Issue 18, part 2, 2025

https://doi.org/10.23939/sisn2025.18.2.163

УДК 004.89:330.43:004.021

ПРОГНОЗУВАННЯ НАПРЯМІВ РОЗВИТКУ IT-РИНКУ З ВИКОРИСТАННЯМ МЕТОДІВ МАШИННОГО НАВЧАННЯ

Ігор Пасемко, Олександр Лозицький

Національний університет «Львівська політехніка»,
кафедра інформаційних систем та мереж, Львів, Україна
E-mail: Ihor.s.Pasemko @lpnu.ua, ORCID: 0009-0000-2055-7170
E-mail: oleksandr.a.lozytskyy@lpnu.ua, ORCID: 0000-0001-8395-8385

У статті досліджено підходи до прогнозування напрямів розвитку IT-ринку на основі методів машинного навчання. Актуальність роботи зумовлена високою динамікою цифрової економіки, швидкими змінами технологічних трендів та потребою у науково обґрунтованих інструментах аналізу IT-сфери. Метою дослідження є побудова моделі прогнозування, здатної виявляти закономірності у соціально-економічних, технологічних та поведінкових показниках, що визначають стан і перспективи розвитку IT-ринку.

У роботі використано методи аналізу часових рядів, кореляційно-регресійного моделювання та машинного навчання – зокрема, алгоритми випадковий ліс, градієнтне бустування і довга короткочасна пам'ять. Для формування вибірки об'єднано макроекономічні індикатори (ВВП, обмінний курс, індекс споживчих цін), показники IT-ринку (рівень заробітної плати, попит на вакансії, обсяг експорту IT-послуг) та поведінкові дані (індекси пошукових запитів Google Trends). Проведено попередню обробку даних, нормалізацію ознак і тестування стаціонарності часових рядів.

Отримані результати засвідчили, що моделі випадковий ліс та довга короткочасна пам'ять забезпечують найвищу точність прогнозу ($R^2 > 0.85$), що дозволяє ефективно виявляти як коротко-, так і середньострокові тенденції розвитку IT-ринку. Встановлено ключові предиктори динаміки ринку – рівень споживчих витрат, середню заробітну плату та індекс популярності IT-запитів.

Практична цінність дослідження полягає у можливості використання запропонованих моделей для підтримки стратегічних рішень у сфері цифрової економіки, планування освітніх і кадрових політик, а також прогнозування технологічних трендів у національному IT-секторі.

Ключові слова – машинне навчання; прогнозування; IT-ринок; часові ряди; економічні індикатори; стаціонарність; сезонно-трендова декомпозиція; випадковий ліс; градієнтне бустування; довга короткочасна пам'ять.

Постановка проблеми

В умовах стрімкого розвитку інформаційних технологій та цифрової трансформації бізнесу, аналітика даних набуває все більшого значення. Одним з ключових напрямків, що активно розвивається, є Data Science − міждисциплінарна галузь, яка поєднує в собі елементи статистики, машинного навчання, аналізу даних, програмування та бізнес-аналітики. Основною метою Data Science є вилучення цінної, дієвої інформації з великих обсягів структурованих і неструктурованих даних, що дозволяє компаніям приймати обґрунтовані управлінські рішення, виявляти тенденції, оптимізувати процеси та досягати конкурентних переваг. Зі зростанням ролі даних у бізнесі, охороні здоров'я, фінансах, освіті та інших

галузях зростає і попит на системних аналітиків. Ці фахівці виконують широкий спектр завдань − від опрацювання первинних даних і побудови аналітичних моделей до впровадження систем прогнозування та автоматизації процедур прийняття рішень. Такий попит призводить до зростання заробітних плат на цьому ринку праці, що, в свою чергу, становить інтерес як для роботодавців, так і для працівників. У цьому контексті прогнозування заробітної плати працівників набуває особливого значення. Для компаній це важливий елемент стратегічного планування. Своєчасна і точна оцінка витрат на персонал дозволяє розраховувати бюджети, адаптувати політику оплати праці, прогнозувати фінансові навантаження і приймати обґрунтовані рішення про наймання персоналу. З іншого боку, для працівників такі прогнози дозволяють оцінити ринкові пропозиції, визначити найкращий час для зміни роботи або підвищення кваліфікації, а також спланувати свої особисті фінанси. Моделювання заробітної плати, зокрема в IT галузі, має також дослідницьке значення. Оскільки заробітна плата є комплексним показником, який залежить від безлічі факторів − досвіду, освіти, місця перебування, економічної ситуації, рівня конкуренції, технічних навичок та галузі − її прогнозування вимагає інтеграції даних з різних джерел. Цей процес вимагає глибокого аналізу часових рядів, використання методів машинного навчання або економетрики, а також врахування зовнішніх макроекономічних факторів. Динамічність ринку праці в IT становить особливий виклик. Зарплати в галузі можуть суттєво змінюватися внаслідок глобальних подій (наприклад, пандемії, війни, економічної кризи), зміни попиту на певні засоби та технології (наприклад, штучний інтелект), політичних рішень (зміни в оподаткуванні, регулюванні фрілансу тощо) та інших факторів. Тому прогнозування має бути не лише точним, але й адаптивним − здатним враховувати зміни у зовнішньому середовищі та оперативно оновлювати оцінки. З огляду на вищезазначене, актуальність теми цього дослідження не викликає сумнівів. Вона має практичне значення для компаній, які прагнуть залишатися конкурентоспроможними на ринку праці

Аналіз останніх досліджень та публікацій

Сучасні дослідження у сфері прогнозування тенденцій розвитку IT-ринку зосереджені на застосуванні методів машинного навчання та аналітики часових рядів для виявлення закономірностей ринкової динаміки, інноваційних циклів і технологічних зсувів. Panchal, Ferdouse і Sultana (2024) порівняли ефективність моделей ARIMA та LSTM для прогнозування цін на акції. У статті підкреслюється, що моделі на основі авторегресивних процесів ефективні для короткострокового прогнозування, проте мають обмеження щодо врахування нелінійних і структурних змін у технологічному середовищі.

З метою подолання цих обмежень дедалі частіше використовуються нейронні мережі глибинного навчання. Так, Hewamalage et al. (2021) довели, що архітектури типу LSTM (Long Short-Term Memory) та GRU (Gated Recurrent Unit) забезпечують точніше моделювання довгострокових тенденцій IT-ринку порівняно з класичними статистичними підходами. Ці моделі здатні враховувати залежності між послідовними періодами, коливання сезонного попиту та вплив глобальних подій (наприклад, цифровізації під час пандемії COVID-19).

