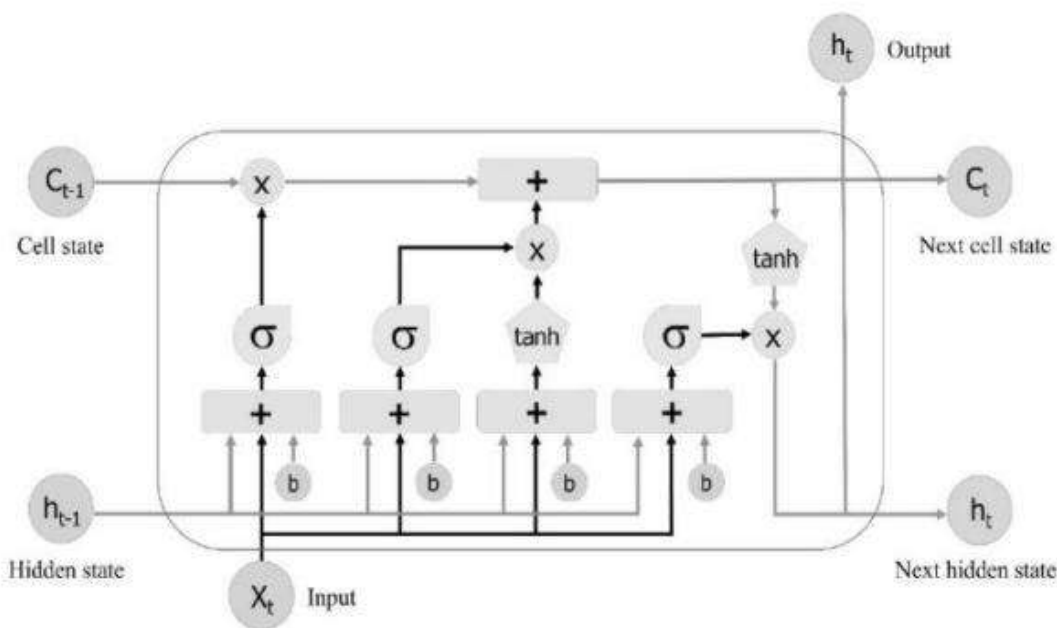


Рис 1. Структура LSTM

3. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Розглянемо функціонування наших розроблених моделей для прогнозу вартості акцій компанії Tesla – глобального гіганта у сфері електричних автомобілів та сонячних батарей. Tesla відома своєю революційною підходом до автомобільного виробництва, звертаючи увагу на стає покращення технологій і виробництва більш доступних електричних автомобілів. Крім того, вони мають величезний вплив на розвиток зеленої енергетики через виробництво сонячних батарей та систем для дому. В контексті прогнозування акцій Tesla, важливо визнати, що це завдання вкрай складне через багатофакторність та велику кількість невизначених змінних, що впливають на фінансові ринки. Tesla, як і багато інших технологічних компаній, піддається впливу різних чинників, таких як ринкова конкуренція, технологічні інновації, зміни у виробництві, регулююча політика та глобальні економічні умови.

Сучасні методи прогнозування акцій включають в себе використання алгоритмів машинного навчання, аналізу глибоких нейронних мереж, а також статистичних методів для аналізу ринкових тенденцій та динаміки. Проте, навіть з використанням передових технологій, точність прогнозування залишається високою ступенем невизначеності через непередбачувані зміни в економіці та інших сферах. Першою моделлю, яка використовувалась для прогнозування є модель ARIMA, для якої потрібно проаналізувати ряд на стаціонарність та обрати найкращу модель.

На основі результатів тестів на стаціонарність ряду, які включають зростаюче середнє та стандартне відхилення на графіку, стає очевидним, що наш ряд не є стаціонарним. Наступним кроком аналізу часового ряду є виконання його перетворення (диференціювання – перехід до попарних різниць сусідніх значень ряду), метою яких є зведення ряду до стаціонарного.

Модель ARIMA включає в себе три основні компоненти: авторегресійну (AR), інтегровану (I) та ковзну середню (MA). Компонента авторегресії AR враховує автокореляцію

між попередніми значеннями часового ряду, інтегрована компонента займається видаленням сезонності та трендів, а компонента ковзного середнього враховує кореляцію між спостереженнями та їхніми випадковими помилками.

Оптимальною моделлю виявилася модель ARIMA(1,1,0), а результати та графік вихідного часового ряду та прогнозу, побудованого за допомогою моделі ARIMA(1,1,0), зображено на рисунку 3.

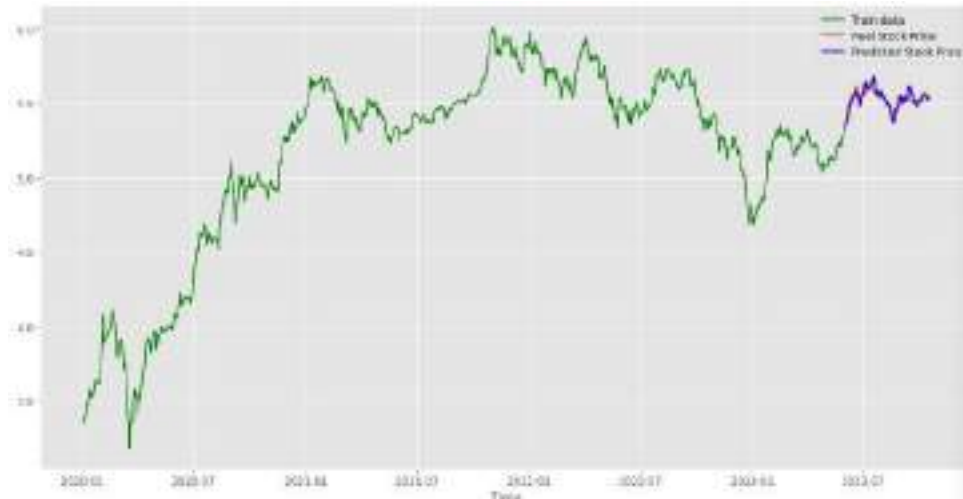


Рисунок 3. Графік фактичних та прогнозних значень

У рамках цього дослідження ми спробуємо різні конфігурації моделі LSTM, і зробимо порівняльний аналіз їхньої ефективності в прогнозуванні майбутніх цін акцій Tesla. Ми також будемо оцінювати точність прогнозів та використовувати різні критерії адекватності, щоб визначити оптимальну модель для даного завдання. Порівняльне дослідження дозволить нам визначити, яка модель та конфігурація є найбільш підходящою для прогнозування цін акцій Tesla в реальних умовах фінансового ринку.

В нашому дослідженні надалі ми будемо використовувати нейронну мережу з наступною архітектурою (Рис. 4):



Рисунок 4. Архітектура НМ

Використання моделі нейронної мережі "LSTM-LSTM-Dense-Dense" для прогнозування часових рядів і послідовностей даних має численні переваги. Ця архітектура відзначається можливістю моделювання довгострокових залежностей в даних завдяки використанню LSTM-шарів. Додатково, вона дозволяє для більш точних та адаптивних прогнозів використовувати різні екзогенні фактори. Важливою перевагою є багаторівнева обробка і абстрагування інформації завдяки повторенню LSTM і Dense шарів. Ця гнучкість дозволяє адаптувати модель до конкретних завдань прогнозування.

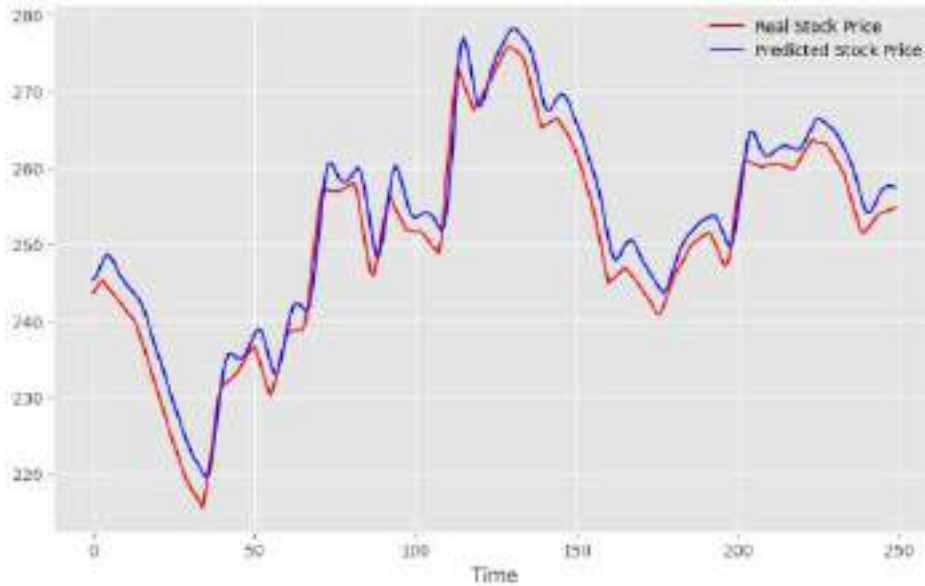


Рисунок 5. Графік фактичних значень та прогнозних за моделлю

Використаємо аугментацію для аналізу нашого часового ряду на прикладі NARX. TSAUG дозволяє створювати більше варіацій та аугментованих прикладів історичних даних для підвищення якості прогнозу. Отже, результати прогнозування після аугментації (Рис. 6):

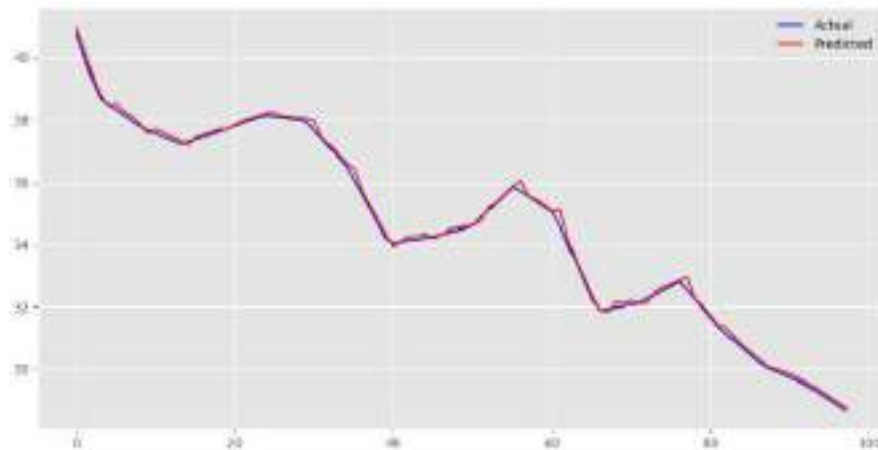


Рисунок 6. Графік фактичних значень та прогнозних за моделлю

Отримані результати свідчать про те, що модель прогнозування акцій Tesla є дуже точною і ефективною. Вона має дуже низькі помилки, високу точність і велику здатність пояснювати зміни в цінах акцій. Це може бути корисним для інвесторів та трейдерів для прийняття обґрунтованих рішень щодо купівлі або продажу акцій Tesla. Хоча початкова модель NARX також показувала чудові результати, в даному випадку вдалось ще покращити результати її роботи (Табл. 1):

Таблиця 1. Характеристики моделей

	ARIMA	LSTM	LSTM+NARX+TSAUG	NARX	NARX+TSAUG
MAE	0,19	10,7302	7,5464	4,224687	0,211251
MSE	0,224	137,504	91,2330	31,55268	0,165859
RMSE	0,3	12,04696	9,5516	5,617177	0,407258
R-Squared	0,9782	0,90434	0,9306	0,976779	0,9998808225637595

Моделі ARIMA та NARX дуже ефективно показують себе на даному набору даних, і їх прогнози можуть бути використані для подальшого прийняття рішення інвестором, але комбінація моделей LSTM+NARX+TSAUG та моделі NARX виявилась ще краще, а додаткове застосування лінійних інтерполяцій значно поліпшує точність прогнозу, і за результатами модель NARX+TSAUG є найкращою.

4. ВИСНОВКИ

Значна частина процесів на фондовому ринку, так само як і динаміка ціноутворення вартостей акцій компанії, є нестационарними часовими рядами, оскільки для них часто є характерними тренд, гетероскедастичність та сезонність, на них впливає велика кількість зовнішніх факторів та їх імовірнісні характеристики змінюються з часом, тобто є функціями часу.

Використання нестационарних даних часових рядів у фінансових моделях може призводити до побудови неякісних моделей та дає ненадійні результати прогнозів з великими похибками. Тому необхідно або зводити часовий ряд до стаціонарного вигляду, або застосовувати моделі, що здатні моделювати різницево-стаціонарні або інтегровані часові ряди. Серед таких можна виділити моделі типу Бокса-Дженкінса (ARIMA, SARIMA, з урахуванням зовнішнього впливу – ARIMAX, SARIMAX, нелінійні – NARMAX), моделі волатильності типу ARCH.

У ході нашого дослідження динаміки цін акцій компанії Tesla з початку 2020 року до жовтня 2021 року було проведено порівняльний аналіз різних моделей прогнозування. Результати показали, що ARIMA та ARIMAX моделі показали високу точність. Однак застосування нелінійної моделі NARMAX з урахуванням впливу зовнішніх факторів, дозволило досягти надзвичайно високої якості прогнозів. Ці результати свідчать про високу ефективність та надійність моделей NARMAX для прогнозування цін акцій Tesla в порівнянні з класичними ARIMA моделями.

Крім того, наше дослідження було збагачено за допомогою методів аугментації даних, які дозволили покращити якість прогнозів. Використання бібліотеки TSAUG перед моделлю NARX сприяло збільшенню обсягу доступних даних та покращило стійкість моделі до шумів і нерегулярностей в даних. Ця аугментація дала можливість створити різноманітні варіанти часового ряду, що сприяє кращому вивченню залежностей та динаміки цін акцій Tesla. Результати цього дослідження демонструють важливість використання аугментації даних як частину процесу прогнозування цін акцій та підкреслюють високу ефективність моделей NARMAX в поєднанні з цим підходом.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. П. І. Бідюк, В. Д. Романенко, та О. Л. Тимошук, Аналіз часових рядів. Київ, Україна: Політехніка, 2010.
2. José Maria P. Menezes Jr., Guilherme A. Barreto. Long-term time series prediction with the NARX network: An empirical evaluation. Neurocomputing, Volume 71, Issues 16–18, 2008, P. 3335-3343

3. R. Hyndman, G. Athanasopoulos, *Forecasting: Principles and Practice*. Melbourne, Australia: OTexts, 2013.
4. Shumway R., Stoffer D., *Time Series Analysis and Its Applications*. New York, USA: Springer, 2011.
5. S. Hochreiter, and J. Schmidhuber, “Long Short-Term Memory”, *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.
6. F. A. Gers, D. Eck, and J. Schmidhuber, “Applying LSTM to Time Series Predictable Through Time-Window Approaches”, in *Proc. of International Conference on Artificial Neural Networks*, Vienna, 2001, pp. 669-676. doi: 10.1007/3-540-44668-0_93.
7. S. Hochreiter, Y. Bengio, and J. Schmidhuber, “Gradient flow in recurrent nets: the difficulty of learning long-term dependencies”. [Online]. Available: <http://www.bioinf.jku.at/publications/older/ch7.pdf> . Accessed on: Dec. 12, 2018.
8. “Understanding LSTM Networks”. [Online]. Available: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>. Accessed on: March 9, 2018.
9. F. Chollet, and J. Allaire, *Deep Learning with R*. New York, USA: Manning, 2018.
10. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, USA: MIT Press, 2016

СИСТЕМА ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ ВИБОРУ S-МОДЕЛЕЙ ЕКОНОМІЧНОГО ЗРОСТАННЯ

Кузьмінчук А.В.¹, Лопатін О.К.²

Національний технічний університет України «Київський політехнічний інститут
імені Ігоря Сікорського», Київ, Україна

¹ anatoliykozminchuk@gmail.com, ² lopatinalexey142@gmail.com

У цій роботі вперше досліджується двухфакторна агрегована S-тренд виробнича функція $y'(t)=P(t)*S(x(t),A,B,a,m,u)$ на основі емпіричних даних, а саме на прикладі індексу ВВП на душу населення Сінгапуру (1960-2022). Тут $y(t)$ вихідний ряд статистичних даних, $P(t)$ -коефіцієнт загальної факторної продуктивності, $y'(t)=S(x(t),A,B,a,m,u)$ детермінована функція, що апроксимує ряд $y(t)$, $S(x(t),A,B,a,m,u) = u + A(1 + B * \exp(-a(t - m)))^{-1}$. Позначимо через $G_y(t), G_p(t), G_s(t)$ річні швидкості зростання у відсотках вихідного ряду, TFP коефіцієнта і внеску фактора "нагромадженням капіталу" Має місце тотожність $G_p(t)=G_y(t)-G_s(t)$. У цій тотожності треба знати $G_s(t)$. Щоб знати $G_s(t)$, потрібно визначити $y'(t)=S(x(t),A,B,a,m,u)$ і отримати оцінку апроксимації $MARE(y(t),y'(t)) = (100\%)/N \sum_{t=1}^N (\frac{y(t)-y'(t)}{y(t)})$, критерій $MARE < 10\%$ вважається відмінним результатом, $10\% < MARE < 20\%$ вважається хорошим результатом.

Ключові слова: S-подібна виробнича функція, сукупна факторна продуктивність, ефективність інвестицій, принцип Чарльза Р. Халтена, фактор продуктивності, фактор нагромадження капіталу.

1. ВСТУП

У сучасній економічній теорії існує широкий спектр методів, які дозволяють розрахувати рівень сукупної факторної продуктивності (далі – СФП), що є одним з ключових показників ефективності виробництва як на рівні окремих фірм, так і на рівні галузей, регіонів та країн. Зростання капіталу, праці та технічного прогресу є трьома основними джерелами економічного зростання країни та регіону. Темпи зростання робочої сили обмежуються темпами зростання населення, особливо в індустріально розвинених країнах, де населення рідко зростає більш ніж на два відсотки на рік, навіть з урахуванням міжнародної міграції. Отже, темпи зростання капіталу (фізичного і людського) і технологічний прогрес є основними джерелами більшої частини економічного зростання.

Модель Солоу (модель Солоу-Свона) – модель екзогенного економічного зростання базується на роботах Солоу [5] і Свона [6], а також на роботі Солоу [7], в якій було введено поняття сукупної продуктивності факторів виробництва (TFP), що отримала назву залишкової Солоу. Модель Солоу вважається відправною точкою для всіх сучасних моделей економічного зростання.

Залишок Солоу все ще залишається, після багатьох десятиліть, робочою конячкою емпіричного аналізу економічного зростання. Було опубліковано тисячі сторінок досліджень, і щороку їх публікується все більше. Приведемо підбірку найбільш актуальних для побудови нових агрегованих виробничих функцій за останні п'ять років: Цонас, Майк Г., Полеміс,

Майкл Л., (2019), Цуніс Ніколас, Стедман Ян; (2021), Френсіс Девід К., Нона Каралашвілі, Хібрет Маемір, Хорхе Родрігес Мез (2020), Вілан Карл (2021); Дандан, Дуду (2020); Харб Жорж, Басіль Шарбе (2023): Узагальнення та подальший розвиток методів вимірювання сукупної факторної продуктивності. Бюро статистики праці США, (випуск 23 березня 2023 року) У цій статті визначено ключові терміни та поняття, які є ключовими для розуміння того, як Бюро статистики праці (БСП) розробляє показники продуктивності для різних рівнів американської економіки.

2. ВИРОБНИЧА ФУНКЦІЯ S-ТРЕНДУ

Нехай досліджується часовий ряд

$$y(t) = y(t_1), \dots, y(t_T). \quad (1)$$

Наприклад, це може бути ВВП на душу населення деякої країни. Для часових рядів прийнято розглядати його рівні як суміш чотирьох компонент – трендової, циклічної, сезонної та випадкової складових, які неможливо виміряти безпосередньо [1]

$$y(t_i) = T(t_i) + C(t_i) + S(t_i) + \varepsilon(t_i).$$

$T(t_i)$ – тренд, основна тенденція розвитку досліджуваного процесу в часі. Цей тренд є детермінованою складовою, незалежною від циклічних, сезонних та випадкових складових. Він може бути представлений у вигляді більш-менш гладкої кривої.

Компоненти часового ряду $T(t_i)$ не є спостережуваними. Вони є теоретичними величинами. Оцінка майбутніх членів ряду зазвичай робиться за допомогою прогнозної моделі. Прогнозна модель – це модель, яка апроксимує тренд. Ми обираємо S-образну криву Верхольста як модель прогнозування тренду (МПТ)

$$y'(t) = u + \frac{A}{1+B \cdot \exp(-a(t-m))} = S(t, A, B, a, m, u) \quad (2)$$

Наслідок 1. Тип прогнозної моделі можна встановити за графіком $(y(t), t)$ вихідного ряду. Отже, вихідні дані (1) мають бути апроксимовані МПТ (2).

Наслідок 2. Точність апроксимації ряду (1) оцінюється за критерієм MAPE.

Формулювання проблеми. Метою даної роботи є розробка нової виробничої функції вигляду $y'(t) = S(x(t), A, B, a, m, u)$, $t=1, \dots, T$.

У цьому випадку факторний показник $x(t)=t$ характеризує внесок компоненти "нагромадження капіталу" в економічний вихід. Це час із кроком в 1 рік (звичайний крок статистичних таблиць).

$$\text{MAPE}(y(t), y'(t)) = \frac{100\%}{N} \sum_{t=1}^N \left| \frac{y(t) - y'(t)}{y(t)} \right|,$$

де $y(t)$ – координати точкового графіка вихідного ряду, $y'(t)$ – координати, що будуються.