Підхід градієнтного бустування дедалі частіше використовується не лише для регресійних, але й для змішаних задач, які поєднують класифікацію та прогнозування кількісних показників. Зокрема, Alwanin et al. (2025) розробили модель одночасної класифікації та регресії на основі Gradient Boosting, що продемонструвала високу узагальнювальну здатність у різних наборах даних. Shah, Bhutia, Manga, Limboo і Rai (2025) відзначають, що сучасні тенденції в ансамблевому навчанні пов’язані з оптимізацією моделей і комбінуванням гібридних алгоритмів. У статті розглянуто інтеграцію методів градієнтного бустингу і випадкового лісу для класифікації напрямів розвитку IT-сектору за категоріями технологій − хмарні обчислення, штучний інтелект, аналітика великих даних, кібербезпека тощо. Такі ансамблеві методи показали високу стійкість до «шумових» даних та ефективність у задачах, пов’язаних із великими багатовимірними наборами ринкових показників (капіталізація, венчурні інвестиції, кількість стартапів, публікаційна активність).

Поряд із моделями машинного навчання розвиваються гібридні підходи, у яких поєднуються ARIMA/SARIMA для моделювання трендових компонентів, і нейронні мережі (LSTM, GRU) для опису нелінійних залежностей. Гібридні підходи, що поєднують статистичні та нейронні моделі, демонструють високу адаптивність у задачах прогнозування часових рядів. Зокрема, Fatima et al. (2025) показали, що об'єднання ARIMA та LSTM підвищує точність прогнозів продажів завдяки поєднанню короткострокових трендів і довготривалих залежностей у даних.

Окремий напрям формують дослідження у сфері Data Mining та прогнозної аналітики бізнесу (Business Intelligence). Концепція прогнозної аналітики, що поєднує методи статистики, машинного навчання та бізнес-аналізу, була вперше детально описана в роботі Elkan (2010). Цей підхід лежить в основі сучасних систем опрацювання даних і моделювання ринкових тенденцій.

Bustos, Pomares-Quimbaya і Stellian (2025) порівняли різні алгоритми машинного навчання для прогнозування динаміки фондового ринку, підкреслюючи їхній вплив на ефективність ринку. Автори відзначають, що поєднання методів кластеризації з аналізом часових рядів дозволяє виокремлювати нові сегменти IT-ринку й оцінювати потенціал інноваційних ніш. Застосування нейронних мереж та трансформерів розширює можливості багатофакторного прогнозування, де враховується не лише динаміка фінансових індикаторів, а й текстова аналітика новин, патентів та публікацій у відкритих джерелах.

Сучасний IT-ринок характеризується високими темпами розвитку, глобальною конкуренцією та постійною трансформацією технологічних трендів. Згідно з аналітичним звітом Gartner (2024), середній цикл оновлення ключових технологій у сфері інформаційних технологій скоротився з 4,2 року у 2015 р. до 2,7 року у 2024 р., що свідчить про підвищення динамічності ринку та зростання складності його прогнозування. За таких умов традиційні економетричні методи, зокрема ARIMA, SARIMA або лінійна регресія, демонструють обмежену здатність адекватно відтворювати нелінійні закономірності у даних, що визначають розвиток IT-індустрії (Box et al., 2016).

Проблема полягає у тому, що класичні підходи не враховують комплексну взаємодію численних чинників, зокрема обсягів венчурних інвестицій, частки використання хмарних сервісів, швидкості впровадження штучного інтелекту, геополітичних ризиків тощо. На відміну від них, моделі машинного навчання (ML) − зокрема ансамблеві методи (Random Forest, XGBoost) та рекурентні нейронні мережі (LSTM) − здатні виявляти приховані закономірності й автоматично адаптуватися до змін структури даних (Goodfellow et al., 2016; Chollet & Allaire, 2022).

Актуальність теми зумовлена стрімким скороченням життєвого циклу технологічних трендів на IT-ринку та потребою у точному, адаптивному та інтелектуальному прогнозуванні на основі великих даних.

Проведений аналіз засвідчує перехід від традиційних статистичних моделей до інтелектуальних систем прогнозування, заснованих на глибинному навчанні, ансамблевих методах і гібридних архітектурах. Водночас залишається актуальним завдання інтеграції прогнозних моделей із реальними потоками ринкових даних (API аналітичних платформ, відкриті бази Crunchbase, Kaggle, CB Insights), що дозволить підвищити точність і адаптивність прогнозів у динамічному середовищі IT-ринку.

Таким чином, застосування методів машинного навчання для прогнозування напрямів розвитку IT-ринку не лише підвищує точність аналітичних оцінок, а й формує основу для прийняття обґрунтованих управлінських рішень у сфері цифрової економіки.

Формулювання цілі статті

Метою дослідження є розроблення науково обґрунтованого підходу, здатного з високою точністю передбачати розмір середньої заробітної плати в IT галузі на основі доступних економічних індикаторів та характеристик працівника. Модель повинна бути стійкою до шуму в даних, здатною узагальнювати на нові спостереження та забезпечувати інтерпретовані результати для практичного застосування.

Досягнення поставленої мети передбачає вирішення таких завдань дослідження:

1. Проаналізувати наукові підходи до моделювання та прогнозування технологічних трендів IT-ринку, визначивши переваги й обмеження класичних статистичних та сучасних нейромережевих методів.
2. Дослідити можливості застосування ансамблевих та гібридних моделей (ARIMA, SARIMA, Random Forest, Gradient Boosting, LSTM, GRU) для опису часової динаміки ринкових показників.
3. Розробити архітектуру прогнозної системи, яка поєднує алгоритми машинного навчання з аналітичними потоками відкритих ринкових даних (API-платформи, бази Crunchbase, Kaggle, CB Insights).
4. Побудувати експериментальну модель прогнозування з використанням часових рядів ключових індикаторів IT-сектору.
5. Оцінити точність і стійкість розробленої моделі.

Виклад основного матеріалу

Гібридна прогнозна система

Моделювання заробітної плати IT-спеціалістів на основі часових рядів є багатогранною задачею, що потребує інтеграції традиційних статистичних методів із методами машинного навчання. Аналіз наукової літератури свідчить про еволюцію методологічних підходів від класичних моделей часових рядів до складних гібридних методів.

Класичні підходи до прогнозування заробітної плати базуються на методах аналізу часових рядів. Авторегресивна інтегрована модель ковзного середнього (AutoRegressive Integrated Moving Average, далі ARIMA) залишається зручним інструментом для прогнозування економічних показників завдяки своїй здатності виявляти та моделювати автокореляційні структури у даних. Розширена модель сезонної авторегресивної інтегрованої моделі ковзного середнього (Seasonal ARIMA, далі SARIMA) використовується для врахування сезонних коливань, що особливо актуально для IT-ринку з його циклічними трендами наймання та встановлення оплати. Модель позначається як:

$$\mathrm{SARIMA}(p, d, q) \times (P, D, Q)_s \quad (1)$$

де p, d, q − параметри звичайної ARIMA;P, D, Q − сезонні параметри;s − довжина сезонного циклу (наприклад, 12 для місячних даних із річною сезонністю).

Сезонна авторегресивна інтегрована модель ковзного середнього − статистична модель, що поєднує авторегресію, різницювання, ковзне середнє та сезонні компоненти для прогнозування часових рядів із регулярною періодичністю. Однак ці методи обмежені лінійністю припущень та неспроможністю враховувати складні взаємодії між різними факторами.

Сучасні методи машинного навчання суттєво розширили можливості прогнозування. Ансамблеві моделі, такі як випадковий ліс та градієнтний бустинг, демонструють високу ефективність у роботі з багатовимірними наборами даних, дозволяючи виявляти нелінійні залежності між предикторами. Нейронні мережі, зокрема мережі з довгою короткостроковою пам’яттю (Long Short-Term Memory, далі LSTM) та рекурентні нейронні архітектури з керованими воротами (Gated Recurrent Unit, далі GRU), показують особливу ефективність у моделюванні довгострокових залежностей у часових рядах, що критично важливо для прогнозування динаміки заробітної плати в умовах швидкозмінного технологічного ландшафту.

Запропонована гібридна система прогнозування призначена для підвищення точності прогнозування тенденцій IT-ринку за рахунок поєднання класичних статистичних моделей часових рядів і сучасних алгоритмів машинного навчання. Система реалізує багаторівневий конвеєр опрацювання даних, у межах якого SARIMA відповідає за моделювання лінійних трендів і сезонності, тоді як модуль машинного навчання (XGBoost або LSTM) моделює нелінійні взаємозв’язки між залишковими компонентами.

Для подолання обмежень лінійних моделей у роботі запропоновано гібридний конвеєр опрацювання даних, який поєднує класичну модель SARIMA з алгоритмами машинного навчання. Такий підхід дозволяє одночасно моделювати лінійну сезонність і нелінійні залежності між показниками.

Конвеєр опрацювання даних описується формулою:

$$\widehat{y_t}=\hat{y}_t^{(\mathrm{SARIMA}),}+\mathrm{fML}(r_t),\ r_t=y_t-\hat{y}_t^{(\mathrm{SARIMA})}, \tag{2}$$

де fML − нелінійна функція, апроксимована алгоритмом машинного навчання; $\hat{y}_t^{(\mathrm{SARIMA}),}$ − прогноз сезонно-трендової складової; r_t − залишкова компонента, що містить нелінійні впливи.

Перший етап конвеєра усуває тренд і сезонність за допомогою $SARIMA(p,d,q)\times(\mathrm{P},\mathrm{D},\mathrm{Q})_s$, другий − моделює нелінійні залишкові залежності через XGBoost або LSTM. Отримані часткові прогнози об'єднуються за принципом адаптивного ансамблювання.

Для підвищення стійкості результатів застосовується адаптивне ансамблювання, у межах якого ваги окремих моделей змінюються пропорційно до їхньої поточної точності. Ваги окремих моделей визначаються пропорційно до їхньої поточної точності на валідаційній підвибірці:

$$w_i(t)=\frac{1/\mathrm{RMSE}_i(t)}{\sum_{j=1}^{k}1/\mathrm{RMSE}_i(t)} \tag{3}$$

де $w_i(t)$ − вага і-тої моделі у момент часу t; $\mathrm{RMSE}_i(t)$ − середньоквадратична помилка на останньому вікні. Фінальний прогноз:

$$\widehat{y_t}=\sum_{i=1}^{k}w_i\,(t)\widehat{y_{i,t}}. \tag{4}$$

У табл.1 продемонстровано вплив окремих економічних і поведінкових чинників на прогнозування динаміки заробітної плати в IT-галузі. Ваги ознак отримані за результатами моделі XGBoost у складі гібридного конвеєра опрацювання даних SARIMA+ML.

Таблиця 1

Важливості ознак у гібридній моделі прогнозування

Ознака	Вага у моделі, %	Інтерпретація
GDP_USD	21,4	Відображає загальний стан економіки
CPI	17,8	Вплив інфляційних очікувань
Exchange_rate	15,2	Вплив валютних коливань
IT_export	12,5	Динаміка експорту IT-послуг
Google_trends_index	11,0	Цифрова активність, попит на спеціалістів
Consumer_spending	9,6	Відображає рівень внутрішнього попиту
Average_salary_UA	7,4	Загальна динаміка заробітних плат у країні
Inflation_expectations	5,1	Вплив макроекономічних очікувань

Джерело: розраховано авторами на основі даних CB Insights (2025), Gartner (2024) та World Bank (2024))

Компоненти системи:

Input – вхідні економічні та поведінкові показники (ВВП, курс валют, CPI, експорт IT, Google Trends).

SARIMA – модуль моделювання лінійної сезонності та трендів..

Residuals – обчислення залишків між фактичними та прогнозними значеннями SARIMA.

ML – модуль машинного навчання (XGBoost або LSTM) для моделювання нелінійних впливів.

Ensemble – адаптивне зважування часткових прогнозів.

Output – фінальний прогноз динаміки IT-показників.

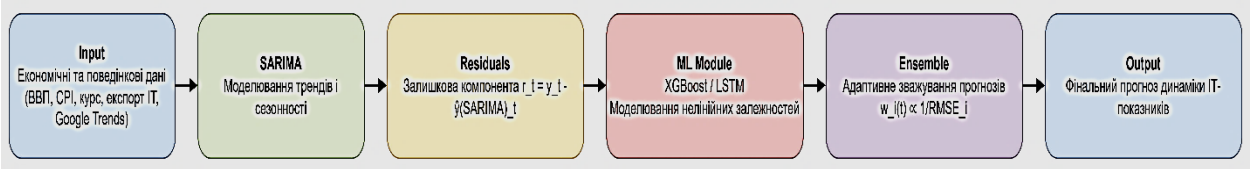

Рис. 1. Архітектура гібридної прогнозної системи SARIMA+ML для аналізу динаміки IT-ринку України

Підхід, запропонований у даній роботі, відрізняється декількома ключовими новаціями. По-перше, застосовується гібридний конвеєр опрацювання даних, що поєднує переваги класичних методів часових рядів із сучасними алгоритмами машинного навчання, дозволяючи одночасно враховувати як лінійні тренди, так і складні нелінійні взаємодії. По-друге, до аналізу включаються специфічні для українського ринку економічні чинники, такі як динаміка IT-експорту, геополітичні ризики та валютні коливання, що раніше не розглядались у комплексному контексті. По-третє, використовується ансамблеве моделювання з адаптивним зважуванням, що дозволяє динамічно коригувати важливість різних компонентів моделі залежно від ринкових умов. Така методологія забезпечує більшу стійкість до структурних змін в економіці та підвищує адаптивність моделі до нових ринкових реалій.

Передбачення заробітної плати працівників IT галузі є комплексною аналітичною задачею, що знаходиться на перетині економетрики, статистичного моделювання та машинного навчання. Заробітна плата як економічна категорія формується під впливом численних макроекономічних та мікроекономічних чинників, включаючи рівень кваліфікації працівника, загальноекономічні показники країни, монетарну політику, ринкові тенденції та інші фактори.