3. МЕТОДОЛОГІЯ: ГРАФО-АНАЛІТИЧНИЙ МЕТОД. РУЧНА ОПТИМІЗАЦІЯ

Діаграма технологічного укладу ВВП на душу населення в Сінгапурі 1960–2022.

Джерело даних (Табл. 1): https://www.google.ru/publicdata/explore?ds=d5bncppjof8f9_

Таблиця 1. ВВП на душу населення Сінгапуру в тисячах доларів (у поточних доларах США)

NO	Year	Y	NO	Year	Y	NO	Year	Y	NO	Year	Y
1	1960	4,28	17	1976	27,59	33	1992	161,36	49	2008	400,09
2	1961	4,49	18	1977	28,46	34	1993	182,90	50	2009	389,27
3	1962	4,72	19	1978	31,94	35	1994	215,52	51	2010	472,37
4	1963	5,11	20	1979	39,01	36	1995	249,15	52	2011	538,91
5	1964	4,80	21	1980	49,28	37	1996	262,33	53	2012	555,48
6	1965	5,17	22	1981	55,97	38	1997	262,76	54	2013	569,67
7	1966	5,67	23	1982	60,78	39	1998	238,29	55	2014	575,65
8	1967	6,26	24	1983	66,33	40	1999	217,97	56	2015	556,46
9	1968	7,09	25	1984	72,28	41	2000	238,53	57	2016	568,96
10	1969	8,13	26	1985	70,02	42	2001	217,00	58	2017	611,65
11	1970	9,26	27	1986	68,00	43	2002	221,60	59	2018	668,37
12	1971	10,71	28	1987	75,39	44	2003	237,30	60	2019	660,70
13	1972	12,64	29	1988	80,14	45	2004	276,08	61	2020	612,74
14	1973	16,85	30	1989	103,95	46	2005	299,61	62	2021	777,10
15	1974	23,42	31	1990	118,62	47	2006	337,68	63	2022	828,08
16	1975	24,96	32	1991	145,02	48	2007	394,33			

Побудова циклів S-трендів.

Наведемо в таблиці 2 розбиття вибірки на ділянки для моделювання, а також в таблиці 3 опишемо загальну структуру циклів.

Таблиця 2. Визначення циклів верхніми та нижніми асимптотами

Theta	1	2	3	4	5	6
Значення	4	58	262	221	555	958

Таблиця 3. Вигляд S-кривих

Цикл 1	Цикл 2	Цикл 3	Цикл 4	Цикл 5
$z_1(t) = \frac{T_2 - T_1}{D_1} + T_1$	$z_2(t) = \frac{T_3 - T_2}{D_2} + T_2$	$z_3(t) = \frac{T_3 - T_4}{D_3} + T_4$	$z_4(t) = \frac{T_5 - T_4}{D_3} + T_4$	$z_5(t) = \frac{T_7 - T_5}{D_5} + T_5$

Уточнення до таблиці 3: $z_i(t) = 1 + B_i \text{EXP}(-a_i * (t - m_i))$, $i = 1, \dots, 5$.

На рисунках 1 та 2 можемо спостерігати результати моделювання S кривих.

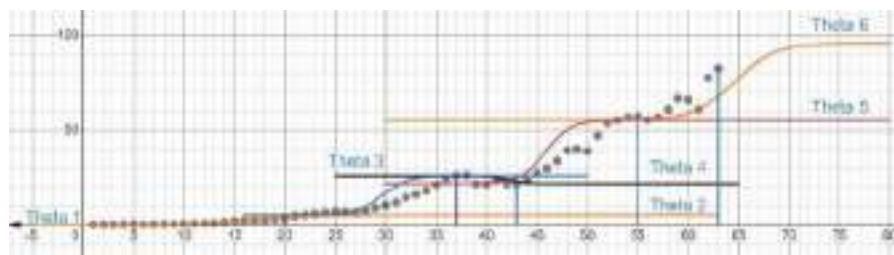


Рисунок 1. Вище наведено $S_i(t) = 1, \dots, 5$, визначені їхніми верхніми та нижніми асимптотами

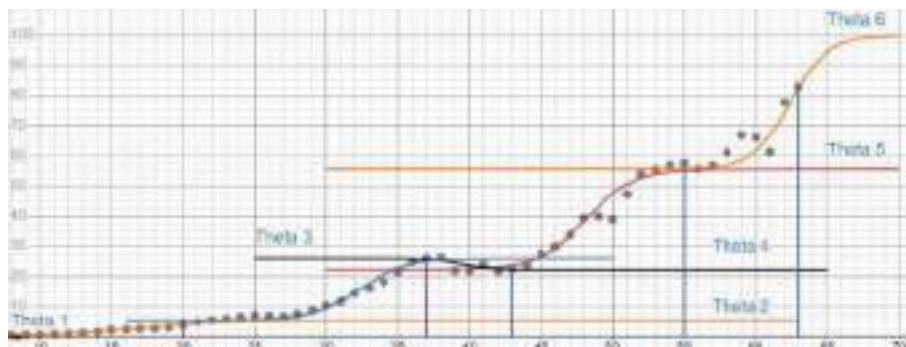


Рисунок 2. Вище наведено оптимізовані $S_i(t) = 1, \dots, 5$, визначені їхніми верхніми та нижніми асимптотами

4. ОГЛЯД РЕАЛІЗОВАНОЇ СППР

Для розв'язання задачі було обрано метод градієнтного спуску, а саме його модифікацію на основі частинних похідних параметрів. Модифікуємо метод так, щоб мінімізувати критерій різниці квадратів [12]:

Визначимо шукану функцію:

$$f(x_i) = \frac{A}{1 + B \exp(a * (m - x_i))}$$

Визначимо формулу різниці квадратів з огляду на поставлену задачу:

$$E = \sum_{i=1}^n (y_i - f(x_i))^2 = \sum_{i=1}^n \left(y_i - \frac{A}{1 + B \exp(a * (m - x_i))} - u \right)^2$$

Для використання градієнтного методу визначимо похідні для всіх параметрів
Визначимо ці похідні:

$$\frac{dE(A, B, a, m, u)}{dA} = \sum_{i=1}^n \left(\left(\frac{-2}{1 + B \exp(a * (m - x_i))} \right) * \left(y_i - \frac{A}{1 + B \exp(a * (m - x_i))} - u \right) \right)$$

$$\frac{dE(A, B, a, m, u)}{dB} = \sum_{i=1}^n \left(\left(\frac{2A \exp(a * (m - x_i))}{(1 + B \exp(a * (m - x_i)))^2} \right) * \left(y_i - \frac{A}{1 + B \exp(a * (m - x_i))} - u \right) \right)$$

$$\frac{dE(A, B, a, m, u)}{da} = \sum_{i=1}^n \left(\left(\frac{2AB(m - x_i) \exp(a * (m - x_i))}{(1 + B \exp(a * (m - x_i)))^2} \right) * \left(y_i - \frac{A}{1 + B \exp(a * (m - x_i))} - u \right) \right)$$

$$\frac{dE(A, B, a, m, u)}{dm} = \sum_{i=1}^n \left(\left(\frac{2aAB(m - x_i) \exp(a * (m - x_i))}{(1 + B \exp(a * (m - x_i)))^2} \right) * \left(y_i - \frac{A}{1 + B \exp(a * (m - x_i))} - u \right) \right)$$

$$\frac{dE(A, B, a, m, u)}{du} = \sum_{i=1}^n \left(-2 * \left(y_i - \frac{A}{1 + B \exp(a * (m - x_i))} - u \right) \right)$$

Наступним етапом потрібно визначити формат кроків ітерації:

$$\begin{cases} A_{n+1} = A_n + \lambda_n^1 \frac{dE(A, B, a, m, u)}{dA} ; \\ B_{n+1} = B_n + \lambda_n^2 \frac{dE(A, B, a, m, u)}{dB} ; \\ a_{n+1} = a_n + \lambda_n^3 \frac{dE(A, B, a, m, u)}{da} ; \\ m_{n+1} = m_n + \lambda_n^4 \frac{dE(A, B, a, m, u)}{dm} ; \\ u_{n+1} = u_n + \lambda_n^5 \frac{dE(A, B, a, m, u)}{du} , \end{cases}$$

де $\lambda_n^i, i = 1 \dots 5$, число менше 1 і більше 0, яке дозволяє впливати на розмір кроку ітерації.

При створенні до СППР було висунуто наступні умови: інтерактивність – система повинна мати інтуїтивно зрозумілий графічний інтерфейс користувача (GUI), який дозволяє користувачам взаємодіяти з системою за допомогою кнопок, меню, графічних елементів тощо; доступність – система повинна бути легкодоступною для користувачів, чим менше користувачу потрібно зробити для початку користування продуктом, тим більше ймовірності привернути його увагу; повнота – в системі має бути можливість провести повноцінний аналіз впливу: починаючи від завантаження даних для тренування моделі, закінчуючи використанням нових даних на створеній моделі та вивантаженням отриманих результатів.

Блок-схема алгоритму роботи з СППР зображена на рисунку 3.

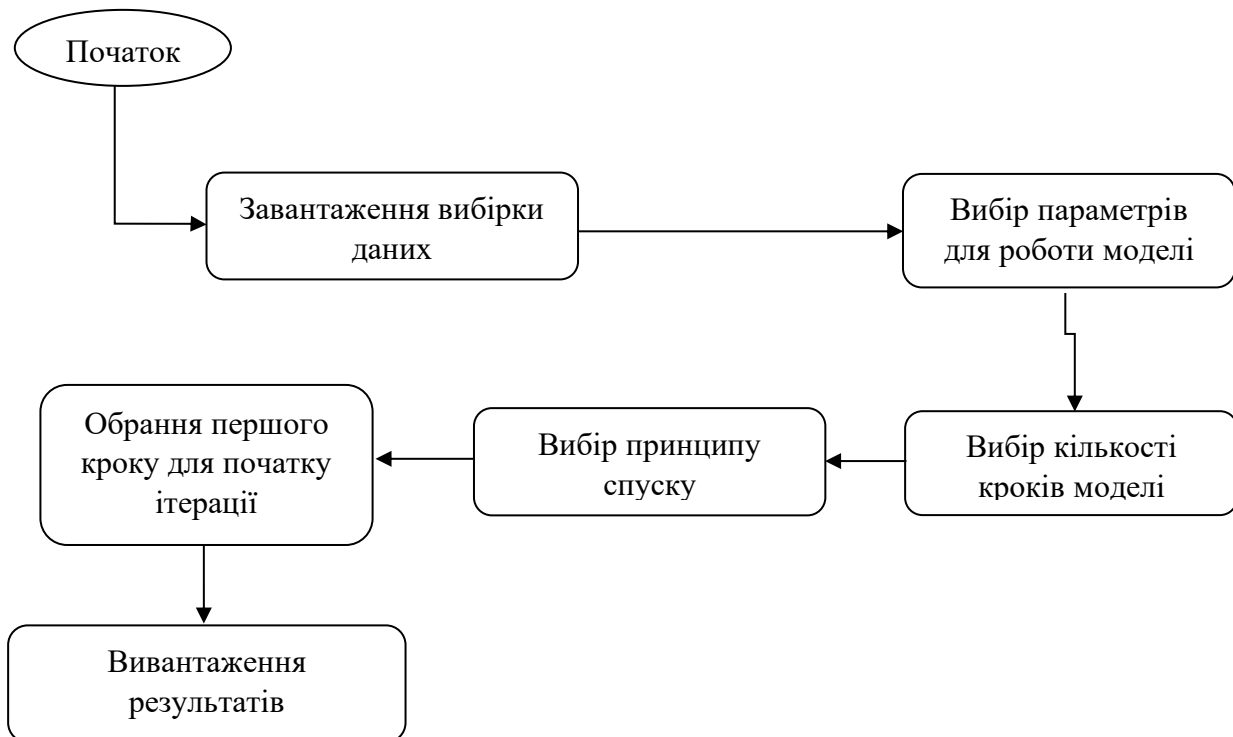


Рисунок 3. Блок-схема алгоритму роботи з СППР

Реалізацію вищезгаданих етапів роботи з СППР зображено на рисунках 4 та 5. На рисунку 4 зображено базовий інтерфейс додатку. На рисунку 5 зображено результат роботи

додатку. За наведеним рисунком можна побачити, що інтерфейс є інтуїтивно зрозумілим і СППР дозволяє провести аналіз, тобто вона відповідає висунутим до неї вимогам.



Рисунок 4. Інтерфейс СППР

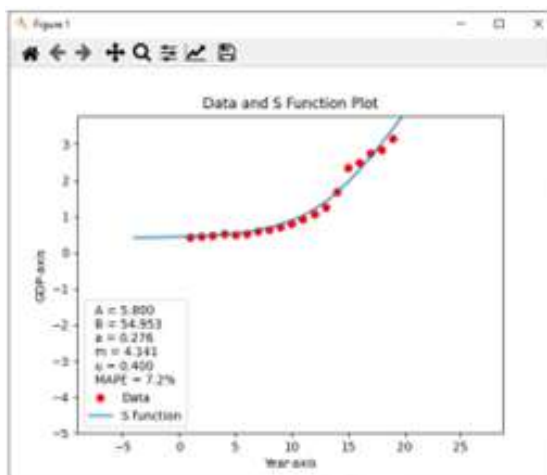


Рисунок 5. Інтерфейс результату роботи СППР

4. РЕЗУЛЬТАТИ РОБОТИ

Виконаємо обробку циклів вище за допомогою СППР, результати представлені на малюнку 5 та в таблиці 3.

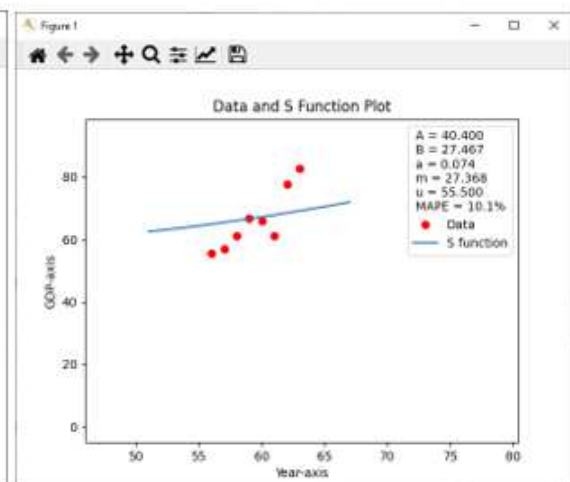
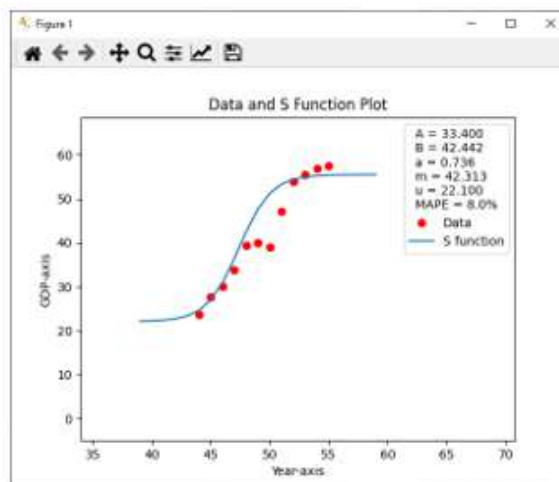
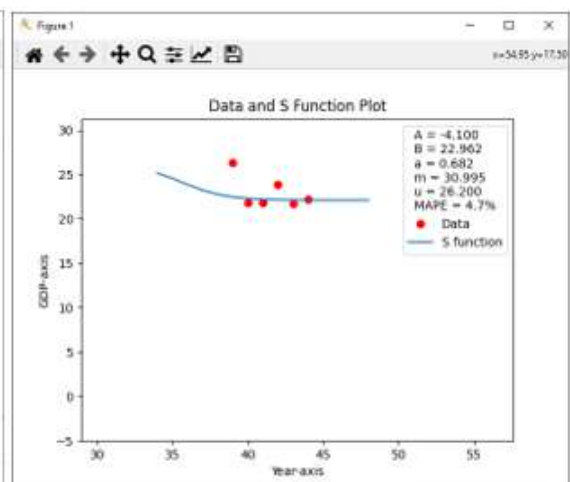
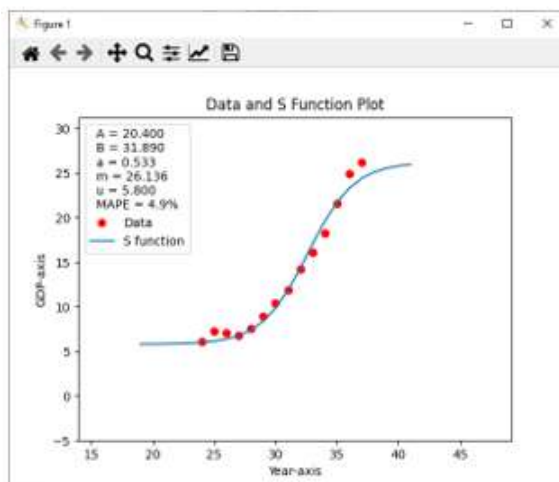
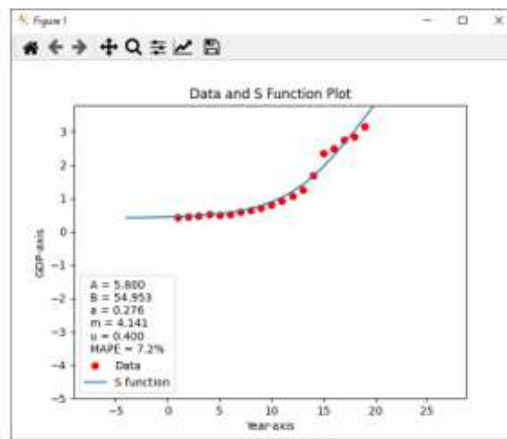


Рисунок 6. Результати обробки за допомогою СППР

Таблиця 4. Зведені результати роботи

Цикл 1				
	B	a	m	MAPE%
general	4	0,5	13	21%
selecting	4	0,35	13	8%
gradient descent	54,953	0,276	4,141	7,2%
Цикл 2				
	B	a	m	MAPE%
general	20	0,8	26	27%
selecting	30	0,6	26,7	8%
gradient descent	31,89	0,53	26,13	4,9%
Цикл 3				
	B	a	m	MAPE%
general	40	-1,2	45	9%
selecting	43	-1	43	5%
gradient descent	22,96	0,68	30,99	4,7%
Цикл 4				
	B	a	m	MAPE%
general	10	0,8	43	22%
selecting	6,4	0,6	45	6%
gradient descent	42,442	0,736	42,313	8,0%
Цикл 5				
	B	a	m	MAPE%
general	100	0,5	60	16%
selecting	6,2	0,75	60	6%
gradient descent	27,467	0,074	27,368	10,1%

5. ВИСНОВКИ

В рамках роботи було проведено роботу над побудовою S-кривої та розроблено СППР на його основі.

Порівняльний аналіз на основі рисунку 6 та таблиці 4 показав, що для розглянутого набору даних СППР прискорив обробку інформації та надав оптимальні результати.

Під час розробки СППР було висунуто вимоги до неї та обрано інструменти для створення. Приклад роботи з СППР та огляд інтерфейсу показують, що СППР відповідає висунутим до неї вимогам.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Бідюк, П. І., Романенко, В. Д., Тимошук, О. Л. (2010). Аналіз часових рядів: навчальний посібник. Київ, Київська область: Політехніка, 317 с.
2. Cıpra T. (2020). Time Series in Economics and Finance. Springer, 353 p

3. Dandan , Doudou (2020) Comparison and Analysis of Measurement Methods of Total Factor Productivity International Journal of Frontiers in Engineering Technology ISSN 2706-655X Vol.2, Issue 1: 18-30, DOI: 10.25236/IJFET.2020.020102
4. Harb Georges, Bassil Charbe (2023). TFP in the Manufacturing Sector: Long-Term Dynamics, Country and Regional Comparative Analysis, *Economies* 2023, 11(2), <https://doi.org/10.3390/economies11020034>
5. Solow R. M. (1956), A contribution to the Theory of Economic Growth, *The Quarterly Journal of Economics*, 70 (1), 65—94.
6. Swan T. W. (1956), Economic growth and capital accumulation, *The Economic Record*, 32(2), 334—361
7. Solow R. M. (1957) Technical change and the aggregate production function, *The Review of Economics and Statistics*, 39(3), \312—320.
8. Tsionas, Mike G., Polemis, Michael L., (2019). On the estimation of total factor productivity: A Bayesian non-arametric approach, *European Journal of Operational Research*, Elsevier, vol. 277(3), pp 886-902.
9. Tsounis Nicholas, Steedman Ian (2021), A New Method for Measuring Total Factor Productivity Growth Based on the Full Industry Equilibrium Approach: The Case of the Greek Economy *Economies* 2021, 9, 114, pp 1-21
10. The Bureau of Labor Statistics (BLS) produces measures of productivity for different levels of the U.S. economy. Total factor productivity (TFP) in the private nonfarm business sector (release March 23, 2023) USDL-23-0540 Technical information: (202) 691-5606 • Productivity@bls.gov • www.bls.gov/productivity Media contact: (202) 691-5902 • PressOffice@bls.gov
11. Wikipedia World Development Indicators." World Bank". https://www.google.ru/publicdata/explore?ds=d5bncppjof89_
12. Wikipedia Gradient descent https://en.wikipedia.org/wiki/Gradient_descent

РОЗРОБКА МОДЕЛЕЙ ОЦІНЮВАННЯ РИЗИКІВ ЗЕЛЕНИХ ПРОЕКТІВ

Кузнєцова Н.В., Шевчук О.С.¹

Національний технічний університет України «Київський політехнічний інститут
ім. Ігоря Сікорського»

¹ oleksii.shevchuk@ukr.net

Забруднення навколишнього середовища є однією з головних глобальних проблем сучасного світу, яка регулярно обговорюється у різних політичних та наукових колах. Світове товариство розуміє катастрофічність можливих наслідків, тому шукає варіанти для запобігання світової екологічної кризи. Розробляються еко-технології, впроваджуються відповідні закони, а також ведеться планування та імплементація зелених проектів. Ці проекти націлені на відновлення та збереження екосистеми, проте постає питання оцінки ефективності та ризикованості таких проектів. Виконуються дослідження щодо оцінки ризикованості зелених проектів як класичними методами, так і на основі нових розроблених методів. У роботі розглянуто класичні методи розв'язання задачі кредитного скорингу, розроблено моделі прогнозування ризиків конкретно зелених проектів

Ключові слова: оцінка ризиків, кредитний скоринг, зелені проекти.