Складність задачі передбачення заробітної плати обумовлена нелінійними взаємозв'язками між економічними показниками, наявністю часових залежностей, впливом зовнішніх чинників та структурних змін в IT галузі. Традиційні економетричні підходи часто виявляються недостатніми для адекватного моделювання таких складних залежностей, особливо в умовах високої волатильності економічних показників.

Методи машинного навчання надають потужний інструментарій для вирішення подібних задач завдяки здатності автоматично виявляти приховані патерни в даних, опрацьовувати нелінійні залежності та адаптуватися до змін у структурі даних. Застосування методів машинного навчання дозволяє створювати більш точні та стійкі прогностичні моделі, що враховують багатофакторну природу формування заробітної плати в IT галузі.

Задача передбачення заробітної плати належить до класу задач регресії з наглядовим навчанням. Вхідними даними служить набір спостережень, кожне з яких характеризується сукупністю економічних та демографічних ознак. Цільовою змінною виступає середня заробітна плата працівників за відповідний період. Сутність задачі полягає у виявленні закономірностей та залежностей між набором вхідних характеристик та розміром заробітної плати. Методи машинного навчання повинні автоматично ідентифікувати патерни в історичних даних, що дозволить створити прогностичну модель для передбачення заробітної плати в нових умовах.

Ключовою особливістю задачі є необхідність врахування часової динаміки різних чинників та їх взаємного впливу на формування рівня заробітної плати. Модель має враховувати як прямі залежності між окремими факторами, так і складні нелінійні взаємодії між групами економічних індикаторів.

Класичний набір даних містить 16 змінних, що характеризують економічний стан IT галузі України за різні часові періоди. Дані структуровані у вигляді часових рядів з періодичністю півріччя, що дозволяє відслідковувати динамічні зміни економічних показників. Проаналізуємо основні з них.

Категоріальна змінна «рівень кваліфікації» відображає кваліфікаційний рівень працівника з трьома можливими значеннями junior, middle та senior. Ця змінна відображає професійну ієрархію та рівень експертизи, що безпосередньо впливає на розмір заробітної плати. Для моделювання необхідне застосування методів кодування категоріальних змінних. Для моделювання використовувались

ключові економічні, монетарні та поведінкові індикатори, що впливають на рівень заробітної плати IT-фахівців. Категоріальна змінна *«рівень кваліфікації»* (junior, middle, senior) відображає професійну ієрархію та рівень експертизи працівників. До макроекономічних показників віднесено *валовий внутрішній продукт (у доларовому еквіваленті)* та *його відносну зміну (%ΔGDP)*, які характеризують загальний розвиток IT-галузі. Монетарний блок включає *обмінний курс гривні до долара США* та *індекс споживчих цін (CPI)*, що відображають вплив інфляційних процесів і валютної стабільності. До зарплатних показників належать *середня заробітна плата у доларах США* та *її відносна зміна (%ΔSalary)*, які дозволяють оцінити динаміку реальних доходів незалежно від девальваційних чинників. Для врахування поведінкових аспектів використано *індекс популярності запиту «Data Science» у Google Trends*, який відображає цифрову активність та попит на IT-спеціалістів в Україні.

Для оцінювання якості регресійних моделей обрано три комплементарні метрики, що забезпечують всебічну характеристику точності передбачень.

Середня абсолютна помилка (MAE) характеризує середнє відхилення передбачених значень від фактичних в абсолютному вираженні:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Перевагою MAE є стійкість до викидів та інтерпретованість у вихідних одиницях вимірювання. Метрика забезпечує рівномірне зважування всіх помилок незалежно від їх величини.

Середньоквадратична помилка (RMSE) надає більшу вагу великим помилкам через квадратичну функцію втрат:

$$\text{RMSE} = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Це особливо важливо для задач, де великі помилки передбачення мають критичне значення. RMSE також виражається в тих самих одиницях, що й цільова змінна.

Коефіцієнт детермінації (R^2) показує частку дисперсії цільової змінної, що пояснюється моделлю:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

де $SS_{\text{res}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$- сума квадратів залишків, тобто відхилення фактичних значень від прогнозованих; $SS_{\text{tot}} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ - загальна сума квадратів, тобто відхилення фактичних значень від їх середнього; y_i— фактичне значення; $\hat{y}_i$ - прогнозоване значення моделі; $\bar{y}$ - середнє значення залежної змінної; n — кількість спостережень.

R^2 є безрозмірною метрикою зі значеннями від 0 до 1, де значення близьке до 1 свідчить про високу якість моделі. Метрика дозволяє порівнювати моделі незалежно від масштабу даних.

Для вирішення задачі регресії попередньо розглядалися базові моделі – лінійна регресія, дерево рішень і випадковий ліс. Порівняльний аналіз показав, що алгоритм XGBoost (eXtreme Gradient Boosting) забезпечує найкращий баланс між точністю, стійкістю та швидкістю обчислень. XGBoost представляє сучасну реалізацію градієнтного бустингу з численними оптимізаціями. Алгоритм послідовно будує ансамбль слабких учнів, де кожна наступна модель виправляє помилки попередніх. XGBoost включає вбудовану L1 та L2 регуляризацію для запобігання перенавчанню, ефективне опрацювання пропущених значень та автоматичну оптимізацію гіперпараметрів. Алгоритм демонструє високу продуктивність на структурованих даних та здатність моделювати складні нелінійні залежності.

Результати реалізації моделі

Перший етап опрацювання даних полягав у автоматичному визначенні структури CSV-файлів, які можуть бути створені з різних джерел. Для цього використано алгоритм Sniffer з модуля csv у Python, який перевіряє перші 1024 байти файлу та намагається визначити, який роздільник (наприклад, кома або крапка з комою) використовується. Якщо визначити формат не вдається, система підставляє запасні варіанти роздільника, щоб уникнути помилок. Вибір обсягу 1024 байти − це баланс між швидкістю обробки та точністю.

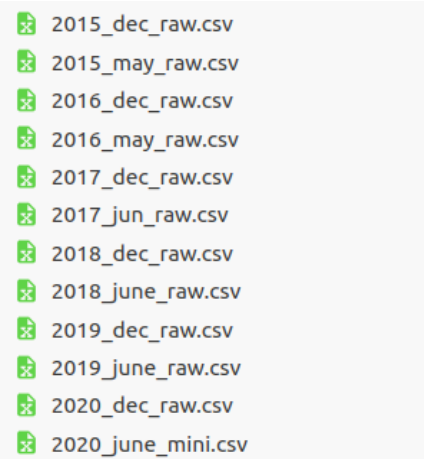

Рис. 2. Приклади назв файлів, у яких збережено зібрані дані

Щоб зрозуміти, за який час зібрані дані, аналізувалися назви файлів (див. рис. 1) за допомогою регулярних виразів. Дані збирались по півріччях, щоб зменшити вплив сезонних коливань і зробити результати більш надійними. Це дозволило краще бачити загальні тенденції, а не випадкові стрибки.

Оскільки дані зібрані з різних джерел, одні й ті самі речі можуть називатися по-різному. Тому було створено словник відповідностей. 45 назв було зведено до 5 основних колонок: зарплата, посада, рівень досвіду, стаж і частота відповідей, а в подальшому звели до 3 колонок: зарплата, посада, рівень досвіду. Це дозволяє об'єднувати різні таблиці в одну без плутанини. Показник *«частота відповідей»* є похідною змінною, що обчислюється як кількість уніфікованих спостережень у кожній категорії (посада × рівень досвіду, за потреби − з розбиттям за періодами).

Він використовувався для зважування навчальної вибірки та контролю якості (відсікання рідкісних груп), але не увійшов до фінального набору предикторів, тому у підсумковому датасеті залишено три змістові колонки: зарплата, посада, рівень досвіду.

В аналіз включалися лише ті записи, де йшлося про фахівців із Data Science. Для цього використовували пошук у тексті − наприклад, "Data Scientist" або "Data Science".

Було створено три групи за рівнем досвіду або стажу роботи:

- junior − 0–2 роки,
- middle − 3–5 років,
- senior − 6+ років або керівні посади.

Якщо чітко не вказувався рівень, використовувався стаж або досвід на поточному місці роботи.

Значення заробітної плати у вихідних даних мали різні формати подання. Вони подавалися у вигляді діапазонів (наприклад, 1500–2500), чисел із символом «+» (наприклад, 2000+), а також числових значень із комами або крапками як роздільниками тисяч чи десяткових дробів. Для уніфікації всіх значень було виконано нормалізацію даних:

- для діапазонів заробітної плати обчислювалося середнє значення;
- для записів із символом «+» додавалася умовна надбавка;
- числові формати було приведено до єдиного стандарту (кома або крапка як роздільник);
- додатково враховувалися бонуси − місячні, квартальні та річні.

Іноді в даних могла бути суперечність − наприклад, у людини вказаний рівень senior, а зарплата нижча, ніж у середнього middle. Такі випадки виправлялися на основі середніх зарплат для кожної групи. Це підвищило якість даних і зменшило кількість помилок.

До даних були додані економічні показники − інфляція, середня зарплата по країні, ВВП, споживчі витрати. Зібрані щорічні дані використовувались для кожного з півріччя року. Також був доданий показник інтересу до професії − наскільки часто здійснювався пошук у Google за ключовим словом "data science". Для забезпечення коректного часово-серійного аналізу даних колонку, що відповідає за часовий період (наприклад, місяць або півріччя), вважали за доцільне встановити як індекс датафрейму. Це дало змогу зручно працювати з часовими рядами, наприклад, будувати графіки динаміки або обчислювати середні значення за певні періоди.

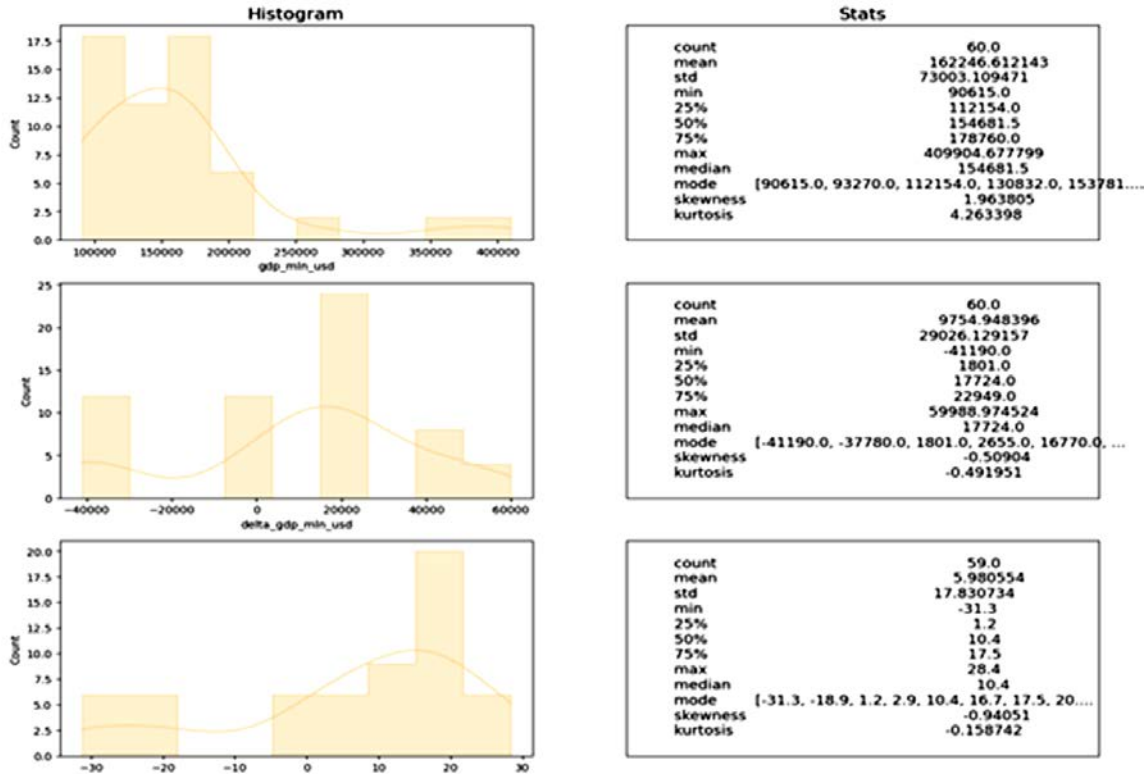

Рис. 3. Порівняння фактичних і прогнозованих значень заробітних плат за рівнями кваліфікації

Під час первинного аналізу числових ознак (див. рис. 3, 4, 5) було виявлено, що деякі з них мають виражену асиметрію (значення розподілені нерівномірно відносно середнього) та підвищену куртозу (наявність викидів або «довгих хвостів» у розподілі). Це означає, що дані значно відрізняються від нормального розподілу. Через це стандартні статистичні методи, які припускають нормальність (наприклад, лінійна регресія), можуть працювати неточно. Натомість краще застосовувати моделі, що не залежать від розподілу даних. Одним із таких варіантів є XGBoost – потужний алгоритм на основі дерев рішень, який добре справляється з нерівномірно розподіленими ознаками.