1. ВСТУП

За час існування планети екосистема зазнавала значних змін, спричинених як зовнішніми чинниками, так і внутрішніми. Найбільш значущим внутрішнім чинником є вплив мешканців планети на її екосистему і основну роль у зміні екосистеми світу відіграє людина. Кількість населення неспинно зростає, що збільшує потреби у природних ресурсах. Століттями людство чинило негативний вплив на екосистему Землі. Однак зростає і рівень освіченості суспільства, що дозволило зрозуміти, що у боротьбі з природою переможцем людина бути не може. Усвідомивши небезпеку та критичність ситуації в екологічній сфері, найбільш розвинені країни світу почали вивчати всі аспекти даної проблеми та шукати способи її вирішення [1].

Новітній і на даний час найбільш перспективний спосіб вирішення глобальної екологічної проблеми – це розробка та впровадження зелених проектів. Чіткого формалізованого визначення зелених проектів немає, але усі варіанти визначення включають те, що дані проекти повинні бути спрямовані на зниження кількості споживання природних або енергетичних ресурсів або на їх відновлення. Крім цього, зелені проекти можуть приносити і економічні переваги та дивіденди, хоча це не є їх основною метою.

У зв'язку із появою зелених проектів постає питання коректної оцінки їх економічної та екологічної доцільності. Класичні методи та моделі оцінки проектів орієнтуються лише на економічну складову. У випадку зелених проектів такий підхід не буде повністю коректним. У такому разі, і рішення задачі кредитного скорингу зелених проектів доцільно переосмислити.

2. ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

Метою даної роботи є дослідження існуючих методів та моделей вирішення задачі кредитного скорингу, побудова моделей вирішення задачі оцінки та прогнозування ризиків зелених проектів. Результатом роботи буде розробка моделей аналізу та оцінки ризиків зелених проектів на основі математичних методів, а також порівнянні їх ефективності. Таким чином, об'єктом дослідження даної роботи є ризики інвестування у зелені проекти, а також способи їх аналізу та оцінки. Предметом дослідження є математичні методи та моделі аналізу та прогнозування ризиків зелених проектів.

3. ОГЛЯД МЕТОДІВ ОЦІНКИ РИЗИКІВ

Одним із класичних рішень задачі оцінки кредитних ризиків є регресійний аналіз. Регресійний аналіз передбачає пошук певних взаємозв'язків між залежною змінною, яку називають результуючою або цільовою та однією чи декількома незалежними змінними – предикторами. Рівняння регресії є зображенням зв'язку між цими змінними у вигляді математичної формули, а самі зв'язки можуть бути явними або неявними. Явні зв'язки, у більшості випадків, можуть бути завчасно описані за допомогою відомих формул. Неявний зв'язок не може бути переданий у вигляді формул, тому, перш за все, цей зв'язок потрібно знайти. Суть методу полягає у побудові рівняння регресії та пошуку його коефіцієнтів. Для цього вирішується система рівнянь за допомогою метода Крамера.

Для задачі класифікації, використовується логістична регресія. Її особливість полягає у використанні сигмоїдної функції:

$$p(x_i) = \frac{1}{1 + \exp(-f(x_i))}$$

Значення $p_i(x)$ часто інтерпретують як прогнозовану ймовірність того, що результат для заданих предикторів дорівнює 1. Отже, $1 - p(x)$ – це ймовірність того, що результат дорівнює 0. У цьому випадку 0 та 1 розглядаються як два різні класи, між якими розподіляються елементи.

Ще одним популярним методом вирішення задачі кредитного скорингу є метод дерев рішень. Він є одним із найбільш вживаних методів інтелектуального аналізу через свою простоту, вільність у неоднозначності та надійність, навіть при відсутності певних вхідних значень. Древа рішень можуть працювати як із дискретними, так і з неперервними величинами як цільової змінної так і предикторів.

Древа рішень складаються із декількох компонент, основними з яких є вузли та гілки. Самі вузли також поділяються на декілька типів: корінь, внутрішні вузли, листки.

Побудова дерева рішень відбувається у декілька кроків. Найважливіші з них:

- розбиття;
- зупинка;
- відсікання.

Задачею розбиття дерева є розділення батьківських вузлів на більш чисті дочірні вузли цільової змінної. Процес побудови моделі відбувається шляхом визначення найважливішого для вузла предиктора і подальшого розбиття множини на дві або більше категорій. Вибір найбільш значущого предиктора відбувається шляхом аналізу характеристик, пов'язаних зі ступенем "чистоти" отриманих дочірніх вузлів. Ці характеристики включають ентропію, індекс Джині, помилку класифікації, інформаційний виграш та коефіцієнт виграшу [2].

Процедура розбиття триває доти, доки не будуть виконані заздалегідь визначені критерії однорідності або зупинки. У більшості випадків не всі потенційні предиктори будуть використані для побудови моделі дерева рішень, але можливі випадки, коли певна вхідна

змінна може бути використана кілька разів на різних рівнях дерева рішень. До загальних параметрів, що використовуються в правилах зупинки, належать:

- мінімальна кількість записів у листку;
- мінімальна кількість записів у вузлі до розбиття;
- глибина (тобто, кількість кроків) будь-якого листка від кореневого вузла.

У деяких ситуаціях правила зупинки не працюють належним чином. Альтернативний спосіб побудови дерева рішень полягає в тому, щоб спочатку виростити велике дерево, а потім обрізати його до оптимального розміру, видаливши вузли, які надають менше додаткової інформації [3].

Метод випадкового лісу отримав свою популярність через зручність та адаптивність, що дозволяє йому вирішувати як задачі класифікації, так і регресії [4]. Основною перевагою методу є його здатність обробляти складні набори даних та запобігати перенавчанню. Випадкових ліс відноситься до ансамблевих методів і використовує бегінг при побудові моделі. Ансамбль складається з набору індивідуально навчених моделей, які потім об'єднуються в одну кінцеву модель. Дослідження показують, що ансамбль часто є більш точним, ніж будь-який з окремих моделей в ансамблі.

Бегінг (bagging) це один із методів створення ансамблів [4]. Бегінг означає бутстреп-агрегування, яке працює на основі концепції бутстреп-вибірок. Якщо початковий навчальний набір даних має розмір N і потрібно згенерувати m індивідуальних моделей для побудови ансамблю, тоді m різних навчальних наборів, кожен з яких має розмір N , генеруються з вихідного набору даних за допомогою вибірки із заміною. Множинні класифікатори, згенеровані в пакетах (bag), є незалежними один від одного.

Випадковий ліс відноситься до ансамблевих методів, тому, у випадку задачі класифікації, модель випадкового лісу буде складатися з колекції моделей дерев рішень, кожна з яких буде віддавати свій голос за обраний нею клас, під час прогнозування. Щоб досягти різноманітності серед базових дерев рішень, використовується підхід рандомізації, який добре працює з бегінгом.

Більш сучасний спосіб вирішення задачі кредитного скорингу – це бустингові методи. Градієнтний бустинг, так само як і випадковий ліс, відноситься до ансамблевих методів. Однак, якщо випадковий ліс використовує бегінг, то градієнтний бустинг використовує бустинг. У випадку бустингу, кожній вибірці присвоюються ваги з навчального набору даних. Якщо потрібно згенерувати m моделей, то вони генеруються послідовно так, щоб за одну ітерацію генерувалась одна модель. Для генерації класифікатора c_i ваги навчальних вибірок оновлюються на основі результатів класифікації класифікатора c_{i-1} . Моделі, згенеровані у випадку використання бустингу, залежать одна від одної [5].

AdaBoost – одна з модифікацій методів градієнтного бустингу, яка часто використовується для задач класифікації. Алгоритм AdaBoost отримує на вхід певну навчальну вибірку і викликає заданий слабкий алгоритм навчання багаторазово на кожній ітерації. Однією з основних ідей алгоритму є підтримка розподілу ваг D_t на навчальній множині. Спочатку всі приклади мають однакову вагу, але з кожною ітерацією вага неправильно класифікованих прикладів збільшується, таким чином слабкий "учень" змушений зосередитися на важких прикладах навчального набору. Задача слабого "учня" полягає у пошуку слабкої гіпотези h_t , яка б підійшла для розподілу D_t . Якість слабкої гіпотези вимірюється її похибкою. Після отримання слабкої гіпотези h_t , AdaBoost обирає параметр $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$ і оновлює розподіл D_t так, щоб збільшити вагу прикладів, класифікованих неправильно, за допомогою гіпотези h_t , і зменшити вагу правильно класифікованих прикладів. Остаточна гіпотеза H є зваженою більшістю голосів з T слабких гіпотез.

4. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Для вирішення задачі кредитного скорингу зелених проектів було створено набір даних, шляхом комбінації двох окремих наборів – набору даних про кредитоспроможність компанії та набору даних про її вплив на навколишнє середовище. Таким чином, результуючий датасет складається з наступних двох:

- датасет кредитного рейтингу компаній та їх фінансових показників;
- датасет впливу діяльності компаній на навколишнє середовище.

Кредитний рейтинг компаній виражає здатність компанії погашати свої борги перед кредиторами і ґрунтується, переважно, на фінансовій звітності. Набір даних кредитного рейтингу компаній містить саме ці метрики, що вказують на економічне становище компанії.

Набір даних впливу діяльності компанії на навколишнє середовище характеризує вплив компанії на навколишнє середовище у грошовому еквіваленті.

Формування навчального набору здійснювалось на основі двох наборів даних за принципом еквівалентності назви компанії, та року, коли було проведено оцінку кредитного рейтингу та оцінку впливу на навколишній світ. У результаті було отримано датасет, що містить 847 відбитків даних, серед яких:

- 229 відбитків із рейтингом “А”;
- 534 відбитків із рейтингом “В”;
- 13 відбитків із рейтингом “С”;

У дослідженні було побудовано чотири моделі: логістичної регресії, дерев рішень, випадкового лісу та AdaBoost. Кожна з моделей будувалась для трьох різних комбінацій вхідних параметрів:

- набір даних лише із кредитним рейтингом компаній та їх фінансовими показниками;
- набір даних лише із показниками впливу компаній на навколишнє середовище;
- комбінований набір даних кредитного рейтингу та впливу компаній на навколишнє середовище.

Результати кожної моделі для кожного набору даних можна побачити у Таблицях 1-4.

Таблиця 1. Метрики моделі логістичної регресії

Фінансові навчальні дані			Дані впливу на навколишнє середовище			Комбіновані навчальні дані		
Точність	0,7164		Точність	0,6824		Точність	0,6824	
Зважений F1 score	0,6968		Зважений F1 score	0,5982		Зважений F1 score	0,6231	
F1-score			F1-score			F1-score		
“А”	“В”	“С”	“А”	“В”	“С”	“А”	“В”	“С”
0,4946	0,8049	0	0,1846	0,8029	0	0,2740	0,7970	0

Таблиця 2. Метрики моделі дерева рішень

Фінансові навчальні дані			Дані впливу на навколишнє середовище			Комбіновані навчальні дані		
Точність	0,8118		Точність	0,8059		Точність	0,8353	
Зважений F1 score	0,8118		Зважений F1 score	0,8063		Зважений F1 score	0,8387	
F1-score			F1-score			F1-score		
“А”	“В”	“С”	“А”	“В”	“С”	“А”	“В”	“С”
0,7273	0,8596	0	0,7207	0,8546	0	0,7719	0,8739	0,5

Таблиця 3. Метрики моделі випадкового лісу

Фінансові навчальні дані			Дані впливу на навколишнє середовище			Комбіновані навчальні дані		
Точність	0,8706		Точність	0,8294		Точність	0,9	
Зважений F1 score	0,8700		Зважений F1 score	0,8234		Зважений F1 score	0,8973	
F1-score			F1-score			F1-score		
“А”	“В”	“С”	“А”	“В”	“С”	“А”	“В”	“С”
0,7963	0,9043	1	0,7217	0,8797	0	0,8546	0,89258	0

Таблиця 4. Метрики моделі Ada Boost

Фінансові навчальні дані			Дані впливу на навколишнє середовище			Комбіновані навчальні дані		
Точність	0,8706		Точність	0,8177		Точність	0,9	
Зважений F1 score	0,8722		Зважений F1 score	0,8111		Зважений F1 score	0,8953	
F1-score			F1-score			F1-score		
“А”	“В”	“С”	“А”	“В”	“С”	“А”	“В”	“С”
0,8104	0,9009	1	0,7010	0,8714	0	0,8432	0,9283	0

Із наведених таблиць видно, що модель логістичної регресії показала найгірший результат із досить низькою точністю. Дана модель є єдиною, яка показала вищу точність на фінансовому наборі даних, ніж на комбінованому. Вона показала непогану точність для “В” рейтингу, але для “А” та “С” точність низька. Особливо важлива низька точність для “А” рейтингу, адже проблеми рейтингу “С” в більшій мірі пов’язані із низькою кількістю даних даного рейтингу в датасеті.

Модель дерева рішень показала значно кращі результати, ніж модель логістичної регресії незалежно від датасету. Точність визначення рейтингу “А” значно виросла на всіх наборах даних у порівнянні із регресією. Найкращі метрики модель показала на комбінованому наборі даних, навіть змогла визначити рейтинг типу “С”, але це більше випадковість. Найгірші результати моделі на наборі даних впливу на навколишнє середовище, але різниця незначна у порівнянні з фінансовим датасетом.

Результати моделей AdaBoost та випадкового лісу майже не відрізняється. Точність незначно краща у другій моделі, але це можна списати на статистичну похибку. В цілому, моделі перевершили за точністю попередні моделі регресії та дерева рішень. На наборі фінансових даних моделі можуть визначати рейтинг “С”, хоч їх частка в датасеті незначна. Найбільші показники точності та F1-score моделі показали саме на комбінованому наборі даних. Незначно гірші метрики моделі показали на фінансовому наборі даних, а навчальний набір даних впливу на навколишнє середовище виявився недостатньо інформативним, хоча точність також досить висока

5. ВИСНОВКИ

Питання оцінки ризиків зелених проектів стає все більш важливим. Кліматичні зміни та збільшення частотності стихійних лих стимулює людство переглядати стратегію поводження з природою планети. В результаті даного дослідження було розглянуто методи вирішення задачі кредитного скорингу та побудовано моделі для прогнозування ризиків зелених проектів. Було проведено порівняння створених моделей та визначено їх ефективність.

У випадку застосування моделі логістичної регресії додавання показників впливу на навколишнє середовище не дало покращень у прогнозуванні. Модель дерева рішень показала значно кращі результати, ніж модель логістичної регресії на усіх варіантах наборів даних, але вона все ще поступається ансамблевим моделям. Результати моделей AdaBoost та випадкового лісу майже не відрізняється. Точність незначно краща у другій моделі, але відмінністю фактично можна знехтувати.

Застосування екологічних метрик при оцінці перспективності та ефективності нових зелених проєктів, безумовно, буде важливим завданням для відбудови України. Вже зараз актуально впроваджувати зелені проєкти, накопичувати дані стосовно їх ефективності та ощадності, стимулювати збільшення таких проєктів шляхом інвестування і видачі вигідних позик та кредитних ліній. Тому, розпочате нами дослідження з метою оцінки можливості застосування класичних підходів та методів, а також розробки нових математичних моделей для оцінювання і врахування як фінансових так і екологічних показників при оцінювання зелених проєктів, безумовно, є початком і перспективних напрямом для подальших досліджень.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Сучасні екологічно чисті технології: Курс лекцій [Електронний ресурс] : навч. посіб. для здобувачів ступеня доктора філософії спеціальності 161 «Хімічні технології та інженерія» / КПІ ім. Ігоря Сікорського ; уклад.: В.М. Павленко, В.Ю. Тобілко, А.І. Бондарева. – Електронні текстові дані (1 файл: 0,945 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2021. – 78 с.
2. Patel N, Upadhyay S. Study of various decision tree pruning methods with their empirical comparison in WEKA. *Int J Comp Appl.* 60(12):20–25.
3. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning.* Springer; 2001. p. 269-272
4. Кузнєцова Н.В., Бідюк П.І. Теорія і практика аналізу фінансових ризиків: системний підхід, монографія. Київ 2020, 400 с.
5. Vrushali Y Kulkarni, Pradeep K Sinha. Random Forest Classifiers. *A Survey and Future Research Directions, International Journal of Advanced Computing, ISSN:2051-0845, Vol.36, Issue.1.* p.1144-1153
6. Euro 7 emission standards, *Insigh: from inferiup international limited*, 18 July 2023, P.8.
7. Євтух О. Типові ризики іпотечного капіталу та управління ними // *Вісник НБУ.* – № 11. – с. 43-46.
8. Грушко В.І., Пилипченко О.І., Пікус Р.В. *Управління фінансовими ризиками.* – Київ: Інститут економіки і права “Крок”, 2000. – с. 24.
9. Яворський Р. Розвиток банківської системи в Україні: Матеріали досліджень переможців всеукраїнського конкурсу “Економічні реформи в Україні”. – Київ, 1999. – с. 107.
10. Чайковський Я. Удосконалення методики комплексної оцінки кредитоспроможності позичальників // *Вісник НБУ.* – 2003. – № 11. – с. 30- 34.
11. Berry MJA, Linoff G. *Mastering Data Mining: The Art and Science of Customer Relationship Management.* New York: John Wiley & Sons, Inc; 1999.
12. Zibran MF. *Department of Computer Science. Diagnostic and Statistical Manual of Mental Disorders – Fourth Edition.* Alberta, Canada: Department of Computer Science, University of Calgary; 2012.
13. Freund Y., Schapire R. E. A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771-780, September, 1999. 14 p.

АНАЛІЗ ВІДКРИТИХ ДАНИХ ПРО ЯКІСТЬ ПОВІТРЯ В МІСЬКОМУ СЕРЕДОВИЩІ ТА РОЗРОБКА ПРОГНОСТИЧНИХ МОДЕЛЕЙ ДЛЯ ПРОГНОЗУ ЗАБРУДНЕННЯ ПОВІТРЯ В МІСТІ

Луцкер Р.О., Гуськова В.Г.

Національний технічний університет України «Київський політехнічний інститут
ім. Ігоря Сікорського»

1. ВСТУП

У сучасному світі існує тенденція активного зростання урбанізації, призводячи до того, що понад 50% населення проживає великих містах. Великі міста, подібно великим пічам, викидають значні обсяги тепла, речовин і відходів у повітря, що призводить до забруднення атмосфери. Проблеми з якістю повітря, особливо гострі у місті Запоріжжя, індустріальному центрі України, стають серйозним викликом для здоров'я людей, оскільки токсичні гази негативно впливають на їхній організм, а індустріальні викиди обтяжують навколишнє середовище.

Для розв'язання проблем забруднення повітря необхідно впроваджувати комплексні заходи. Це включає розподіл міста на рівномірні райони зі станціями моніторингу якості повітря в реальному часі, хімічний аналіз повітря у різних районах та встановлення очисних споруд на критичних об'єктах. Забруднення може бути спричинене різними факторами, тому важливо проводити аналіз через різні організації, зосереджуючись на створенні комплексних систем моніторингу та прогнозування.

Наступним кроком є розробка систем прогнозування та оповіщення на основі актуальних даних про атмосферу. Для цього необхідні виміри, методи та моделі, включаючи метеорологічні та хімічні, забезпечуючи більш точне уявлення про розподіли динамічних та хімічних компонентів. Прогнозування повинно враховувати рівень хімічних елементів у минулому, використовуючи методи, такі як нейронні мережі та комбінування різних вхідних даних для досягнення максимально точного прогнозу рівня забруднення атмосфери.

2. ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

Мета роботи полягає в проведенні досліджень та розробці методів прогнозування забруднення атмосферного повітря з використанням програмної системи. Зазначена система має на меті створення моделі прогнозування за певним методом, надання двотижневого прогнозу та візуалізацію результатів за допомогою веб-інструментів. Для досягнення цієї мети перед розробником стоять завдання, такі як аналіз аналогів програмного забезпечення для визначення вимог, дослідження методів прогнозування часових рядів, розробка технічного завдання, вибір інструментів розробки, проектування структури програмного забезпечення, розробка самого ПЗ, його тестування, відлагодження та економічне обґрунтування розробки.

3. ОГЛЯД КЛАСИЧНИХ МЕТОДІВ ПРОГНОЗУВАННЯ ЯКОСТІ ПОВІТРЯ

Емпіричні моделі використовують прості алгебраїчні відношення для визначення різних величин, таких як концентрація забруднення. Наприклад, методика ОНД-86 використовує цей підхід для прогнозування рівнів забруднення повітря в Україні. Емпіричний

метод модової декомпозиції розкладає часовий ряд на прості складові, відомі як власні характерні форми.

Хоча безпосереднє використання емпіричних моделей є простим підходом, вони не враховують фізичний зміст і не забезпечують узагальнення при зміні вхідних даних.

Статистичні моделі є прогнозними моделями, які будуються на основі аналізу результатів спостережень за допомогою статистичних методів. Для їхньої побудови важлива значна кількість вимірювань, що вимагає витрат часу на отримання та оброблення даних.

У групі статистичних моделей виділяють регресійні та авторегресійні моделі. Регресійні моделі можуть бути простою лінійною, багатовимірною або нелінійною. Серед авторегресійних моделей найвідомішою є ARIMA (інтегрована модель авторегресії рухомого середнього).

Лінійна регресія дозволяє знаходити одне прогнозоване значення за моделлю, тоді як авторегресійна модель використовує результати попередніх спостережень як вхідні дані для рівняння регресії на наступних кроках. Загалом, в контексті прогнозування забруднення атмосферного повітря, такі моделі призначені для короткострокових прогнозів на декілька годин або діб.

Нейронні моделі складаються з нейронних мереж, які представляють собою розділ штучного інтелекту. Використання інструментів, аналогічних явищам у живих організмах, є основним принципом нейронних мереж. Важливою особливістю цих мереж є їхня здатність до паралельної обробки інформації всіма ланками, що значно прискорює процес обробки даних. При великому числі міжнейронних з'єднань мережа також виявляє стійкість до помилок, які можуть виникнути на деяких лініях.

Ще однією важливою характеристикою є здатність нейронних мереж до навчання та узагальнення накопичених знань. Мережа, яка пройшла тренування на обмеженій множині даних, може ефективно узагальнювати інформацію та демонструвати добрі результати на нових даних, які не використовувалися під час навчання.

3. ОГЛЯД МОДЕЛІ LSTM

Long Short Term Memory (LSTM) – це різновид рекурентної нейронної мережі. У RNN вихід з останнього кроку подається як вхід на поточний крок. LSTM була розроблена Hochreiter & Schmidhuber. Вона вирішила проблему довгострокових залежностей ШНМ, коли ШНМ не може передбачити слово, що зберігається в довгостроковій пам'яті, але може давати більш точні прогнози на основі нещодавньої інформації. Зі збільшенням довжини проміжку RNN не дає ефективної роботи. LSTM за замовчуванням може зберігати інформацію протягом тривалого періоду часу. Він використовується для обробки, прогнозування та класифікації на основі даних часових рядів.

Довготривала короткочасна пам'ять (LSTM) – це тип рекурентної нейронної мережі (RNN), яка спеціально розроблена для обробки послідовних даних, таких як часові ряди, мова і текст. LSTM-мережі здатні вивчати довгострокові залежності в послідовних даних, що робить їх добре придатними для таких завдань, як переклад мови, розпізнавання мови та прогнозування часових рядів.

Традиційний ШНМ має один прихований стан, який проходить через час, що може ускладнити для мережі навчання довгострокових залежностей. LSTM вирішують цю проблему шляхом введення комірки пам'яті, яка є контейнером, що може зберігати інформацію протягом тривалого періоду часу. Коміркою пам'яті керують три вентиля: вхідний вентиль, вентиль забування та вихідний вентиль. Ці вентиля вирішують, яку інформацію додавати в комірку пам'яті, видаляти з неї та виводити з неї.

Вхідний клапан контролює, яка інформація додається до комірки пам'яті. Клапан забування контролює, яка інформація видаляється з комірки пам'яті. А вихідний клапан контролює, яка інформація виводиться з комірки пам'яті. Це дозволяє LSTM-мережам вибірково зберігати або відкидати інформацію, коли вона проходить через мережу, що дозволяє їм вивчати довгострокові залежності.

LSTM має ланцюгову структуру, яка містить чотири нейронні мережі та різні блоки пам'яті, які називаються комірками.

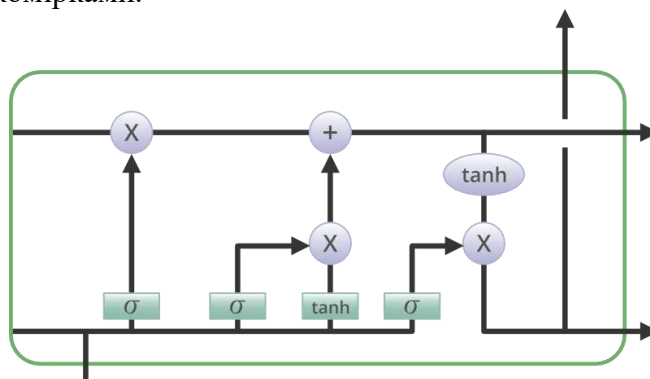


Рисунок 1. Структура нейронної мережі LSTM

Інформація зберігається в комірках, а маніпуляції з пам'яттю здійснюються за допомогою клапанів. Існує три види клапанів:

Клапан забування: Цей механізм дозволяє вилучити інформацію, яка вже не є корисною для стану комірки. Два входи, x_t та h_{t-1} , піддаються обробці клапаном, де їхні вагові матриці та зсуви визначають, яка частина інформації буде збережена або видалена.

Вхідний клапан: Вхідний клапан відповідає за додавання корисної інформації до стану комірки. Інформація регулюється сигмоїдною функцією, яка фільтрує значення, що повинні бути запам'ятовані. Функція \tanh генерує вектор, який містить всі можливі значення з попереднього вихідного стану та поточного входу.

Вихідний клапан: Вихідний клапан відповідає за вилучення корисної інформації з поточного стану комірки для представлення на виході. Генерується вектор за допомогою функції тангенса, а потім сигмоїдна функція фільтрує та визначає, які значення слід зберегти чи видалити, використовуючи входи h_{t-1} та x_t .

4. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Представлені результати використовують концентрацію частинок PM2.5, проте важливо відзначити, що аналогічні тенденції та результати можна спостерігати і для інших показників якості повітря. Ці дані служать лише експериментальним прикладом та ілюстрацією, але схожі закономірності відслідковуються і для інших забруднюючих речовин (табл. 1, рис. 2).

Таблиця 1. Результати прогнозування

Метод прогнозування	RMSE	MAE	Точність, %
Модель ARIMA	15,33	10,95	86,35
Штучні нейронні мережі	19,10	12,87	83,96
Рекурентні нейронні мережі	14,95	10,63	86,75
Нейронні мережі типу «Довга короткострокова пам'ять»	13,87	9,13	88,62

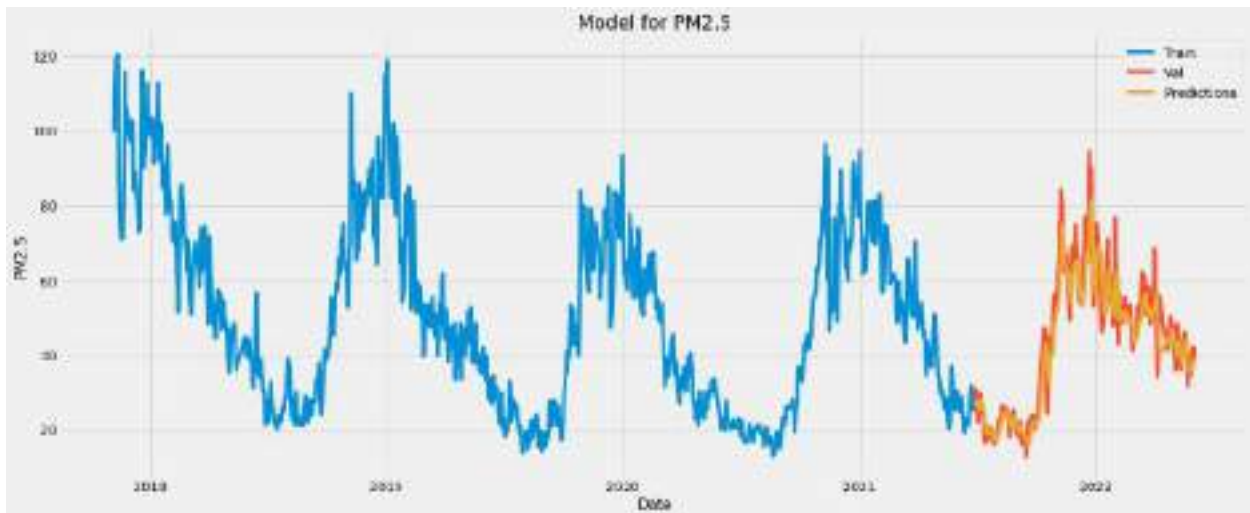


Рисунок 2. Результати прогнозування

Метод "Довга короткострокова пам'ять" (LSTM) виявився найефективнішим у порівнянні з іншими методами прогнозування, демонструючи найменше значення середньої квадратичної помилки (RMSE) та середньої абсолютної помилки (MAE), а також досягнувши найвищу точність прогнозування на рівні 88,62%. Рекурентні нейронні мережі та модель ARIMA також продемонстрували добрі результати, але в усіх трьох аспектах LSTM перевершила їх. Застосування штучних нейронних мереж виявилось менш ефективним у порівнянні з іншими методами, враховуючи значення RMSE, MAE та точність. Отже, отримані результати свідчать про високу ефективність методу LSTM у прогнозуванні якості повітря, роблячи його переважним вибором серед розглянутих методів. Нижче представлений графік прогнозування, отриманого за допомогою методу LSTM.

5. ВИСНОВКИ

У дослідженні порівняли різні методи прогнозування якості повітря за концентрацією частинок. Метод "Довга короткострокова пам'ять" (LSTM) виявився найефективнішим з найменшими помилками та точністю 88,62%. Рекурентні нейронні мережі та ARIMA показали гарні результати, але LSTM виявився кращим. Штучні нейронні мережі були менш ефективними. Узагальнено: LSTM – оптимальний вибір для прогнозування якості повітря з високою точністю. Графік прогнозування LSTM ілюструє його переваги. Це дослідження важливе для сфери моніторингу повітря.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Біляєв, М.М. Моделювання і прогнозування стану довкілля: підручник для студентів вищих навчальних закладів [Текст] / М. М. Біляєв, В.В. Біляєва, П.С Кіріченко. – Кривий Ріг : Вид. Р.А. Козлов, 2016. – 207 с.
2. Applications of Hilbert-Huang transform to non-stationary financial time series analysis [Text] / N. E. Huang, M. Wu, W. Qu, S. R. Long, S. S. P. Shen. // Applied Stochastic Models in Business and Industry. – 2003. – Vol. 19, № 3. – Pp. 245–268.
3. Moustiris K.P., Nastos P.T., Larissi I.K., Paliatsos A.G. Application of Multiple Linear Regression Models and Artificial Neural Networks on the Surface Ozone Forecast in the Greater Athens Area, Greece [Електронний ресурс] / K.P. Moustiris., P.T. Nastos, I.K. Larissi, A.G. Paliatsos – Режим доступу: <https://www.hindawi.com/journals/amete/2012/894714/>

ПОРІВНЯЛЬНИЙ АНАЛІЗ МОДЕЛЕЙ ДЛЯ МЕТОДІВ ПРОГНОЗУВАННЯ

Макухін Є.І.¹, Макаренко О.С.², Бідюк П.І.

Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського»

¹ makukhin.yevhen@lil.kpi.ua, ² makalex51@gmail.com

Прогнозування є важливим елементом сучасного світу, де швидкі зміни і нестабільні умови вимагають точних та ефективних стратегій планування. Особливо це стосується прогнозування фінансових ринків. Створення якісної моделі є ключовою складовою якісного прогнозування. Метою роботи є порівняння різних моделей для методів прогнозування. Результатом роботи є програмне забезпечення що виконує прогнозування ціни акція і дозволяє порівняти результати різних моделей. У роботі використано теоретичні та емпіричні методи дослідження.

Ключові слова: моделі прогнозування, методи прогнозування, ARIMA, LSTM, N-BEATS, аналіз даних.

1. ВСТУП

На сьогоднішній день фондовий ринок є незамінним інструментом, що дозволяє фінансовим аналітикам, інвесторам та трейдерам інвестувати в цінні папери та оцінювати вартість активів. У той же час, для більш ефективних інвестицій користувачі повинні мати якісні прогнози. Для цього постає задача знаходження та використання методів та моделей для аналізу та побудови точних прогнозів динаміки ціноутворення об'єктів фондового ринку.

Знаходження відповідних моделей дозволяє зробити точний прогноз і допомогти в прийнятті рішення щодо інвестицій. Використання різних моделей дозволить знайти метод який найкраще підходить для відповідних даних. При цьому важливо враховувати усі особливості та властивості прогнозованих даних.

Для порівняння прогнозів було розглянуто математичні методи і machine learning моделі, що застосовують для вирішення практичних задач аналізу та прогнозування нестационарних процесів, а саме: модель авторегресії інтегрованого ковзного середнього ARIMA, LSTM, N-BEATS та ансамблеву модель.

2. МЕТОДИ ПРОГНОЗУВАННЯ

Метод ARIMA (Autoregressive Integrated Moving Average) є популярним і ефективним інструментом для прогнозування часових рядів. ARIMA включає в себе три основні складові: авторегресію (AR), інтегрування (I) та ковзне середнє (MA). ARIMA розшифровується як Авторегресійна інтегрована модель ковзного середнього. Вона належить до класу моделей, які пояснюють заданий часовий ряд на основі його власних минулих значень, тобто власних лагів і помилок прогнозування. Рівняння можна використовувати для прогнозування майбутніх значень. Будь-який "несезонний" часовий ряд, який демонструє закономірності і не є випадковим білим шумом, може бути змодельований за допомогою моделей ARIMA. ARIMA моделі задаються трьома параметрами порядку: (p, d, q), де

- p – порядок AR-члена;
- q – порядок члена MA;

- d – кількість диференціювань, необхідних для того, щоб зробити часовий ряд стаціонарним.

Авторегресія $AR(p)$ – регресійна модель, яка використовує залежний зв'язок між поточним спостереженням і спостереженнями за попередній період. Авторегресійний компонент ($AR(p)$) означає використання минулих значень у рівнянні регресії для часового ряду. $I(d)$ Інтегрування – використовує диференціювання спостережень (віднімання спостереження від спостереження на попередньому часовому кроці) для того, щоб зробити часовий ряд стаціонарним. Диференціювання передбачає віднімання поточних значень ряду від його попередніх значень d разів. $MA(q)$ Moving Average – модель, яка використовує залежність між спостереженням і залишковою похибкою від моделі ковзного середнього, застосованої до запізнених спостережень. Компонент ковзного середнього відображає похибку моделі як комбінацію попередніх членів похибки. Порядок q показує кількість членів, які включаються в модель. В даній роботі буде використаний метод `auto_arima` з бібліотеки `pmdarima` мови програмування Python для автоматизації процесу прогнозування ARIMA. Метод `auto_arima()` використовує покроковий підхід для перебору декількох комбінацій параметрів p, d, q і вибирає найкращу модель, яка має найменший AIC (інформаційний критерій Акаїке). Вона працює шляхом проведення тестів на диференціювання (наприклад, Квятковського-Філіпса-Шмідта-Шина, розширеного Дікі-Фуллера або Філіпса-Перрона) для визначення порядку диференціювання, d , а потім підбирає моделі в межах визначених діапазонів `start_p`, `max_p`, `start_q`, `max_q`. Якщо увімкнено опцію сезонності, `auto_arima` також намагається визначити оптимальні гіперпараметри P та Q після проведення тесту Канови-Хансена для визначення оптимального порядку сезонного диференціювання, D .

Мережі з довготривалою короткочасною пам'яттю (ДКЧП, англ. long short-term memory, LSTM), – це особливий тип рекурентних нейронних мереж, що здатний навчатися довготривалих залежностей. Вони були введені Хохрайтером і Шмідгубером (1997), і були вдосконалені і популяризовані багатьма людьми в наступних роботах. Вони надзвичайно добре працюють над широким спектром проблем і зараз широко використовуються. Як і більшість РНМ, мережа LSTM є універсальною в тому сенсі, що за достатньої кількості вузлів мережі вона може обчислювати будь-що, що може обчислювати звичайний комп'ютер, за умови, що вона має належну матрицю вагових коефіцієнтів, що може розглядатися як її програма. На відміну від традиційних РНМ, мережа LSTM добре підходить для навчання з досвіду з метою класифікації, обробки або передбачення часових рядів в умовах, коли між важливими подіями існують часові затримки невідомої тривалості. LSTM явно розроблені для того, щоб уникнути проблеми довготривалої залежності.

Блоки LSTM містять три або чотири «вентилі» (англ. gates), які вони використовують для керування плином інформації до або з їхньої пам'яті. Звичайний блок LSTM складається з комірки, вхідних вентилів (Input Gate), вихідних вентилів (Output Gate) і вентилів забування (Forget Gate). Комірка запам'ятовує значення протягом довільних часових інтервалів, а три вентилі регулюють потік інформації в комірку і з неї. Вентилі забування вирішують, яку інформацію відкинути з попереднього стану, присвоюючи попередньому стану, порівняно з поточним входом, значення від 0 до 1. (Округлене) значення 1 означає збереження інформації, а значення 0 – її відкидання. Вхідні вентилі вирішують, які частини нової інформації зберігати у поточному стані, використовуючи ту ж систему, що і вентилі забування. Вихідні вентилі керують тим, яку інформацію з поточного стану виводити, присвоюючи інформації значення від 0 до 1, враховуючи попередній і поточний стан. Вибірковий вивід релевантної інформації з поточного стану дозволяє мережі LSTM підтримувати корисні, довгострокові залежності для прогнозування, як в поточному, так і в майбутньому часі.

Основна ідея застосування LSTM полягає в здатності мережі "вирішувати", яку інформацію тримати та яку забути, що дозволяє їй ефективно моделювати довгострокові залежності. Це здатність особливо корисна в задачах прогнозування часових рядів, обробці природної мови та інших завданнях, де важлива контекстуальна інформація. У Python для реалізації LSTM можна використовувати бібліотеки, такі як TensorFlow чи PyTorch.

N-BEATS (Neural Basis Expansion Analysis and Transformation System) - це архітектура нейронних мереж для задач прогнозування часових рядів, яка була представлена в статті "N-BEATS: Neural Basis Expansion Analysis and Transformation System" в 2020 році. Ця архітектура відрізняється своєю гнучкістю та здатністю адаптуватися до різноманітних завдань прогнозування. Ключові особливості N-BEATS:

- Підтримка декількох часових рядів: N-BEATS можна навчати на декількох часових рядах, кожен з яких представляє різний розподіл.
- Швидке навчання: Модель не містить жодних рекурентних шарів або шарів самоуваги – таким чином, швидше навчання та стабільний градієнтний потік.
- Багатогоризонтне прогнозування: Модель виробляє багатокрокові прогнози.
- Інтерпретованість: Автори розробили 2 версії моделі: загальну та інтерпретовану. Інтерпретована версія може виводити інтерпретовані прогнози щодо тренду та сезонності (рис. 1).

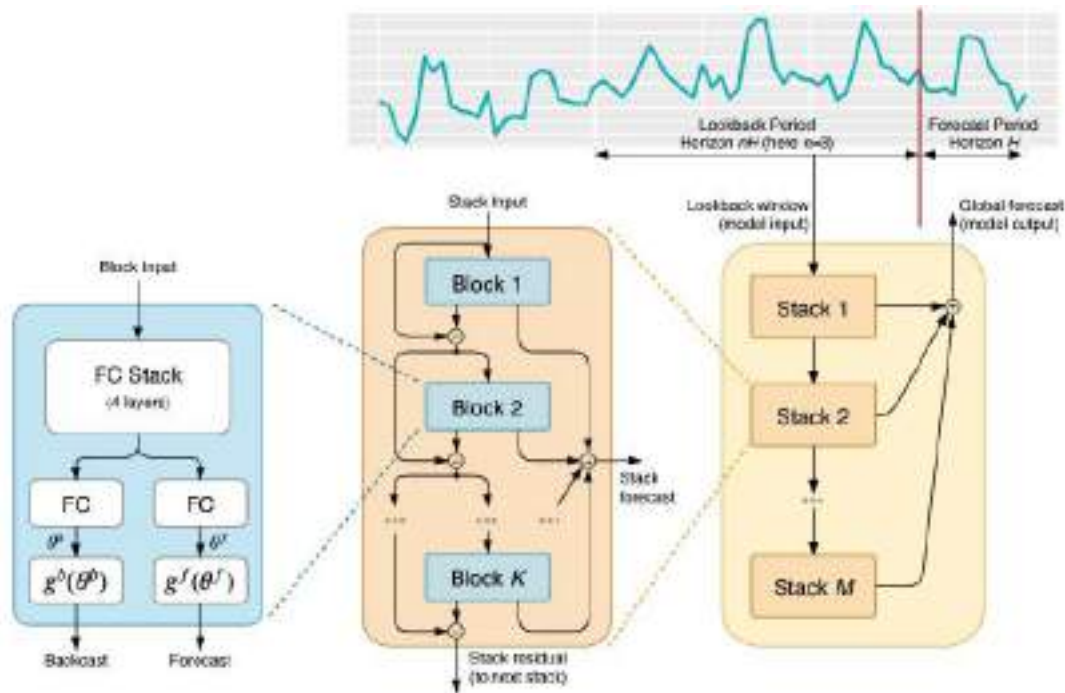


Рисунок 1. Запропонована архітектура моделі

Основні елементи:

- Блок (синій колір) – основна одиниця обробки.
- Стек (помаранчевий колір) – колекція блоків.
- Кінцева модель (жовтий колір) – сукупність стеків.