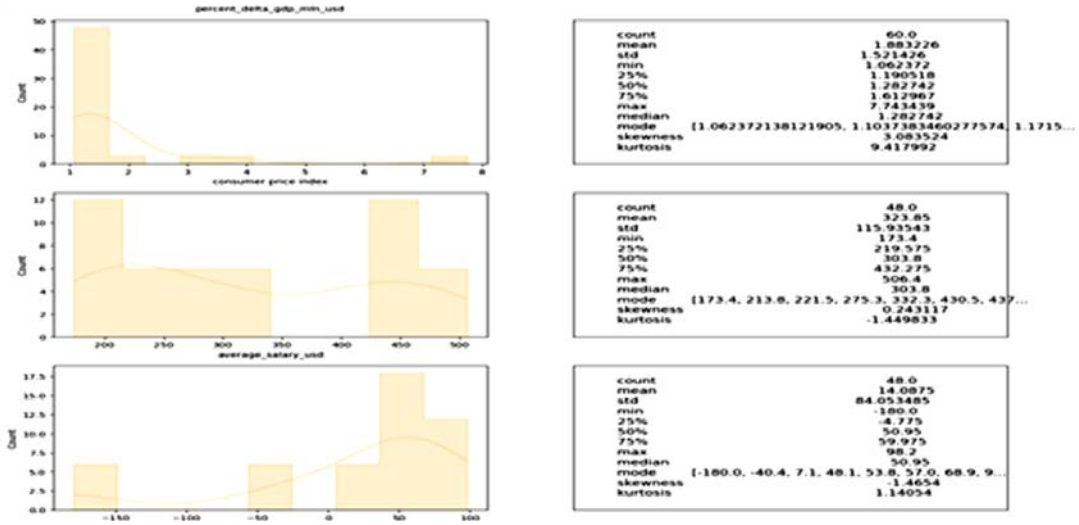

Рис. 4. Порівняння фактичних і прогнозованих значень заробітних плат за рівнями кваліфікації

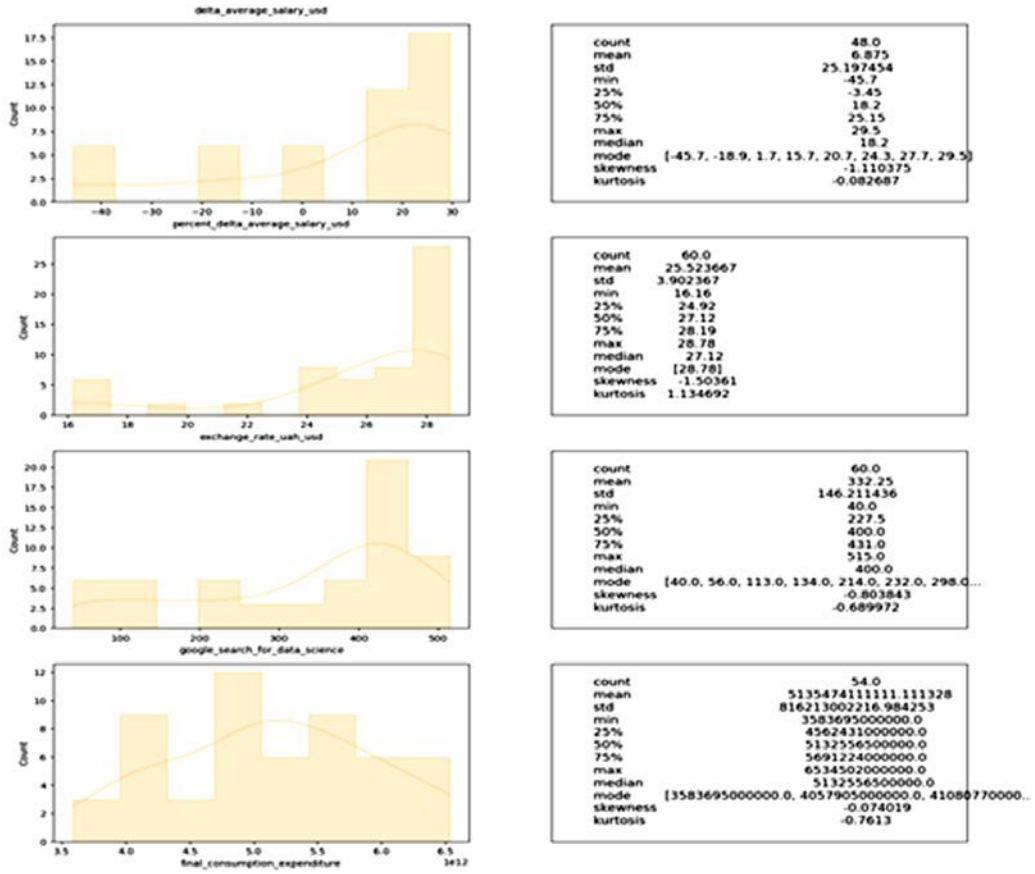

Рис. 5. Порівняння фактичних і прогнозованих значень заробітних плат за рівнями кваліфікації

Як видно з рис. 2–4, що подають порівняння фактичних і прогнозованих значень заробітних плат за рівнями кваліфікації, прогноз гібридної моделі відтворює переломні моменти 2022–2023 рр., демонструючи кращу адаптацію до структурних змін порівняно з базовою SARIMA. На цих рисунках подано розподіли та описову статистику кількох показників, що використовувались у моделі прогнозування (гібридній SARIMA+ML). Ліва частина − графіки розподілу (histogram + KDE). По осі X − значення відповідного показника, по осі Y − кількість спостережень (*Count*), тобто, скільки разів значення цього показника трапляється в даних.

У правій частині таблиці описових статистик для кожного показника наведено *count* − кількість спостережень; *mean* − середнє значення; *std* − стандартне відхилення; *min, 25 %, 50 %, 75 %, max* − квартилі (медіана − 50 %); *mode* − мода (найпоширеніше значення); *skewness* − асиметрія (знак показує напрям "хвоста"); *kurtosis* − ексцес (гостровершинність розподілу).