Ансамблеві методи використовують кілька алгоритмів навчання, щоб отримати кращу прогностичну ефективність, ніж можна було б отримати від будь-якого з складових алгоритмів навчання окремо. Для створення ансамблевих моделей буде використано комбінацію:

- Різних функцій втрат (MAE, MSE, MAPE)
- Випадково ініціалізованих моделей

Буде створено набір різних моделей, які намагатимуться моделювати одні й ті самі дані та функції для списку різних моделей, навчених з різними функціями втрат. Кожен шар в ансамблі моделей буде ініціалізовано випадковим нормальним (гаусівським) розподілом за допомогою Не нормальної ініціалізації, це допоможе оцінити інтервали прогнозування пізніше.

3. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Порівняльний аналіз моделей прогнозування буде здійснюватися на прикладі прогнозування вартості акцій компанії Apple.

При побудові ARIMA моделі потрібно вибрати параметри p, q, d . Для цього ми використовуємо `auto_arima` функцію з бібліотеки `rmadarma`, щоб дозволяє отримати найкращі параметри, навіть не будуючи графіки АКФ та ЧАКФ.

Після розділення вибірки на тестову та навчальну у співвідношенні 90 на 10 знайдемо оптимальну модель (рис. 2).

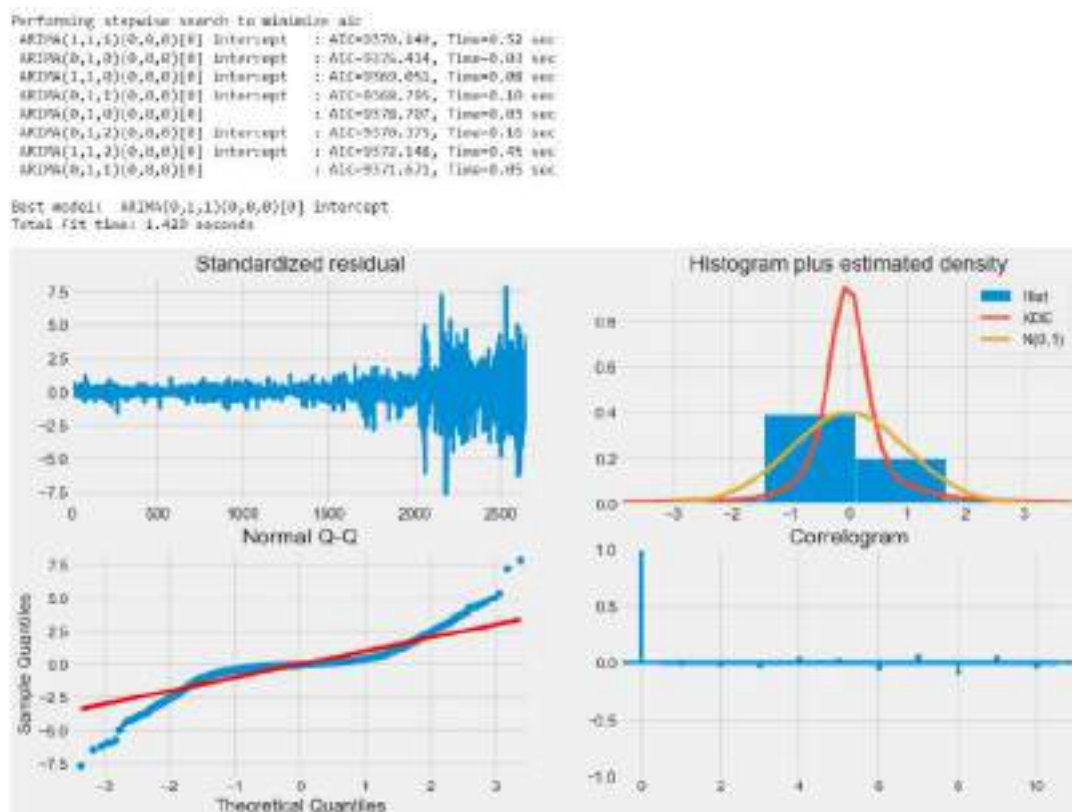


Рисунок 2. Результати пошуку оптимальної моделі

Оптимальною моделлю виявилася модель ARIMA(0,1,1). Далі використаємо модель для прогнозування значень з 95% довірчим інтервалом і оцінимо модель (рис. 3). Для оцінки знайдемо значення MAE (середня абсолютна похибка), RMSE (середньоквадратична похибка), MAPE (середня абсолютна відсоткова похибка) і MASE (середня абсолютна масштабована похибка).



Рисунок 3. Результати прогнозу значень з довірчим інтервалом

Було отримано критерії якості прогнозу, що дорівнюють:

$MAE = 15,124319$, $RMSE = 18,654106$, $MAPE = 10,100647$, $MASE = 7,449288$.

Як можна помітити за результатами, модель має не досить точний прогноз.

Далі переходимо до моделей нейромереж. Спочатку розділимо часовий ряд за допомогою вікон. Вікно – це метод перетворення набору даних часового ряду в керовану навчальну задачу. Іншими словами, ми хочемо використовувати вікна минулого для прогнозування майбутнього. Будемо використовувати розмір горизонту 1 і розмір вікна 7. Потім перетворимо наші вікна на навчальні та тестові розбиття. Замість того щоб розбити на вікна існуючі навчальні та тестові вибірки, враховуючи природу розбиття на вікна (розбиття на вікна часто вимагає зсуву в певній точці даних), зазвичай краще спочатку розбити дані на вікна, а потім розділити їх на навчальні та тестові вибірки. В результаті для кожної з моделей маємо розділення вибірки на навчальну та тестову у співвідношенні 80 на 20.

Модель LSTM має функцію активації ReLu, 100 епох та розмір партії 128. За допомогою порівняння середньої абсолютної похибки знаходимо найкращу модель і прогнозуємо значення (рис. 4).

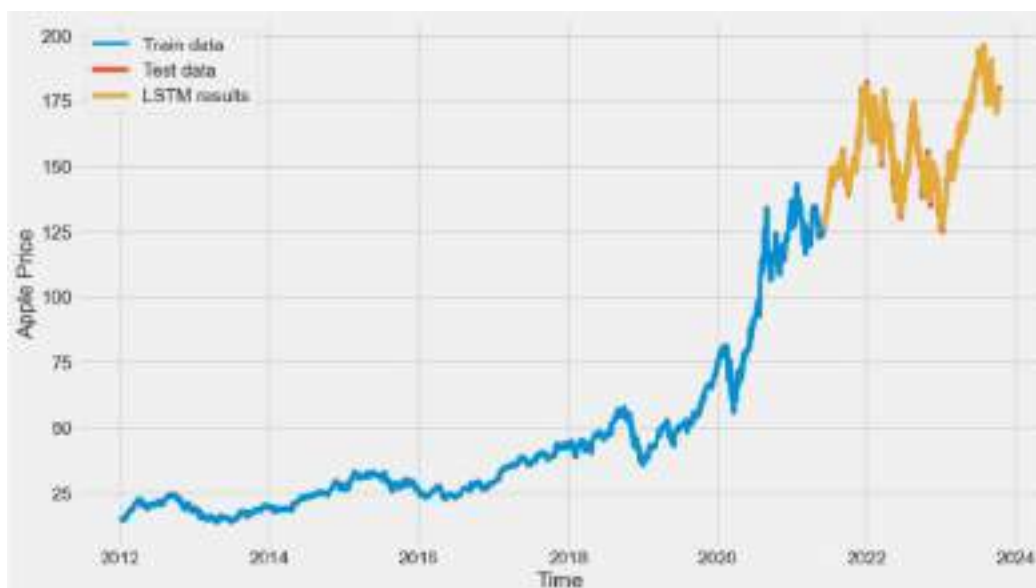


Рисунок 4. Результати прогнозу значень для моделі LSTM

Було отримано критерії якості прогнозу, що дорівнюють:

MAE = 2,206478, RMSE = 2,8834975, MAPE = 1,410206, MASE = 1,0534115

За результатами можна помітити що модель досить добре прогнозує значення ціни акцій.

Далі будемо оцінювати модель N-BEATS в якій для покращення продуктивності моделі збільшена кількість шарів у порівнянні з іншими аналогічними моделями. Почнемо з побудови блочного шару N-BEATS. Створимо основний будівельний блок для архітектури N-BEATS. Цього разу, оскільки ми будемо використовувати більшу архітектуру моделі, щоб забезпечити максимально швидке навчання моделі, ми налаштуємо наші набори даних за допомогою API `tf.data`. Налаштування гіперпараметрів для алгоритму N-BEATS візьмемо зі схеми на рисунку 5, що описує гіперпараметри, які використовуються для різних варіантів N-BEATS. Ми використовуємо N-BEATS-G, що означає загальну версію N-BEATS.

Table 18: Settings of hyperparameters across subsets of M4, M3, TOURISM datasets.

Parameter	M4						M3				TOURISM		
	Yly	Qly	Mly	Wly	Dly	Hly	Yly	Qly	Mly	Other	Yly	Qly	Mly
N-BEATS-I													
L_H	1.5	1.5	1.5	10	10	10	20	5	5	20	20	10	20
Iterations	15K	15K	15K	5K	5K	5K	50	6K	6K	250	30	500	300
Losses	sMAPE/MAPE/MASE						sMAPE/MAPE/MASE				MAPE		
S-width							2048						
S-blocks							3						
S-block-layers							4						
T-width							256						
T-degree							2						
T-blocks							3						
T-block-layers							4						
Sharing	STACK LEVEL												
Lookback period	2H, 3H, 4H, 5H, 6H, 7H												
Batch	1024												
N-BEATS-G													
L_H	1.5	1.5	1.5	10	10	10	20	20	20	10	5	10	20
Iterations	15K	15K	15K	5K	5K	5K	20	250	10K	250	30	100	100
Losses	sMAPE/MAPE/MASE						sMAPE/MAPE/MASE				MAPE		
Width							512						
Blocks							1						
Block-layers							4						
Stacks							50						
Sharing	NO												
Lookback period	2H, 3H, 4H, 5H, 6H, 7H												
Batch	1024												

Рисунок 5. Рекомендовані налаштування гіперпараметрів моделі

Після налаштування гіперпараметрів, тепер, перш ніж ми створимо модель N-BEATS, нам потрібно пройти через два шари, які відіграють велику роль в архітектурі. Саме вони роблять можливим подвійне залишкове укладання N-BEATS:

- `tf.keras.layers.subtract(inputs)` – віднімає список вхідних тензорів один від одного
- `tf.keras.layers.add(inputs)` – додає список вхідних тензорів один до одного

Залишковий зв'язок (також званий "пропускним зв'язком") передбачає, що більш глибокий шар нейронної мережі отримує виходи, а також входи більш поверхневого шару нейронної мережі. У випадку N-BEATS архітектури використовуються залишкові зв'язки, які:

- віднімають зворотні виходи попереднього блоку від зворотних входів поточного блоку
- додають прогнозні виходи з усіх блоків разом у стек

В результаті для створення і навчання моделі робимо наступне:

1. Створимо екземпляр шару блоків N-BEATS за допомогою `NBeatsBlock` (це буде початковий блок для мережі, решта будуть створені як частина стеків)
2. Створимо вхідний шар для стека N-BEATS (для цього ми будемо використовувати `Keras Functional API`)
3. Зробіть початкові беккаст та прогноз для моделі з шаром, створеним у (1)

4. Створимо за допомогою циклу стеки блокових шарів
 5. Використаємо клас NBeatsBlock у циклі, створеному в пункті (4), для створення блоків, які повертають беккасти та прогнози на рівні блоків
 6. Створюємо подвійний залишковий стек, використовуючи віднімання та додавання шарів
 7. Об'єднаємо входи та виходи моделі за допомогою `tf.keras.Model()`
 8. Скопіюємо модель з оцінкою MAE та оптимізатором Adam
 9. Підготуємо модель N-BEATS для 5000 епох, і оскільки вона має таку кількість епох, ми використаємо декілька зворотних викликів:
 - `tf.keras.callbacks.EarlyStopping()` – зупиняє навчання моделі, якщо вона не покращує валідаційні втрати за 200 епох, і відновлює найкращі ваги, використовуючи `restore_best_weights=True`
 - `tf.keras.callbacks.ReduceLROnPlateau()` – якщо втрата валідації моделі не покращується протягом 100 епох, зменшити швидкість навчання у 10 разів, щоб спробувати допомогти їй зробити поступові покращення (чим менша швидкість навчання, тим менші оновлення намагається зробити модель)
- Таким чином, для цієї моделі маємо такі результати прогнозування (рис. 6).

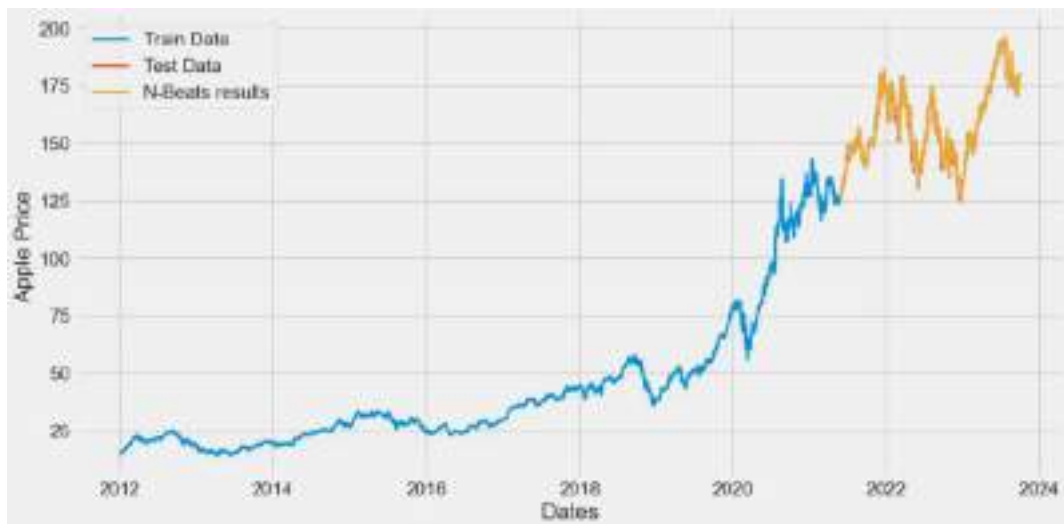


Рисунок 6. Результати прогнозу значень для моделі N-BEATS

Було отримано критерії якості прогнозу, що дорівнюють:

MAE = 2,1958368, RMSE = 2,8769698, MAPE = 1,4034737, MASE = 1,0483311

За результатами можна помітити що якість прогнозу трохи покращилась, але не суттєво.

Наступний експеримент - створення ансамблю моделей. Ансамбль передбачає навчання та об'єднання декількох різних моделей.

Для створення наших ансамблевих моделей будемо використовувати комбінацію:

- Різних функцій втрат (MAE, MSE, MAPE)
- Випадково ініціалізованих моделей

По суті, ми створимо набір різних моделей, які намагатимуться моделювати одні й ті самі дані. Створимо функції для створення списку різних моделей, навчених з різними функціями втрат. Кожен шар в ансамблі моделей буде ініціалізовано випадковим нормальним (гаусівським) розподілом за допомогою Не нормальної ініціалізації, це допоможе оцінити інтервали прогнозування пізніше. Перевіримо для 5 ітерацій та 1000 епох. Це дасть нам 15 моделей (по 5 для кожної функції втрат). Однак, оскільки ми навчаємо 15 моделей, то отримаємо 15 наборів прогнозів. Замість того, щоб порівнювати кожен набір прогнозів з істиною в останній інстанції, візьмемо медіану.

Отже, маємо такі результати прогнозування:

MAE = 2,1159441, RMSE = 2,8000813, MAPE = 1,3533982, MASE = 1,0101889

Маємо графік прогнозу з довірчим інтервалом на рисунку 7.

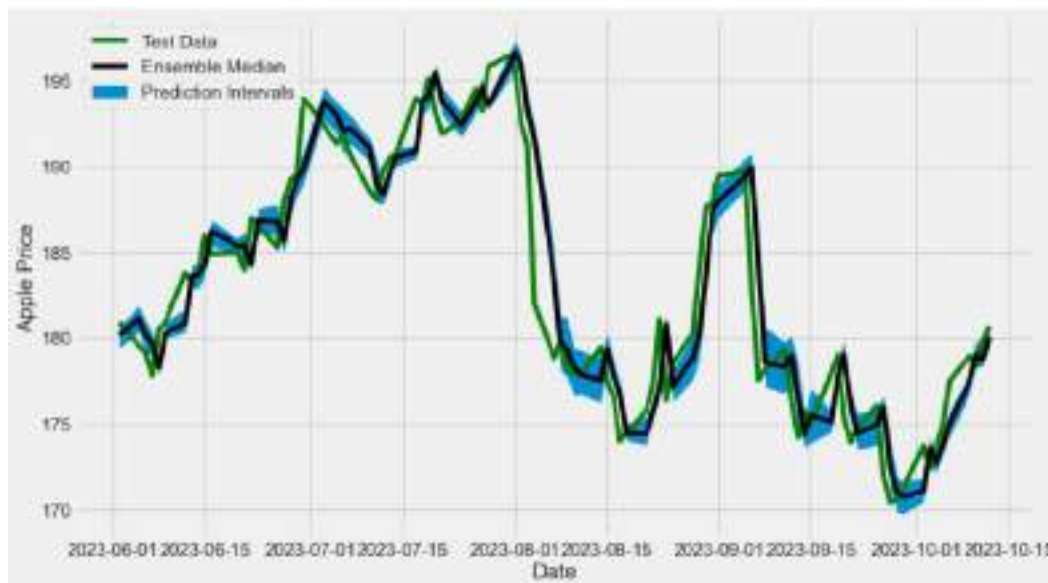


Рисунок 7. Результати прогнозу для ансамблю моделей

Тепер порівняємо результати кожної з моделей на рисунку 8.

	mae	mse	rmse	mape	mase
model_1_arima	15.124319	347.975677	18.654106	10.100647	7.449288
model_2_LSTM	2.206478	8.314558	2.883497	1.410206	1.053411
model_3_NBEATs	2.195837	8.276956	2.876970	1.403474	1.048331
model_4_ensemble	2.115944	7.840455	2.800081	1.353398	1.010189

Рисунок 8. Критерії якості прогнозу для кожної з моделей

Можна побачити, що найкращі результати має ансамблева моделей, хоча результати LSTM та N-BEATS майже аналогічні і усі 3 моделі мають достатньо високу якість прогнозування.

4. ВИСНОВКИ

На сьогоднішній день фінансові ринки стають все більш доступними для користувачів, тому виникає необхідність знаходження більш ефективних інструментів аналізу та прогнозування для подальшого інвестування.

Було проведено порівняльний аналіз методів прогнозування для нестационарних і нелінійних процесів фондових ринків та зроблено огляд моделей прогнозування, досліджено їх результати.

Було розглянуто математичні методи і machine learning моделі, що застосовують для вирішення практичних задач аналізу та прогнозування нестационарних процесів, а саме: модель авторегресії інтегрованого ковзного середнього ARIMA, LSTM, N-BEATS та ансамблеву модель.

Для порівняння моделей була реалізована програма що дозволяє побудову структури даних моделей, знаходження прогнозу та його результатів. У якості об'єкту експериментальних досліджень було обрано акції компанії Apple, та найкращі результати показала ансамблева модель машинного навчання, яка дозволила отримати дуже якісні прогнози. Інші моделі показали схожі результати, тому прогнози нейромереж можна вважати достатньо точними.

Результатом є реалізація та порівняння моделей прогнозування, що вирішують складні, але дуже важливі проблеми бізнесу, дають змогу ефективно вести інвестиційну політику та впроваджувати штучний інтелект в будь-якій сфері.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Бідюк П. І., Романенко В. Д., Тимошук О. Л. Аналіз часових рядів: навч. посіб. ННК «Інститут прикладного системного аналізу» Національний технічний університет України «Київський політехнічний інститут». 2010. 317 с.
2. Бідюк П. І. Економетричний аналіз часових рядів. Київ: Політехніка, 2007. 250 с.
3. Канторович Г.Г. Анализ временных рядов: лекционные и методические материалы. Москва: Экономический журнал ВШЭ, 2002. 129 с.
4. Бідюк П. І. Часові ряди: моделювання і прогнозування: монографія. Київ: ЕКМО, 2003. 144 с.
5. Магнус Я.Р., Катышев П.К., Песесецкий А.А., Магнус Я.Р. Эконометрика: Начальный курс: учеб. 6-е изд., перед. и доп. Москва: Дело, 2004. 576 с.
6. Вербик М. Путеводитель по современной эконометрике / пер. с англ. Банникова В.А.; научн. ред. и пред. Айвазяна С.А. Москва: Научная книга, 2008. 616 с.
7. Бокс Дж., Дженкинс Г. Анализ временных рядов, прогноз и управление / пер.с англ. Москва: Мир, 1974. 406 с.
8. Simonyan K. Very Deep Convolutional Networks for Large-Scal Image Recognition – Режим доступу до ресурсу:
<http://arxiv.org/abs/1409.1556>
9. Boris N. Oreshkin, Nicolas Chapados, Dmitri Carпов, Yoshua Bengio N-BEATS:Neural basis expansion analysis for interpretable time series forecasting – Режим доступу до ресурсу:
<https://arxiv.org/pdf/1905.10437>
10. Luke B. Godfrey, Michael S. Gashler Neural Decomposition of Time-Series Data for Effective Generalization - Режим доступу до ресурсу:
<https://arxiv.org/pdf/1705.09137>
11. Daniel L. Marino, Kasun Amarasinghe, Milos Manic Building Energy Load Forecasting using Deep Neural Networks - Режим доступу до ресурсу:
<https://arxiv.org/ftp/arxiv/papers/1610/1610.09460>
12. Soofi A.S., Liangyue C. Modelling and Forecasting Financial Data. Techniques of Nonlinear Dynamics. Boston: Springer US. 2002. 488 p.
13. Mujtaba S. M., Nadeem M. Analyzing Stock Markets using Data Warehousing. Journal of Independent Studies and Research. Jan. 2006. Vol. 4. P. 8.
14. Mondal D. A., Maji G., Goto T., Debnath N.C., Sen S. Data Warehouse Based Modelling Technique for Stock Market Analysis. International Journal of Engineering & Technology. 2018. Vol. 7 (3.13). P. 165-170.
15. Selva Prabhakaran ARIMA Model – Complete Guide to Time Series Forecasting in Python
<https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/>

ПІДХОДИ ДО ПРОГНОЗУВАННЯ ФІНАНСОВОГО СТАНУ ПІДПРИЄМСТВА ТА ОЦІНКИ ІНВЕСТИЦІЙНОЇ ПРИВАБЛИВОСТІ СІЛЬСЬКОГОСПОДАРСЬКОГО ПІДПРИЄМСТВА НА ОСНОВІ СТАТИСТИЧНИХ ДАНИХ

Мілявський Ю.Л.¹, Павлуша А.О.²

Національний технічний університет України “Київський політехнічний інститут ім. Ігоря Сікорського”

¹ yuriy.milyavsky@gmail.com, ² nata.pavelichanko@gmail.com

У роботі розглянуто проблему прогнозування фінансових показників (зокрема, прибутку) для сільськогосподарських підприємств на прикладі конкретного елеватора. Проведено аналіз фінансових даних, досліджено кореляційну матрицю, побудовано ряд регресійних моделей для прогнозування виручки, вибрано найкращу модель (з регресорами і трендом) на основі аналізу сукупності показників якості. На основі обраної моделі здійснено прогнозування. Запропоновано стартап проєкт для надання послуг з прогнозування фінансових показників сільськогосподарських підприємств для використання їх керівництвом, інвесторами та державою.

Ключові слова: фінансові показники, сільськогосподарське підприємство, часові ряди, авторегресія, показники якості моделей

1. ВСТУП

Прогнози оточують нас майже всюди: від всім відомого прогнозу погоди, до прогнозу рівня життя на Землі через 1000 років. Всі, хто користується прогнозом скаже, що це дуже зручно. Це можна помітити на прикладі прогнозування розвитку хвороб. Вчасно помітивши тенденцію розвитку захворювання, медики та вчені можуть попередити епідемію відкривши вакцину, чи ввівши рекомендації для людей по запобіганню хвороби.

Незважаючи на зручність прогнозів та існування багатьох підходів до прогнозування досі представники аграрного сектора в Україні запевняють, що неможливо побудувати прогноз на прибуток сільськогосподарського підприємства. Аргументують таку позицію тим, що на прибуток сільськогосподарського підприємства впливає врожай, а на врожай комбінація погодних умов, які майже неможливо спрогнозувати на рік вперед та більше.

Сільське господарство має вагомий вплив на українську економіку. Налагодження сільськогосподарської сфери таким чином, щоб українці мали можливість харчуватися продукцією власного виробництва, може сприяти пониженню цін на продукти для населення (оскільки власне вирощення вимагає менше затрат, ніж покупні у інших країн продукти), як наслідок це може спричинити підвищення рівня проживання в країні та стати одним із способів подолання бідності. Якщо після досліджень агропромислового комплексу знайти можливості працювати не лише на задоволення власних потреб, а й додати можливість експорту власної продукції за кордон, то це очевидне збільшення ВВП, що має безпосередній вплив на економіку держави. Науковці проводили багато досліджень, щоб вивчити стан аграрної сфери в Україні [1, 2]. Це важливо для розуміння необхідності прогнозів саме в цій сфері. Якщо застосована в цій роботі технологія виявиться ефективною, то в подальшому, держава може використовувати дану технологію при плануванні бюджету на наступний

календарний рік. Тобто, спеціалісти зможуть спрогнозувати прибутки визначених підприємств, та надати отримані результати державі, які вона в свою чергу може використовувати при плануванні витрат на новий рік.

2. ПІДХОДИ ДО ПОБУДОВИ ПРОГНОЗІВ ПРИБУТКУ ПІДПРИЄМСТВА

Для побудови прогнозів зручно використовувати часові ряди. Дані, що були отримані від підприємства змінюються з часом, тому даний підхід можливий в цьому випадку.

В роботі проводився аналіз роботи та побудова прогнозу виручки для підприємства «Уманьнасінтрав». Це елеватор в Уманському районі, Черкаська область. Дохід елеватора від переробки зерна в крупу, мучку та дрібку. При переробці використовуються різні технології, а саме тривалість лушення та різний рівень вологості. Підприємство обробляє твердозернові та м'якозернові типи зерна. Усього значень для аналізу 126. Дані зібрані помісячно за 10 років і 6 місяців. Частина даних представлена на рис. 1.

Температура повітря, °C	Вологість, %	Температура зерна, °C	Тип зерна	Вміст волок			Виробництво, т/т	Виробництво, т/т	Виробництво, т/т	Виробництво, т/т	Виробництво, т/т	Виробництво, т/т	Виробництво, т/т	Виробництво, т/т	Виробництво, т/т	Виробництво, т/т	Виробництво, т/т	Виробництво, т/т	
				%	%	%													
20	10	10	II Твердозерновий	91,2	0,1	0,7	14523080	78400	1307123	18000892,3	74700	0	812300	0	253800	1000000	1000	1000	1000
40	10	10	II Твердозерновий	90,8	2,1	1,1	14452900	78400	2054723	18001397,5	74700	0	812300	0	253800	1000000	1000	1000	1000
60	10	10	II Твердозерновий	90,9	2,6	1,5	14322960	78400	2052123	18000892,6	74700	0	812300	0	253800	1000000	1000	1000	1000
80	10	10	II Твердозерновий	91,0	3,9	2,5	12000840	64500	4058723	9818053,5	74700	0	812300	0	253800	1000000	1000	1000	1000
100	10	10	II Твердозерновий	91,7	5,1	3,3	13000800	58110	50580	9702179	74700	0	812300	0	253800	1000000	1000	1000	1000
120	10	10	II Твердозерновий	90,0	6,5	4,6	12500960	58075	20805	8450843	74700	0	812300	0	253800	1000000	1000	1000	1000
140	10	10	II Твердозерновий	91,3	5,8	4,9	13308480	27330	91307,5	9006052,3	74700	0	812300	0	253800	1000000	1000	1000	1000
160	10	10	II Твердозерновий	90,9	9,9	6,1	10808800	25772,5	215000	8000000	74700	0	812300	0	253800	1000000	1000	1000	1000
180	10	10	II Твердозерновий	91,1	8,6	7,1	12000400	32888	1325323	8063123	74700	0	812300	0	253800	1000000	1000	1000	1000
20	13,5	10	II Твердозерновий	91,1	2,3	0,7	14500740	88170	1307523	18044481,5	74700	7,55573404	832950	457,911	253225	1000000	1000	1000	1000
40	13,5	10	II Твердозерновий	90,9	2,1	1	14478800	78400	18075	18219478	74700	7,55573404	832950	457,911	253225	1000000	1000	1000	1000
60	13,5	10	II Твердозерновий	90,8	2,8	1,4	14182530	109688	28245	20082145	74700	7,55573404	832950	457,911	253225	1000000	1000	1000	1000
80	13,5	10	II Твердозерновий	91,0	3,8	2,4	14011730	148508	44818	6842878	74700	7,55573404	832950	457,911	253225	1000000	1000	1000	1000
100	13,5	10	II Твердозерновий	90,9	6,1	3	13878380	108110	94823	8782818	74700	7,55573404	832950	457,911	253225	1000000	1000	1000	1000
120	13,5	10	II Твердозерновий	91,7	5,1	3,2	12000800	208480	18400	8462008	74700	7,55573404	832950	457,911	253225	1000000	1000	1000	1000
140	13,5	10	II Твердозерновий	90,5	6	4,7	13301430	218118	87772,3	9008791,3	74700	7,55573404	832950	457,911	253225	1000000	1000	1000	1000
160	13,5	10	II Твердозерновий	90,7	7,1	5,2	12922980	265110	115000	8570908	74700	7,55573404	832950	457,911	253225	1000000	1000	1000	1000
180	13,5	10	II Твердозерновий	91,3	8,7	7	12204430	31690	130725	8092508	74700	7,55573404	832950	457,911	253225	1000000	1000	1000	1000
20	16,5	10	II Твердозерновий	91,0	1,9	0,5	14526940	78965	8017,5	18004241,5	74700	14,18024460	3149000	915,8021	256400	1000000	1000	1000	1000
40	16,5	10	II Твердозерновий	91,1	1,9	1	14530740	78965	18071	8719880	74700	14,18024460	3149000	915,8021	256400	1000000	1000	1000	1000
60	16,5	10	II Твердозерновий	90	2,8	1,3	14182430	109688	28118	2111898	74700	14,18024460	3149000	915,8021	256400	1000000	1000	1000	1000
80	16,5	10	II Твердозерновий	91	3,7	2,1	14038480	138118	87972,3	8927242,0	74700	14,18024460	3149000	915,8021	256400	1000000	1000	1000	1000
100	16,5	10	II Твердозерновий	91,3	4,8	3	13908880	142908	34823	8760115	74700	14,18024460	3149000	915,8021	256400	1000000	1000	1000	1000
120	16,5	10	II Твердозерновий	90,0	9	6,1	12000900	189708	70597,3	8826215	74700	14,18024460	3149000	915,8021	256400	1000000	1000	1000	1000
140	16,5	10	II Твердозерновий	90,8	5,0	4,7	13118120	209678	87772,3	8928215	74700	14,18024460	3149000	915,8021	256400	1000000	1000	1000	1000
160	16,5	10	II Твердозерновий	91,5	8,7	6	13049630	258290	112338	9047415	74700	14,18024460	3149000	915,8021	256400	1000000	1000	1000	1000
180	16,5	10	II Твердозерновий	91,0	8,8	6,4	12601120	318888	118728	8708215	74700	14,18024460	3149000	915,8021	256400	1000000	1000	1000	1000

Рисунок 1. Отримані дані від підприємства

Для початку роботи виконано фільтрацію даних, що є важливим елементом дослідження. Підприємство під час надання даних повідомило про всі параметри, що можуть впливати на прибуток. Проте насправді не всі дані мають вплив на фінальний результат. Наприклад, було видалено з аналізу дані «тривалість відволожування», оскільки значення є константою. Це не буде впливати на зміну прибутку.

Для обробки даних було побудовано кореляційну матрицю залежності виручки від вологості та часу лушення (рис.2).

Correlation			
	BY MON	TIME	WET
BY MON	1.000000	-0.868251	0.109463
TIME	-0.868251	1.000000	0.000000
WET	0.109463	0.000000	1.000000

Рисунок 2. Матриця кореляції

Побудова матриці кореляції дозволила визначити важливу закономірність та побудувати припущення стосовно ефективності роботи підприємства. Оскільки залежність виручки від тривалості лушення має від'ємне значення, а залежність виручки від вологості зерна має

плюсовий показник, тобто це свідчить про обернену лінійну залежність для показників виручка-час та пряму лінійну залежність для показників виручка-вологість [4, 8]. Тобто при збільшення показника «вологість зерна» (можливі показники вологості 13%-16%, з кроком 0,5%) збільшується показник виручки. Також виручка збільшується при зменшенні часу лушення зерна (можливі показники 20с-180с з кроком 20с.). Таким чином, максимальну виручку можна буде отримати при вологості насіння 16% та за час лушення зерна 20с. Якщо повернемося до даних та перевіримо припущення шляхом пошуку максимального значення, отримаємо наступні результати, що наведено на рисунках 3 та 4.

Тривалість лушення, с (X1)	Вологість, % (X2)	Тип зерна	Виручка, грн	Виручка, млн грн
20	16	М'якезерней	960009,334	9,60
20	15,5	М'якезерней	967540,345	9,68
20	15	М'якезерней	966040,779	9,67
20	14,5	М'якезерней	964120,207	9,64
20	14	М'якезерней	979660,638	9,8
20	13,5	М'якезерней	975720,569	9,76
20	13	М'якезерней	974992,3	9,75
40	16	М'якезерней	971509,834	9,72
40	15,5	М'якезерней	967380,345	9,68
40	15	М'якезерней	9647540,779	9,65
40	14,5	М'якезерней	962670,207	9,63
40	14	М'якезерней	958040,138	9,59
40	13,5	М'якезерней	95703,15	9,56
40	13	М'якезерней	956309,569	9,57
60	16	М'якезерней	951902,414	9,52
60	15,5	М'якезерней	948520,845	9,49
60	15	М'якезерней	947940,779	9,48
60	14,5	М'якезерней	945670,207	9,46
60	14	М'якезерней	943480,138	9,44
60	13,5	М'якезерней	941609,569	9,42
60	13	М'якезерней	937840,034	9,38
80	16	М'якезерней	937422,3	9,38
80	15,5	М'якезерней	936047,845	9,37
80	15	М'якезерней	934801,779	9,34
80	14,5	М'якезерней	933063,207	9,3
80	14	М'якезерней	927371,638	9,28
80	13,5	М'якезерней	925487,069	9,26
80	13	М'якезерней	922685	9,23
100	16	М'якезерней	919219,034	9,2

Рисунок 3. Пошук найбільшої виручки для показників обробки м'яких сортів зерна

Тривалість лушення, с (X1)	Вологість, % (X2)	Тип зерна	Виручка, грн	Виручка, млн грн
20	16	Твердозерней	10067720,28	10,07
20	14,5	Твердозерней	10058226,25	10,06
20	13,5	Твердозерней	10056667,84	10,06
20	16	Твердозерней	10024884,91	10,02
20	14	Твердозерней	10046856,64	10,05
40	16	Твердозерней	10039715,78	10,04
40	16	Твердозерней	10038177,41	10,04
40	15,5	Твердозерней	10035807,84	10,03
40	14,5	Твердозерней	10017141,21	10,02
20	13	Твердозерней	10001707,3	10,01
20	13,5	Твердозерней	9988790,589	9,99
40	14	Твердозерней	9981494,138	9,99
40	13,5	Твердозерней	9960787,069	9,97
40	13	Твердозерней	9949417,3	9,95
60	16	Твердозерней	9932507,414	9,94
60	15,5	Твердозерней	9920360,345	9,93
60	15	Твердозерней	9908990,778	9,91
60	14,5	Твердозерней	9875211,207	9,88
60	14	Твердозерней	9852504,138	9,86
60	13	Твердозерней	9841102,5	9,85
60	13,5	Твердозерней	9830062,089	9,84
80	15,5	Твердозерней	9738212,845	9,74
80	15	Твердозерней	9720373,276	9,73
80	16	Твердозерней	9708067,414	9,71
80	14,5	Твердозерней	9677256,207	9,68
100	16	Твердозерней	9657207,414	9,66
100	15,5	Твердозерней	9643970,345	9,65
100	15	Твердозерней	9639328,276	9,63
80	14	Твердозерней	9628861,638	9,61

Рисунок 4. Пошук найбільшої виручки для показників обробки твердих сортів зерна

Проте припущення не дає чіткої впевненості в ефективності роботи підприємства після вибору даного сценарію роботи.

Так як отримані дані є часовим рядом, то на основі відфільтрованих даних було побудовано різні типи рівнянь авторегресії [3, 5, 6, 7]. В результаті порівняння оцінок було виявлено найкращу модель для подальшого прогнозування прибутку. Частина даних наведено в таблиці 1.

Таблиця 1. Показники, отримані для різних моделей авторегресії.

Модель	R^2	СКП	DW	СеКП	САП	САПП	U
AP(1)	0,301908	37,15439	1,807913	0,647216	0,537113	6,010733	0,983385
AP(4)	0,331257	31,80088	1,994592	0,650588	0,538119	6,037392	0,979757
AP(7)	0,351645	33,67155	2,109945	0,655933	0,655933	6,141945	0,976059
AP(11)	0,972906	1,351561	2,031354	0,296172	0,243550	2,736651	0,427008
AP(12)	0,972983	1,335510	1,968940	0,294263	0,240904	2,709157	0,423287
AP(14)	0,973359	1,309259	1,998872	0,298760	0,245208	2,459805	0,425987
AP(16)	0,974120	1,271503	1,989156	0,283336	0,232562	2,610530	0,399008
AP(17)	0,974058	1,269783	1,997185	0,288819	0,240220	2,695588	0,405153
AP(8) +регресори	0,999631	0,019121	0,268604	0,012783	0,009921	0,210940	0,018059
AP(8)+регре сори+тренд	0,999900	0,005179	0,929176	0,006644	0,005033	0,055541	0,009603

Для вибору найкращої моделі було розглянуто показники R^2 , середньо квадратична похибка статистична (СКП), коефіцієнт Дарбіна-Уотсона, показники якості такі, як середня квадратична похибка (СеКП), середня абсолютна похибка, середня абсолютна похибка в процентах, коефіцієнт Тейла [3, 6].

3. ОТРИМАНІ РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Для перевірки правильності роботи прогнозу, побудуємо прогноз на період з уже відомими даними. Тестовий прогноз продемонстрував високу якість, про що свідчить рис. 5.

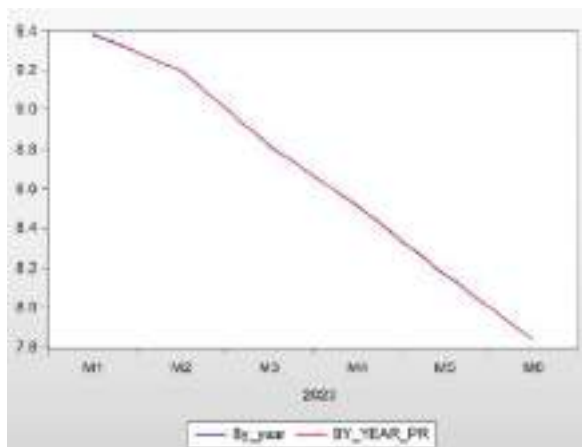


Рисунок 5. Графік реального та прогнозованого значення за 6 місяців

На графіку синім кольором позначено реальні дані, червоним прогнозовані. Як видно з рисунку, прогнозовані значення майже повністю відповідають реальним. Це свідчить про високу точність побудованого прогнозу.

Побудуємо динамічний прогноз, використовуючи найкращу модель, знайдену у результаті проведеного дослідження [8]. Обираємо саме динамічний прогноз, оскільки він дозволяє прогнозувати значення на обраний час. Перше значення прогнозу буде побудоване на основі реальних історичних даних, всі подальші будуються вже з урахуванням прогнозованих значень. Оскільки в результаті маємо надати інвестору прогноз мінімум на рік, вибираємо період 1 рік. Для наглядності результатів, наведемо отриманий прогноз разом з реальними даними за попередній період (таблиця 2).

Таблиця 2. Порівняння прогнозованої виручки з історичними даними за попередній період.

Місяць	Виручка, млн.грн (2022-2023)	Виручка, млн.грн (2023-2024)
Липень	8,49	8,63
Серпень	8,13	8,60
Вересень	7,8	8,56
Жовтень	9,9	8,52
Листопад	9,72	8,48
Грудень	9,52	8,44
Січень	9,38	8,40
Лютий	9,2	8,36
Березень	8,82	8,33
Квітень	8,52	8,32
Травень	8,17	8,27
Червень	7,84	8,23
Річний	105,49	101,14

4. ПРАКТИЧНЕ ЗАСТОСУВАННЯ ТА ВИСНОВКИ

Прогнозування широко використовується для соціальних досліджень, в метеорологічній сфері, в медицині, тощо. Проте дуже часто можна зустріти людей, які кажуть, що неможливо надати відносно точний прогноз для роботи деяких підприємств. В основному це стосується підприємств сільського господарства. Така позиція викликає труднощі при аналізі роботи підприємства. Також, відсутність прогнозів відлякують інвесторів від сільськогосподарської сфери.

В результаті дослідження було показано, що обрана модель та обраний метод прогнозування показали високу точність, що дозволяє робити прогноз на майбутнє. Отриманий прогноз пропонується представити потенційному інвестору. Таким чином людина, яку зацікавила робота конкретного підприємства може зробити висновок, як швидко інвестор зможе отримати прибуток з власної інвестиції, чи доцільно робити вкладення саме зараз в це підприємство, чи не будуть уже налаштовані процеси збитковими тощо. Також, в результаті проведеного аналізу роботи підприємства, в даному випадку це був елеватор, можна запропонувати інвестору чи керівнику підприємства сценарії роботи, при яких підприємство може отримувати більший прибуток, ніж для технології, яку використовують на проведення аналізу. Для підтримки точності прогнозу, рекомендується регулярно оновлювати історичні дані. В разі великого відхилення очікуваних результатів від реальних прогноз буде перероблено.

Під час вивчення потенційного ринку для стартапу, було виявлено три основні цільові аудиторії. Керівники підприємств могли б після замовлення послуги отримати аналіз роботи

їх підприємства, потенційний прибуток на період, що був обраний для прогнозування, а також можливі сценарії розвитку, для забезпечення більшого прибутку. Інвесторам, що замовили б аналіз підприємства, що їх цікавить, був би наданий схожий звіт, як керівникам підприємства, але інвестори могли б зробити висновок, чи цікавить їх дане підприємство з точки зору інвестування. Проектні організації пропонують багато послуг, в тому числі побудова нового підприємства та введення його в експлуатацію або оптимізація роботи вже існуючого підприємства. На початку роботи над подібними проектами компанії буде корисно отримати прогнози стосовно можливого прибутку, щоб зрозуміти доцільність побудови нового підприємства чи змін в роботі вже існуючого, та прогнози стосовно запитів на переробку та зберігання врожаю, щоб налаштувати логістичні процеси. Для подальшого розвитку використання технології можна також розглядати вектор взаємодії з подібними компаніями, з подальшою побудовою власної компанії з унікальним переліком послуг.