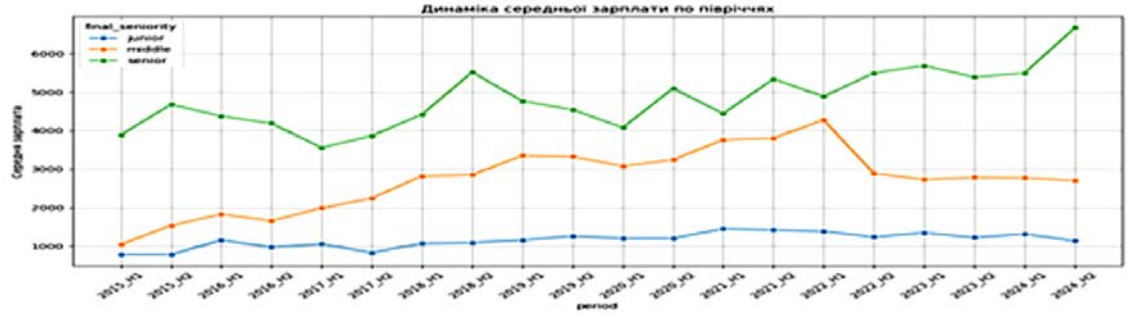

Рис. 6. Динаміка середньої зарплати по півріччях

На рис. 6 наведено розподіли ключових незалежних змінних, що були використані у моделі прогнозування, зокрема *зміна середньої заробітної плати, відносна зміна середньої зарплати (%ΔSalary), обмінний курс гривні до долара США, індекс пошукових запитів "data science" у Google Trends та кінцеві споживчі витрати.* На горизонтальній осі кожного графіка відкладено значення відповідного показника, на вертикальній − кількість спостережень у вибірці. Аналіз показує, що більшість змінних мають виражену правосторонню асиметрію (skewness > 0), тобто характерні одиничні періоди із різкими стрибками показників. Зокрема, розподіл зміни середньої зарплати має два піки – навколо невеликих негативних значень (періоди спаду) та приростів у діапазоні +25–30 %, що відображає кризові коливання 2022 року. Курс гривні до долара зосереджений у межах 26–28 грн/USD, з поодинокими викидами до 40 грн, що відповідає періодам девальваційних коливань. Індекс Google Trends для запиту "data science" варіює у межах 100–500 пунктів, що свідчить про поступове зростання інтересу до тематики машинного навчання серед українських користувачів. Показник кінцевих споживчих витрат має розподіл із плавним зміщенням праворуч, що підтверджує стабільне зростання внутрішнього попиту. Аналіз динаміки середньої зарплати за півріччями (рис. 6) показує, що у 2022 році спостерігається переломний момент. Для рівнів junior і middle темпи зростання заробітних плат сповільнюються або взагалі зупиняються. У той самий час для senior рівня продовжується стабільне зростання. Це може свідчити про те, що попит на досвідчених фахівців залишається високим навіть у складні періоди, тоді як менш досвідчені працівники більше залежать від зовнішніх умов на ринку праці.

Таким чином, аналіз розподілів виявив відмінності у масштабах і варіаціях показників, що підтверджує доцільність їх нормалізації або логарифмування перед навчанням моделі.

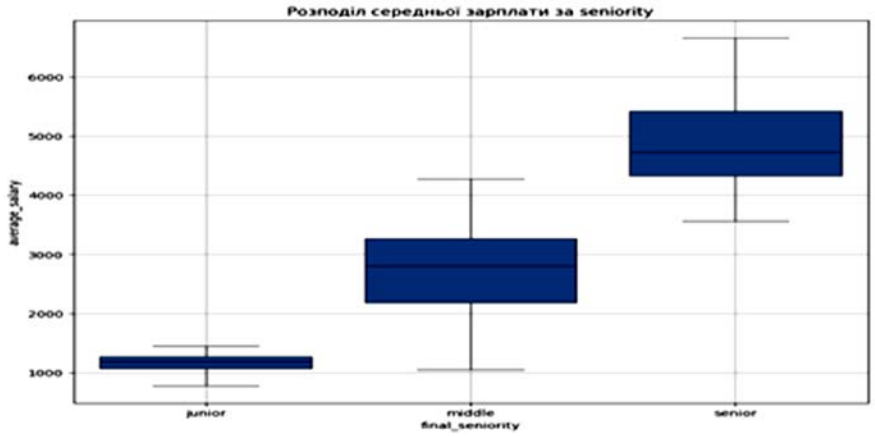

Рис. 7. Розподіл середньої зарплати за seniority у динаміці за період ***2020–2024 рр з кроком агрегування – 1 місяць***

Рис. 7 демонструє динаміку та структуру заробітних плат IT-фахівців різних рівнів кваліфікації протягом 2020–2024 рр**.** Крок агрегування − 1 місяць, що дозволяє відстежувати середньострокові зміни без надмірного коливання показників. На осі X подано часову шкалу (місяці або квартали), на осі Y − середню заробітну плату у доларах США. Дані відображають усереднені значення по вибірках опитувань та HR-звітів, агреговані з місячною частотою. Аналіз показує чітке розшарування заробітних плат за рівнями seniority. Середні зарплати junior-фахівців коливаються у межах 600–900 USD з помітним падінням у II–III кв. 2022 р. Middle рівень зберігає стабільність у діапазоні 1800–2500 USD, із короткостроковим зниженням під час воєнних подій, senior рівень демонструє найбільшу стійкість − середня зарплата перевищує 3500 USD, із подальшим поступовим зростанням у 2023–2024 рр. Помітно, що після 2022 року криві для всіх трьох рівнів поступово відновлюють зростання, що свідчить про адаптацію IT-ринку України до нових економічних умов та відновлення активності експорту IT-послуг. Різниця між рівнями залишається значною, однак із тенденцією до зближення у 2024 р., що може бути пов'язано зі зростанням попиту на middle-

спеціалістів і стабілізацією ринку праці. Графік boxplot демонструє, як змінюється розподіл середньої зарплати в залежності від рівня спеціаліста. Можна зробити два важливих висновки. Із підвищенням кваліфікації зростає як середня зарплата, так і розкид значень. Найбільший розкид зарплат спостерігається на рівні senior, що означає різні умови оплати для досвідчених фахівців. Для рівня junior зарплати більш однорідні, тобто їх легше передбачити. Ці особливості важливо враховувати при побудові моделей прогнозування зарплат або автоматичного визначення рівня спеціаліста за даними.

Комбінований аналіз макроекономічних, поведінкових та зарплатних індикаторів показав, що вибірка, яка охоплює 60 місяців (2020–2024 рр.), дає змогу виявляти середньострокові тренди. Більшість показників мають нелінійні, асиметричні розподіли, тому застосування гібридного підходу (SARIMA + XGBoost) є обґрунтованим. Рівень заробітної плати зростає пропорційно seniority, але темпи приросту після 2022 р. помітно відрізняються між групами, що вказує на структурні зміни ринку праці.

Перед побудовою прогнозних моделей перевірено, чи є часові ряди стаціонарними. Стаціонарний ряд характеризується постійними статистичними властивостями в часі: середнє значення, дисперсія та кореляції не змінюються.

Нестаціонарні ряди можуть призводити до хибних висновків при моделюванні, оскільки їхні статистичні характеристики змінюються з часом. Це робить параметри моделі нестабільними та знижує якість прогнозів.

Для надійної оцінки стаціонарності застосовано два статистичні тести:

Тест Дікі-Фуллера (ADF) дозволяє перевіряти гіпотезу про наявність тренду в даних. Якщо тест показує наявність тренду, то ряд є нестаціонарним. Тест базується на аналізі коефіцієнтів авторегресійного рівняння.

Тест Квятковського–Філліпса–Шмідта–Шина (KPSS) використовується для перевірки стаціонарності часових рядів. На відміну від тесту Дікі–Фуллера (ADF), який перевіряє наявність одиничного кореня, тест KPSS має нульову гіпотезу про стаціонарність, що дозволяє застосовувати його як додаткову перевірку стабільності процесу. Використання двох тестів з різними підходами забезпечує більш надійні висновки про властивості часових рядів. Якщо результати обох тестів узгоджуються, це підтверджує достовірність висновку.