Таким чином, проведене дослідження продемонструвало можливість якісного прогнозування фінансових показників сільськогосподарського підприємства та виявило шляхи подальшого застосування отриманих прогнозів.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Засоба Сергій, *Розвиток агропромислового комплексу України*, Український журнал прикладної економіки та техніки, 2020, 402-409.
2. Агропромисловий комплекс України: сучасний стан розвитку та особливості. URL: <https://www.profbuild.in.ua/uk/stati-2/871-agropromislovij-kompleks-ukrajini-suchasnij-stan-rozvitku-ta-osoblivosti> (дата звернення 10.09.2023).
3. Бідюк П. І. Аналіз часових рядів (навчальний посібник) / Бідюк П.І., Романенко В.Д., Тимощук О.Л. // Аналіз часових рядів (навчальний посібник) — Київ: Політехніка, 2010. — 317 с.
4. Бідюк П. І. Системний підхід до побудови математичних моделей на основі часових рядів / П. І. Бідюк, І. В Баклан., В. М Рифа. // Системні дослідження та інформаційні технології. – 2002. – № 3. – 131 с.
5. Бідюк П. І. Часові ряди: моделювання та прогнозування / П. І. Бідюк, О. І. Савенков, І. В Баклан. – Київ : ЕКМО, 2004. – 144 с.
6. Enders W. Applied econometric time series / Enders W. – New York: John Wiley & Sons, Inc., 1995. – 434 p
7. Dehove A, Commault J, Petitclerc M, Teissier M, Macé J. Economic analysis and costing of animal health: a literature review of methods and importance. Rev Off Int Epizoot. (2012) 31:605–17. 10.20506/rst.31.2.2146
8. Berentsen P, Dijkhuizen A, Oskam A. A dynamic model for cost-benefit analyses of foot-and-mouth disease control strategies. Prev Vet Med. (1992) 12:229–43. 10.1016/0167-5877(92)90052-H
9. Gethmann J, Probst C, Sauter-Louis C, Conraths FJ. Economic analysis of animal disease outbreaks - BSE and Bluetongue disease as examples. Berl Munch Tierarztl Wochenschr. (2015) 128:478–82. 10.2376/0005-9366-128-478

ПОРІВНЯЛЬНИЙ АНАЛІЗ ТА ПОКРАЩЕННЯ МОДЕЛЕЙ ПРОГНОЗУВАННЯ ЦІН АКЦІЙ НА ФІНАНСОВОМУ РИНКУ

Муравльов А.Д., Гуськова В.Г.

Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського»

1. ВСТУП

У динамічному середовищі фінансового ринку точне прогнозування цін на акції вже давно перебуває в центрі уваги інвесторів, аналітиків та дослідників. Безперервні припливи і відпливи ринкових тенденцій під впливом безлічі факторів – від економічних показників до геополітичних подій – роблять прогнозування цін на акції складним завданням. Оскільки фінансовий світ стає все більш взаємопов'язаним і керованим даними, попит на надійні моделі прогнозування зростає.

Метою є дослідження сфери прогнозування цін на акції через призму порівняльного аналізу та вдосконалення. Заглиблюючись у тонкощі існуючих моделей, ми розглядаємо різноманітні методології, що застосовуються для прогнозування цін на акції, та досліджуємо їхню ефективність у фінансовому ландшафті, що постійно змінюється. Основна мета полягає в тому, щоб виявити не лише сильні та слабкі сторони, притаманні сучасним моделям прогнозування, але й запропонувати вдосконалення, які можуть сприяти розробці більш точних та надійних інструментів прогнозування.

Шляхом всебічного вивчення домінуючих підходів, статистичних методів та алгоритмів машинного навчання, ця стаття має на меті висвітлити досягнення у сфері прогнозування фондових цін. Здійснюючи порівняльний аналіз, ми прагнемо виокремити ключові ідеї, які проливають світло на найефективніші стратегії прогнозування руху цін на акції. Крім того, дослідження шляхів вдосконалення існуючих моделей відкриває шлях до підвищення точності прогнозування та зменшення ризиків, пов'язаних з прийняттям фінансових рішень.

2. ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

Проблема дослідження:

Фінансові ринки є складними та волатильними системами, де ціни акцій піддаються впливу різних факторів, таких як економічні новини, геополітичні події, інтереси інвесторів та багато інших. Прогнозування цін акцій є критично важливим завданням для інвесторів, брокерів, фондових аналітиків та компаній, що публікують акції. Точні та надійні моделі прогнозування цін акцій можуть допомогти приймати кращі рішення з управління інвестиціями та ризиками.

Мета дослідження:

Метою є проведення порівняльного аналізу різних моделей прогнозування цін акцій на фінансовому ринку з метою вдосконалення їх точності та надійності. Основні завдання дослідження включають:

1. Вивчення і аналіз існуючих моделей прогнозування цін акцій, таких як моделі часових рядів, машинного навчання та штучні нейронні мережі.
2. Порівняльний аналіз різних методів та моделей на історичних даних ринку для визначення їхньої точності та відповідності реальним умовам.
3. Розробка покращень та оптимізація обраних моделей з метою зменшення помилок та покращення прогностичної здатності.

4. Практичне застосування найкращої моделі прогнозування на реальних даних ринку для перевірки її ефективності в реальних умовах.

Висновки та рекомендації для інвесторів та фінансових аналітиків щодо використання найкращої моделі прогнозування для прийняття рішень щодо інвестицій та управління ризиками на фінансовому ринку.

3. ОГЛЯД МЕТОДІВ ПРОГНОЗУВАННЯ

На сьогоднішній день існує низка сучасних методів прогнозування цін акцій, і вони можуть бути поділені на кілька груп в залежності від підходу та методології.

Методи машинного навчання:

– **Лінійна регресія:** Моделює залежність між ціною акцій та факторами, такими як обсяг торгівлі, індикатори ринку тощо.

– **Дерева рішень та випадкові ліси:** Використовуються для розробки моделей, які можуть враховувати нелінійні взаємозв'язки між факторами та цінами акцій.

RF (random forest) – це безліч вирішальних дерев. У задачі регресії їх відповіді усереднюються, в завданні класифікації приймається рішення голосуванням за більшістю. Всі дерева будуються незалежно за наступною схемою:

– Вибирається підвибірка навчальної вибірки розміру – по ній будується дерево (для кожного дерева – своя підвибірка);

– Для побудови кожного розщеплення в дереві переглядаємо `max_features` випадкових ознак (для кожного нового розщеплення – свої випадкові ознаки);

– Вибираємо найкращу ознаку і розщеплення по ній (за задалегідь заданим критерієм). Дерево будується, як правило, до вичерпання вибірки (поки в листі не залишаться представники тільки одного класу), але в сучасних реалізаціях є параметри, які обмежують висоту дерева, число об'єктів в листі і число об'єктів в підвибірці, при якому проводиться розщеплення

Зрозуміло, що така схема побудови відповідає головному принципу ансамблювання (побудови алгоритму машинного навчання на базі кількох, в даному випадку вирішальних дерев): базові алгоритми повинні бути хорошими і різноманітними (тому кожне дерево будується на своїй навчальній вибірці і при виборі розщеплення є елемент випадковості).

Чим більше дерев, тим краща якість, але час налаштування і роботи RF також пропорційно збільшуються. Часто при збільшенні кількості дерев якість на навчальній вибірці підвищується (може навіть доходити до 100%), а якість на тестовій вибірці виходить на асимптоту.

Метод випадкового лісу заснований на методі вирішальних дерев. Випадковий ліс - це безліч вирішальних дерев, а клас об'єкта, що проходить класифікацію, вибирається голосуванням більшістю.

Згенеруємо випадкову вибірку S розміру l по вихідній навчальній вибірці $D = \{x_i y_i\}_{i=1}^l$.

За вибіркою S індукувати неусічене дерево рішень T_i з мінімальною кількістю спостережень в термінальних вершинах рівним n_{min} , рекурсивно слідуючи до наступного підалгоритма:

- з вихідного набору p ознак випадково вибрати p ознак,
- з p ознак вибрати ознаку, яка забезпечує найкраще рішення,
- розщепити вибірку, відповідну до оброблюваної вершини, на дві підвибірки.

В результаті отримуємо ансамбль дерев рішень $\{T_i\}_{i=1}^B$.

Класифікація нових спостережень здійснюється наступним чином: нехай $\hat{y}_l(x) \in \{y_1, \dots, y_l\}$ – клас, передбачений деревом рішень T_i , тобто $T_i(x) = \hat{y}_l(x)$, тоді $\widehat{y}_{rf}^B(x)$ – клас, найбільш часто зустрічається в множині $\{\hat{y}_b(x)\}_{b=1}^B$.

– **Метод опорних векторів:** шукає оптимальну розділяючу гіперплощину, яка максимізує відстань між двома наборами даних різних класів.

Лінійна класифікація

Нехай є навчальна вибірка:

$$(\bar{x}_1, y_1), \dots, (\bar{x}_m, y_m), \bar{x}_i \in R^n, y_i \in \{-1, 1\}.$$

Метод опорних векторів буде класифікаційну функцію F у вигляді:

$$F(x) = \text{sign}(\bar{w} \cdot \bar{x} - b),$$

де $\bar{w} \cdot \bar{x}$ – скалярний добуток, \bar{w} — нормальний вектор до розділювальної гіперплощини, b – зсув. Ті об'єкти, для яких $F(x) = 1$, потрапляють в один клас, а об'єкти з $F(x) = -1$ – в інший.

Якщо навчальна вибірка містить два класи даних, які можна лінійно розділити, то ми можемо обрати дві паралельні гіперплощини, які розділяють два класи даних так, що відстань між ними якомога більша. Область, обмежена цими двома гіперплощинами, називається «розділенням», а максимально розділова гіперплощина це гіперплощина, яка лежить посередині між цими двома. Ці гіперплощини може бути описано рівняннями:

$$\begin{aligned} \bar{w} \cdot \bar{x} - b &= 1; \\ \bar{w} \cdot \bar{x} - b &= -1. \end{aligned}$$

З геометричної точки зору, відстанню між цими двома гіперплощинами є $\frac{2}{\|\bar{w}\|}$, тому для максимізації відстані між ними нам треба мінімізувати $\|\bar{w}\|$. Оскільки ми також маємо завадити потраплянню точок даних до розділення, ми додаємо таке обмеження: для кожного i ,

$$\begin{aligned} \text{або } \bar{w} \cdot \bar{x}_i - b &\geq 1 \text{ якщо } y_i=1; \\ \text{або } \bar{w} \cdot \bar{x}_i - b &\leq -1 \text{ якщо } y_i=-1. \end{aligned}$$

Ці обмеження стверджують, що кожна точка даних мусить лежати з правильного боку розділення. Ці дві нерівності можна записати як одну:

$$y_i(\bar{w} \cdot \bar{x}_i - b) \geq 1, \text{ для всіх } 1 \leq i \leq n.$$

З цього отримуємо задачу оптимізації “Мінімізувати $\|\bar{w}\|$ за умови $y_i(\bar{w} \cdot \bar{x}_i - b) \geq 1$ для $i=1, \dots, n$.”

Значення \bar{w} та b які розв’язують цю задачу — визначають класифікатор:

$$\bar{x} \rightarrow \text{sgn}(\bar{w} \cdot \bar{x} - b).$$

Очевидним, але важливим наслідком цього геометричного опису є те, що максимально розділова гіперплощина повністю визначається тими \bar{x}_i , які лежать найближче до неї. Ці \bar{x}_i називають опорними векторами.

Нелінійна класифікація

На практиці випадки, коли дані можна розділити гіперплощиною, або, як ще кажуть, лінійно, досить рідкісні.

У цьому разі чинять так: усі елементи навчальної вибірки вкладають у простір X вищої розмірності за допомогою спеціального відображення

$$\varphi: R^n \rightarrow X.$$

При цьому відображення φ вибирають так, щоб у новому просторі X вибірка була лінійно роздільна.

Класифікаційна функція F набуває вигляду

$$F(x) = \text{sign}(\bar{w} \cdot \varphi(\bar{x}) - b).$$

Вираз $K(\bar{x}_i, \bar{x}_j) = \varphi(\bar{x}_i) \cdot \varphi(\bar{x}_j)$ називається ядром класифікатора. З математичної точки зору **ядром** може слугувати будь-яка позитивно визначена симетрична функція двох змінних. Позитивна визначеність необхідна для того, щоб відповідна функція Лагранжа в задачі оптимізації була обмежена знизу, тобто задача оптимізації була б коректно визначена. Слід зазначити, що робота в просторі ознак високої вимірності збільшує похибку узагальнення опорно-векторних машин, хоча за достатньої кількості зразків цей алгоритм все одно працює добре.

Точність класифікатора залежить, зокрема, від вибору ядра. Найчастіше на практиці зустрічаються такі ядра:

– Поліноміальне. При його використанні обчислюються усі можливі комбінації вихідних ознак об'єкта до певної степені;

$$K(\bar{x}_i, \bar{x}_j) = (\bar{x}_i \cdot \bar{x}_j + C)^d.$$

– Радіальна базисна функція. При його використанні обчислюються усі можливі поліноміальні комбінації усіх степенів щодо вихідних ознак об'єкта.

$$K(\bar{x}_i, \bar{x}_j) = e^{-\gamma \|\bar{x}_i - \bar{x}_j\|^2}.$$

– Сигмоїдне

$$K(\bar{x}_i, \bar{x}_j) = \tanh(k\bar{x}_i \cdot \bar{x}_j + C).$$

Ядро пов'язано з перетворенням $\varphi(\bar{x})$ через рівняння $K(\bar{x}_i, \bar{x}_j) = \varphi(\bar{x}_i) \cdot \varphi(\bar{x}_j)$. Значення w також знаходиться в перетвореному просторі:

$$\bar{w} = \sum_i \alpha_i y_i \varphi(\bar{x}_i).$$

Скалярні добутки з w для класифікації, знов-таки, може бути обчислювано за допомогою ядрового трюку, тобто:

$$\bar{w} \cdot \varphi(\bar{x}) = \sum_i \alpha_i y_i \varphi K(\bar{x}_i, \bar{x}).$$

4. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Результати прогнозування за метриками представлено в таблиці 1.

Таблиця 1. Результати прогнозування за метриками

Метод	RMSE	MSE	MAE	MAPE	R2 Score
RF(10 trees)	3,62	13,152	2,86	0,0172	0,8525
RF(100 trees)	3,75	14,067	2,975	0,0178	0,8422
RF(1000 trees)	3,638	13,236	2,864	0,01718	0,8516
RF(10000 trees)	3,610	13,037	2,842	0,01705	0,853
SVM(linear)	2,97	8,84	2,35	0,014	0,9008
SVM(polynomial)	7,28	53,044	5,722	0,034	0,405
SVM(rbf)	3,69	13,66	2,931	0,0175	0,8468
SVM(sigmoid)	3679,8	13541002	2718	15	-151812

5. ВИСНОВКИ

Внутрішня структура моделей:

SVM: SVM будує лінійну або нелінійну гіперплощину, що розділяє дані, і залежно від типу задачі використовує різні ядерні функції для перетворення даних у вищорозмірний простір. Ідея полягає в тому, що дані можуть бути лінійно нероздільними у вихідному просторі, але можуть бути розділені гіперплощиною у вищорозмірному просторі.

Random Forest: Випадковий ліс складається з декількох рішачих дерев, які побудовані на основі підмножини даних та підмножини ознак. Кожне дерево незалежно приймає рішення, шляхом проходження вниз по дереву за допомогою послідовних тестів на ознаках. Остаточне рішення випадкового лісу визначається більшістю голосів дерев.

Обробка даних:

SVM: SVM вимагає числових значень ознак та масштабування даних. Категоріальні ознаки потребують попереднього перетворення в числові.

Random Forest: Випадковий ліс може працювати з числовими та категоріальними ознаками без попереднього масштабування даних.

Обидва методи мають свої переваги та обмеження, і вибір між ними залежить від характеристик даних, розміру вибірки та вимог до точності та швидкодії моделі. SVM часто ефективний для завдань з невеликими даними та коли важлива межа розділення класів. Випадковий ліс може бути корисним у великих наборах даних, коли потрібна висока точність та стійкість до перенавчання.

На датасеті, на якому порівнювалися моделі, Random Forest показав середні результати, помітна просадка точності при першому збільшенні кількості дерев, однак на більшій кількості точність підвищилась. Random Forest показав себе як більш стабільний варіант, однак класифікація за допомогою лінійного класифікатора SVM надала кращі результати.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Орельен Жерон. Прикладное машинное обучение с помощью Scikit-Learn ч TensorFlow. 2018. 688 с.
2. Случайный лес (Random Forest). URL: <https://dyakonov.org/2016/11/14/случайный-лес-random-forest/>.
3. Чистяков С.П. Случайные леса: обзор // Труды Карельского научного центра РАН. 2013. №1. С. 117-136.
4. Метрики в задачах машинного обучения. URL: <https://habr.com/ru/company/ods/blog/328372/>.
5. Ben-Hur, Asa; Horn, David; Siegelmann, Hava; Vapnik, Vladimir N. "Support vector clustering" (2001);"
6. Зайченко Ю.П. Основи проектування інтелектуальних систем. 2004. с.49-66
7. Random forest. URL: https://ru.wikipedia.org/wiki/Random_forest.
8. <https://www.kaggle.com/competitions/nlp-getting-started/overview>

МОДЕЛІ ОПТИМАЛЬНОГО РОЗПОДІЛУ ДАНИХ

Мухін В.Є., Яковлева А.П., Шмідт А.Є.¹

Національний технічний університет України «Київський політехнічний інститут
ім. Ігоря Сікорського»

¹ shmidt.anatolii@lil.kpi.ua

Мета дослідження полягає у розробці та оптимізації математичних моделей для ефективного розподілу даних у різноманітних структурах мережі, таких як ієрархічна, кільцева та решітчаста. Використовуючи теоретичні та емпіричні методи дослідження, надано новий підхід до оптимізації розподілу даних. Наукова новизна полягає у вдосконаленні математичних концепцій для різних типів мереж та впровадженні нової стратегії оптимізації. Застосування розроблених моделей може позитивно вплинути на широкий спектр областей, де важливий оптимальний розподіл даних, включаючи технологічні, комунікаційні та наукові сфери.

Ключові слова: розподіл даних, розподілені бази даних, математична модель, мережева структура, оптимізація.

1. ВСТУП

У контексті зростаючого обсягу даних у сучасному світі, ефективний розподіл і управління ними стає невід'ємною складовою оптимального функціонування різноманітних інформаційних систем. Особливо важливим стає це завдання в рамках різних мережевих структур, таких як ієрархічні, кільцеві та решітчасті мережі.

У цьому дослідженні ставиться за мету розгляд і оптимізацію математичних моделей для оптимального розподілу даних у зазначених типах мереж. Зосереджуючись на теоретичних аспектах та використовуючи математичні моделі, дослідження спрямоване на створення інноваційних підходів, призначених вдосконалити процеси управління даними в різноманітних мережах та сприяти підвищенню їхньої продуктивності.

2. МАТЕМАТИЧНІ МОДЕЛІ РОЗПОДІЛУ ДАНИХ

Перш за все, дамо визначення мережам, що будуть для яких побудуємо моделі.

Розглядатимуться розподілені бази даних. Розподілена база даних — сукупність логічно взаємопов'язаних баз даних, розподілених у комп'ютерній мережі. Логічний зв'язок баз даних в розподіленій базі даних забезпечує система управління розподіленою базою даних, яка дозволяє управляти розподіленою базою даних таким чином, щоб створювати у користувачів ілюзію цілісної бази даних [1].

Моделі побудовані для наступних типів (топологій) мереж: кільцевої, решітчастої, ієрархічної.

Кільцева топологія — це послідовний ланцюжок у замкнутому циклі. Дані переміщуються по кільцю в одному напрямку. Коли один вузол надсилає дані іншому, дані проходять через кожен проміжний вузол кільця, поки не досягнуть місця призначення. Проміжні вузли повторюють (ретранслюють) дані, щоб сигнал був сильним [2].

Плоска решітчаста топологія передбачає, що кожен її вузол сполучений із найближчим сусідом (Рисунок 1) [3]. У цій роботі моделюється багаторівнева плоска решітчаста мережа, де на кожному рівні 4 вузла.

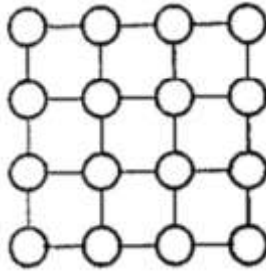


Рисунок 1. Плоска решітчаста мережа

Топологія дерева — це мережа, в якій кожен вузол пов'язаний з іншими в ієрархії. Її також називають ієрархічною топологією, оскільки в цій топології всі елементи розташовані як гілки дерева. У топології дерева будь-які два пов'язані вузли можуть мати лише одне взаємне з'єднання, отже, між ними може бути лише одне з'єднання [4].

Перейдемо до загальної моделі розподілу даних. Математична задача полягає у знаходженні мінімальної кількості серверів в заданій мережі при якій час відповіді теж буде мінімальний. Зі зміною типу мережі змінюється середній шлях повідомлення (кількість переходів між вузлами). Для спрощення моделі припускаємо, що дані розподілені між вузлами мережі рівномірно та сервери мають однакову пропускну здатність.

Для заданої задачі доцільно сформуванати наступне рівняння:

$$T(N_s) = \frac{N}{N_s} * T_s + L(N_s) * T_r.$$

$T(N_s)$ – це функція середнього часу відповіді від мережі в залежності від кількості серверів;

N_s – це кількість серверів;

N – це кількість записів в розподіленій базі даних;

T_s – це час пошуку на одному сервері;

$L(N_s)$ - це функція середнього шляху повідомлення в мережі в залежності від кількості серверів;

T_r – це час переходу повідомлення між серверами.

Це загальне рівняння, що описує загальну модель. Для кожного типу мережі необхідно визначити власну функцію середнього шляху $L(N_s)$.

Для кільцевої мережі визначаємо наступну функцію середнього шляху повідомлення ($\lceil \cdot \rceil$ -оператор округлення до меншого цілого числа):

$$L^1(N_s) = \frac{N_s \lceil \frac{N_s+1}{2} \rceil}{N_s}.$$

Розглядається багаторівнева решітчаста мережа, що на кожному рівні має 4 вузли. Для цієї мережі визначаємо наступне рівняння:

$$L^2(N_s) = \frac{N_s (\lceil \frac{N_s}{4} \rceil + 1)}{N_s - 1}.$$

Для ієрархічної мережі введемо параметр I , що визначатиме кількість підпорядкованих вузлів для кожного вузла вищого рангу. Визначимо рівняння:

$$L^3(N_s) = 2 * \log_I N_s.$$

3. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Для дослідження було створено програму, що реалізує описані математичні моделі. Програма створена у хмарному середовищі розробки Google Colab на мові програмування Python.

Для експерименту було використано наступні параметри: Кількість записів (N) – 1000; Час пошуку (T_s) – 0,0021 с; Час переходу (T_r) – 0,01 с.

Проведемо аналіз отриманих результатів кільцевої мережі (Рисунок 2). Середній час відповіді T зменшується до певного моменту, до оптимального значення кількості серверів N_s^* , а далі середній час збільшується. При збільшенні кількості серверів збільшується час передачі, тому після досягнення оптимального значення N_s^* сенсу додавати нові сервери немає.

Розглянемо графік на меншому відрізку (Рисунок 3) в який входить $N_s^*=21$. Бачимо, що для парної кількості серверів середній час відповіді T більший ніж сусідніх непарних значень. Це можна пояснити тим, що шлях запиту для непарної кількості серверів, наприклад 19, такий же як і для парної, наприклад 20, а якщо додати ще один сервер, то середнє навантаження зменшиться і відповідно середній час відповіді T .

Мінімальне значення середнього часу відповіді T^* досягається при N_s^* , $T^* = 0,2152$ с.

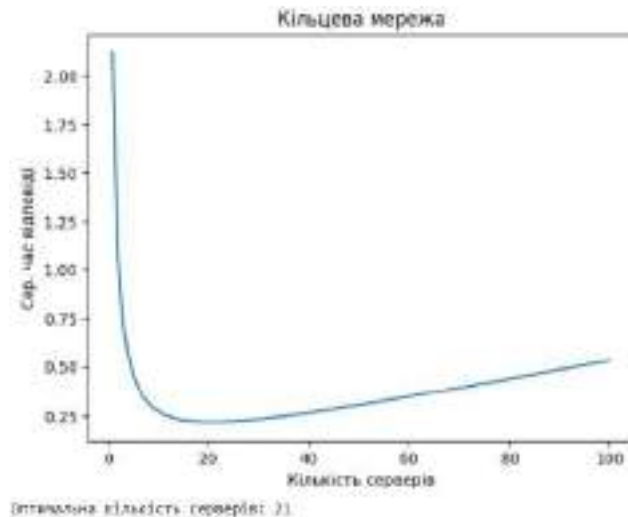


Рисунок 2. Результати кільцевої мережі

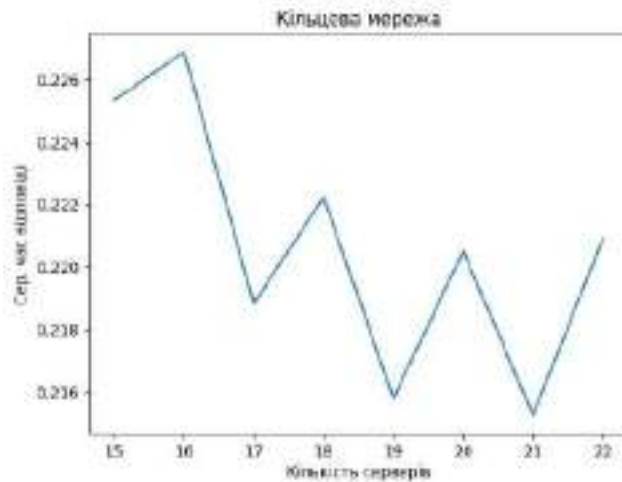


Рисунок 3. Результати кільцевої мережі на меншому відрізку

Проведемо аналіз отриманих результатів для решітчастої мережі (Рисунок 4). Середній час відповіді T зменшується до певного моменту, до оптимального значення кількості серверів Ns^* , а далі середній час збільшується.

Розглянемо графік на меншому відрізку (Рисунок 5) в який входить $Ns^*=32$. На цьому графіку краще помітно, що значення T є однаковим на одному рівні. Рівень складається з 4 серверів.

Мінімальне значення середнього часу відповіді T^* досягається при Ns^* , $T^* = 0,1606$. Порівняно з кільцевою мережею оптимальний середній час відповіді менший, але і серверів потрібно значно більше.

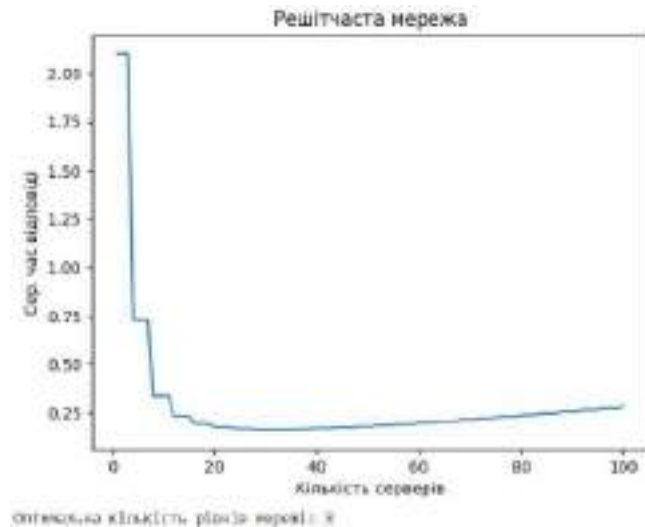


Рисунок 4. Результати решітчастої мережі

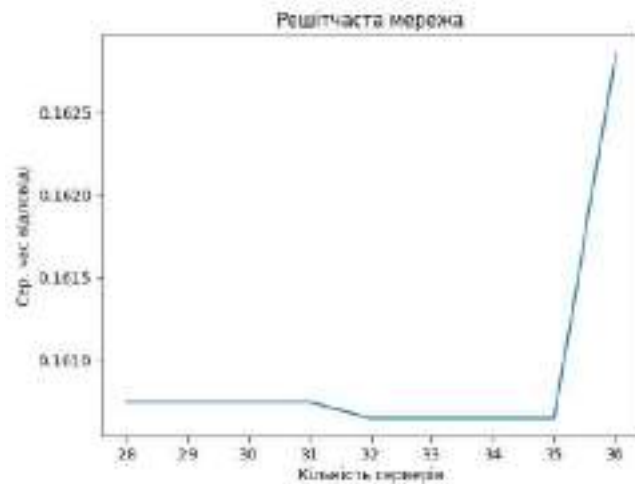


Рисунок 5. Результати решітчастої мережі на меншому відрізку

Проведемо аналіз отриманих результатів (Рисунок 6). Згідно встановлених параметрів сервери в ієрархії мають по два підпорядковані сервери меншого рангу ($I = 2$). Середній час відповіді T досягає оптимального значення кількості серверів Ns^* , а далі більше оптимального.

Розглянемо графік на меншому відрізку (Рисунок 7) в який входить $Ns^*=64$. На графіку помітно, що при збільшенні рівня ієрархії суттєво збільшується середній час відповіді T . Це

викликано тим, що збільшується кількість переходів які потрібно пройти для доступу до даних, що збільшує час передачі.

Мінімальне значення середнього часу відповіді T^* досягається при Ns^* , $T^* = 0,1728$. Порівняно з кільцевою мережею оптимальний середній час відповіді менший, але і серверів потрібно значно більше. Решітчаста мережа показала кращі результати по обом показникам.

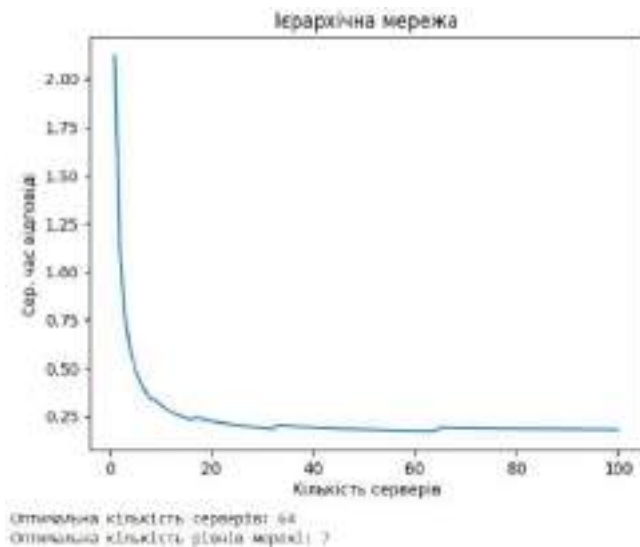


Рисунок 6. Результати ієрархічної мережі

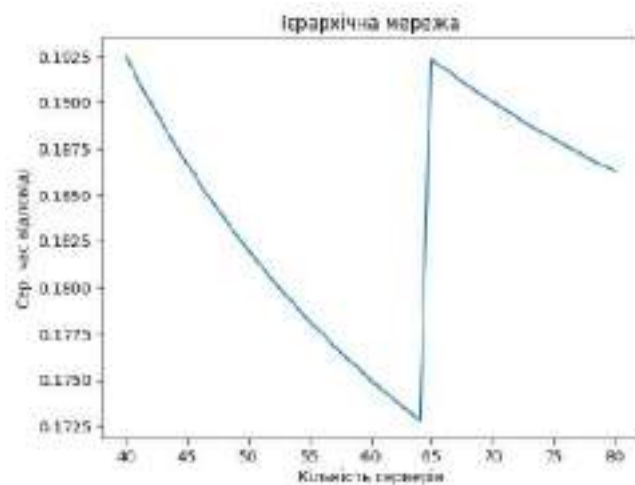


Рисунок 7. Результати ієрархічної мережі на меншому відрізку

У попередніх експериментах було проведено порівняння при наступних параметрах: кількість записів (N) – 1000; час пошуку (T_s) – 0,0021 с; час переходу (T_r) – 0,01 с. Для цих параметрів оптимальною за середнім часом відповіді виявилась решітчаста мережа.

Проведемо експеримент для вдвічі більшої кількості записів. Отже, параметри наступні: кількість записів (N) – 1000; час пошуку (T_s) – 0,0021 с; час переходу (T_r) – 0,01 с; кількість зв'язків ієрархічної мережі (I) – 2.

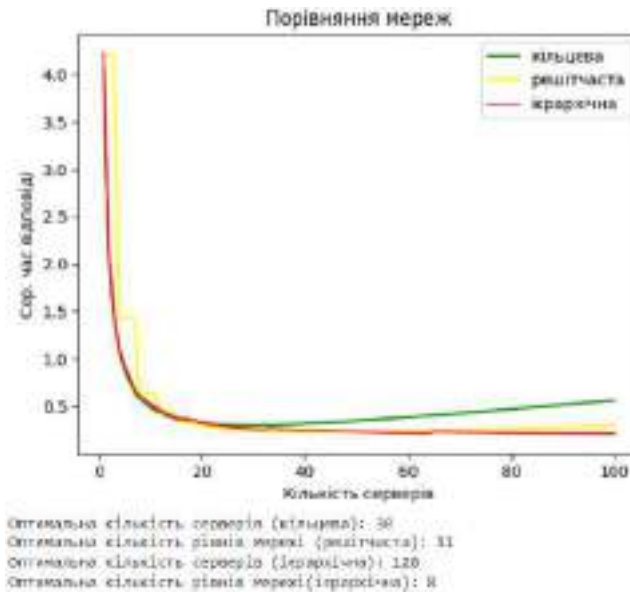


Рисунок 8. Результати для другого набору параметрів

Найменший оптимальний середній час відповіді $T^* = 0,1928$ досягається ієрархічною мережею, але за рахунок великої кількості серверів.

4. ВИСНОВКИ

У висновку важливо підкреслити, що використання математичних моделей є критично важливим для вирішення складних завдань в галузі оптимізації розподілу даних у серверних мережах. Математичні моделі надають структурований та науково обґрунтований підхід до вивчення складних систем, дозволяючи аналізувати взаємодію різних елементів та прогнозувати їхню поведінку.

В результаті дослідження вдалося провести експерименти та проаналізувати їх. Експерименти показали, що ієрархічна мережа найкраще проявляє себе при великій кількості даних або при малій швидкості обробки інформації. Кільцева та решітчаста мережі проявляють себе краще при менших об'ємах даних.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. К. Дж. Дейт Введение в системы баз данных / An Introduction to Database Systems. 8-ме вид. "Вильямс", 2005. 1328 с.
2. Networking complete. 2-ге вид. San Francisco : Sybex, 2001. 877 с.
3. Комп'ютерні системи : методичні вказівки до лабораторних робіт / укл.: Баловсяк С. В., Одайська Х. С. Чернівці : Чернівецький національний університет ім. Ю. Федьковича, 2021. 72 с
4. What is Tree Topology? Definition and Explanation - javatpoint. www.javatpoint.com. URL: <https://www.javatpoint.com/what-is-tree-topology> (дата звернення: 29.11.2023).к

ПРОГНОЗУВАННЯ КРЕДИТНОЇ СПРОМОЖНОСТІ КЛІЄНТІВ БАНКУ НА ОСНОВІ АНАЛІЗУ ФІНАНСОВИХ ДАНИХ

Петровський В.Є., Гуськова В.Г.

Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського»

1. ВСТУП

У сучасних умовах економічної нестабільності та високого рівня конкуренції, важливість ефективного управління ризиками, пов'язаними з кредитуванням, набуває першочергового значення для фінансових установ. Одним із ключових аспектів цього управління є прогнозування кредитної спроможності клієнтів банку. Здатність ефективно визначити й управляти ризиками неплатоспроможності має велике значення для забезпечення стійкості та сталого розвитку фінансових установ.

Однією з перспективних стратегій є використання аналізу фінансових даних для створення точних та надійних моделей прогнозування кредитної спроможності. Розвиток технологій у галузі обробки даних та штучного інтелекту відкриває нові можливості для створення комплексних та інноваційних підходів до оцінки ризиків.

Дана наукова стаття присвячена вивченню сучасних методів та технік прогнозування кредитної спроможності клієнтів банку на основі аналізу фінансових даних. Автори ставлять за мету розглядати ключові аспекти цього процесу, включаючи вибір та обробку релевантних фінансових показників, застосування різноманітних методів аналізу, інтеграцію алгоритмів машинного навчання та оцінку впливу нестандартних даних на точність прогнозів.

Подальший розвиток ефективних стратегій прогнозування кредитної спроможності сприятиме покращенню якості управління кредитним портфелем, зменшенню ризиків та підвищенню фінансової стійкості банківських установ.

2. ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

У сучасному банківському секторі, де висока ступінь конкуренції та ризику вимагає від фінансових установ надзвичайної уважності до управління кредитним портфелем, прогнозування кредитної спроможності клієнтів є стратегічно важливим завданням. За останні роки, з розвитком технологій обробки даних та аналізу, виникає потреба у вдосконаленні методів прогнозування з метою забезпечення надійного управління ризиками.

Основною метою цього дослідження є розробка та впровадження ефективних методів прогнозування кредитної спроможності клієнтів банку на основі аналізу різноманітних фінансових даних. За допомогою сучасних підходів, включаючи алгоритми машинного навчання та статистичні методи, планується розробити моделі, які не лише точно визначатимуть ризик неплатоспроможності, але й дозволять уникнути надмірного відмовлення в кредитах солідним клієнтам.

Задачі дослідження включатимуть:

1. Визначення релевантних фінансових показників: Виділення ключових факторів, що впливають на кредитну спроможність, та їхнє визначення в контексті аналізу даних.
2. Розробка методів обробки та очищення даних: Створення ефективних та надійних методів для обробки фінансових даних, враховуючи можливі аномалії та неповноту інформації.
3. Вибір та оптимізація моделей прогнозування: Порівняння та оптимізація алгоритмів машинного навчання та статистичних методів для побудови точних та ефективних моделей прогнозування.

4. Інтеграція моделей в банківську практику: Розробка механізмів для ефективної інтеграції розроблених моделей в банківські системи та процеси управління кредитами.
5. Оцінка ефективності та надійності: Проведення комплексної оцінки розроблених моделей з метою визначення їхньої точності, чутливості та специфічності у прогнозуванні різних категорій клієнтів.

Це дослідження спрямоване на створення інноваційних підходів до прогнозування кредитної спроможності, які сприятимуть зниженню ризиків та підвищенню ефективності управління кредитним портфелем фінансових установ.

3. ОГЛЯД КЛАСИЧНИХ МЕТОДІВ ПРОГНОЗУВАННЯ КРЕДИТНОЇ СПРОМОЖНОСТІ КЛІЄНТІВ БАНКУ

1. **Multiple Linear Regression:** Метод Multiple Linear Regression (MLR) є класичним статистичним методом, що використовується для вивчення лінійних залежностей між залежною змінною та набором незалежних змінних. У контексті прогнозування кредитної спроможності, MLR дозволяє моделювати взаємозв'язок між різними фінансовими параметрами та ризиком неплатоспроможності.

Перевагами даного підходу є простота та легкість інтерпретації, а також відображення лінійних залежностей між змінними.

Обмеженнями є те, що модель працює краще в умовах лінійності, що може бути недостатнім для складних взаємозв'язків у фінансових даних.

2. **Gradient Boosting Regression:** Gradient Boosting є ансамблевим методом машинного навчання, який покращує прогнози, комбінуючи прості моделі, зосереджуючись на помилках попередніх моделей. У випадку прогнозування кредитної спроможності, Gradient Boosting може враховувати нелінійні та взаємодіючі ефекти між фінансовими параметрами.

Перевагами є те, що даний підхід добре вирішує проблему нелінійних взаємозв'язків та висока точність прогнозів. Обмеженнями є Схильність до перенавчання, що може виникнути при недостатньому контролі параметрів.

3. **Random Forest Regression.** Random Forest є іншим ансамблевим методом, який використовує декілька дерев рішень для отримання прогнозів. Кожне дерево обробляє підвибірки даних та випадковий вибір фіч, що робить модель менш схильною до перенавчання.

Перевагами є висока точність та стійкість до перенавчання, а також здатність обробляти великі набори даних з численними фічами.

Обмежена інтерпретованість порівняно з MLR та затратна обчислювально при великій кількості дерев.

4. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

За результатами дослідження було проведено ряд експериментів, що показали результати, представлені у таблиці 1.

Таблиця 1. Порівняльна таблиця методів: Multiple Linear Regression, Gradient Boosting Regression, та Random Forest Regression

Метрика	Multiple Linear Regression	Gradient Boosting Regression	Random Forest Regression
Середньоквадратична помилка (MSE)	0,2	0,05	0,04
Середня абсолютна помилка (MAE)	0,15	0,08	0,07
Коефіцієнт детермінації (R-squared)	0,3	0,9	0,92

5. ВИСНОВКИ

Multiple Linear Regression підходить для лінійних взаємозв'язків, але обмежений у моделюванні складних нелінійностей. Gradient Boosting Regression та Random Forest Regression мають високу точність, але Gradient Boosting потребує уважного налаштування, а Random Forest – ефективний при обробці великих обсягів даних. Всі методи досягають низьких помилок, проте Gradient Boosting та Random Forest найбільш точні за кількісними метриками. Їхня помилковість та дисперсія показують низький Bias та помірну Variance, вказуючи на бажаний баланс. Вибір методу залежить від конкретних вимог задачі: Multiple Linear Regression для простих сценаріїв, Gradient Boosting та Random Forest – для складних взаємозв'язків.

Загальною тенденцією є те, що для точних та складних прогнозів, особливо в умовах невизначеності, Gradient Boosting та Random Forest можуть бути більш ефективними методами порівняно з Multiple Linear Regression. Однак, враховуючи інтерпретованість та обмежену обчислювальну складність, Multiple Linear Regression може залишатися важливим інструментом у деяких сценаріях.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Christopher Pal, Mark Hall, Eibe Frank, Ian Witten. Data Mining: Practical Machine Learning Tools and Techniques, 4rd ed. / Morgan Kaufmann, 2016.
2. Jason Bell. Machine Learning: Hands-On for Developers and Technical Professionals / John Wiley & Sons, 2014.
3. Дивак М. П. Методичний посібник з дисципліни «Системний аналіз» / М. П. Дивак. – Тернопіль: ТАНГ. – 2004. – 136 с.