Проведені статистичні тести (див. рис. 8) показали, що досліджувані часові ряди є нестаціонарними. Тест Дікі–Фуллера (ADF) виявив наявність тренду, тоді як тест Квятковського–Філліпса–Шмідта–Шина (KPSS) підтвердив відсутність стаціонарності. Узгоджені результати обох тестів однозначно свідчать про нестаціонарний характер вихідних даних. Для усунення нестаціонарності застосовано диференціювання першого порядку − стандартний підхід до перетворення нестаціонарних часових рядів, який полягає в обчисленні різниць між сусідніми спостереженнями для усунення тренду. Після диференціювання (див. рис. 9) повторна перевірка підтвердила досягнення стабільності. Перетворені ряди демонструють стабільні статистичні характеристики (див. рис. 10).

```
--- JUNIOR ---
ADF p-value: 0.0003819759721485734
KPSS p-value: 0.022553821061996163

--- MIDDLE ---
ADF p-value: 0.9243427088444492
KPSS p-value: 0.04892354629097834 4

--- SENIOR ---
ADF p-value: 0.9974812993511674
KPSS p-value: 0.01
```

Рис. 8. Результати тестування стаціонарності часових рядів

На рис. 8 подано результати тестів стаціонарності рядів заробітних плат для трьох рівнів кваліфікації (junior, middle, senior). Тест Дікі–Фуллера (ADF) показав, що лише ряд junior має $p < 0.05$, тобто є квазі-стаціонарним після усунення тренду. Для middle і senior ADF > 0.05, а тест KPSS < 0.05, що підтверджує наявність трендового компонента та одиничного кореня. Отже, базові

ряди мають нестаціонарну природу, що обґрунтовує необхідність попередньої диференціації та сезонного згладжування перед моделюванням.

```
=== JUNIOR (DIFFERENCED) ===
ADF p-value: 3.4967839938782357e-09
KPSS p-value: 0.041666666666666755

=== MIDDLE (DIFFERENCED) ===
ADF p-value: 0.9989251337131194
KPSS p-value: 0.1

=== SENIOR (DIFFERENCED) ===
ADF p-value: 9.53986949698269e-07
KPSS p-value: 0.1
```

Рис. 9. Результати тестів на стаціонарність часових рядів після диференціювання першого порядку.

На рис. 9 подано результати тестів стаціонарності після застосування першої різниці (d=1). Для рядів Junior і Senior тест Дікі–Фуллера (ADF) показав $p < 0.05$, а тест KPSS – $p \approx 0.1$, що підтверджує досягнення стаціонарності після першого диференціювання. Для Middle ADF > 0.05, отже ряд залишається трендовим; необхідно застосувати додаткову сезонну різницю (D=1, s=12). Таким чином, більшість рядів можна вважати стаціонарними у слабкому сенсі, що є передумовою для побудови адекватних моделей SARIMA та їх гібридизації з ML-методами.

Отримані диференційовані ряди (рис.10) свідчать про досягнення стаціонарності для більшості рівнів кваліфікації. Для junior і senior коливання мають випадковий характер, тоді як для middle спостерігається одиничний структурний злам у 2022 р., після якого ряд повертається до стабільної поведінки.Таким чином, перша різниця (d=1) є достатньою для стабілізації часових рядів заробітних плат перед побудовою моделі SARIMA + XGBoost.

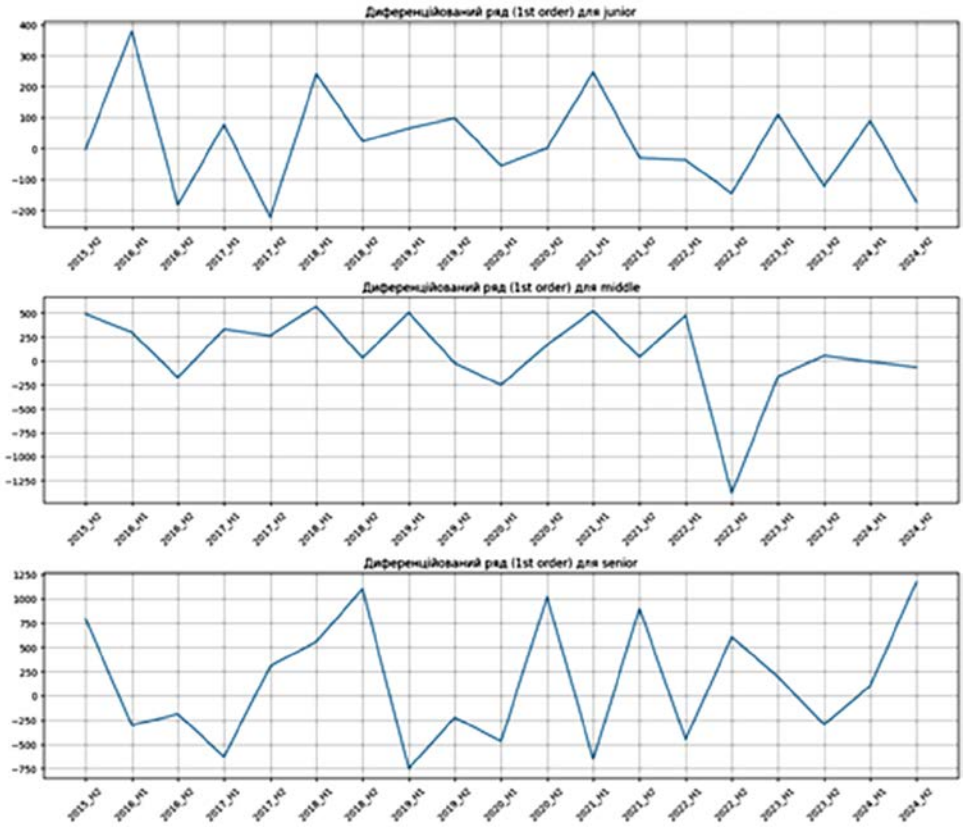

Рис.10. Диференційовані часові ряди

Декомпозиція часових рядів середньої заробітної плати за рівнями кваліфікації дозволила виявити основні закономірності динаміки для кожної професійної категорії.

Тренд для висококваліфікованих фахівців (рис. 11) свідчить про стабільне зростання показника впродовж аналізованого періоду. Після короткого періоду зниження у 2016-2017 роках спостерігається постійне підвищення зарплат, яке особливо активізувалося з 2018 року. Починаючи з 2021 року, тренд поступово вирівнюється, що свідчить про стабілізацію показника. Сезонна складова має великі коливання (±200 одиниць). Залишкова складова показує помірні відхилення без систематичних змін, що підтверджує стабільність ринку праці для цього сегменту.

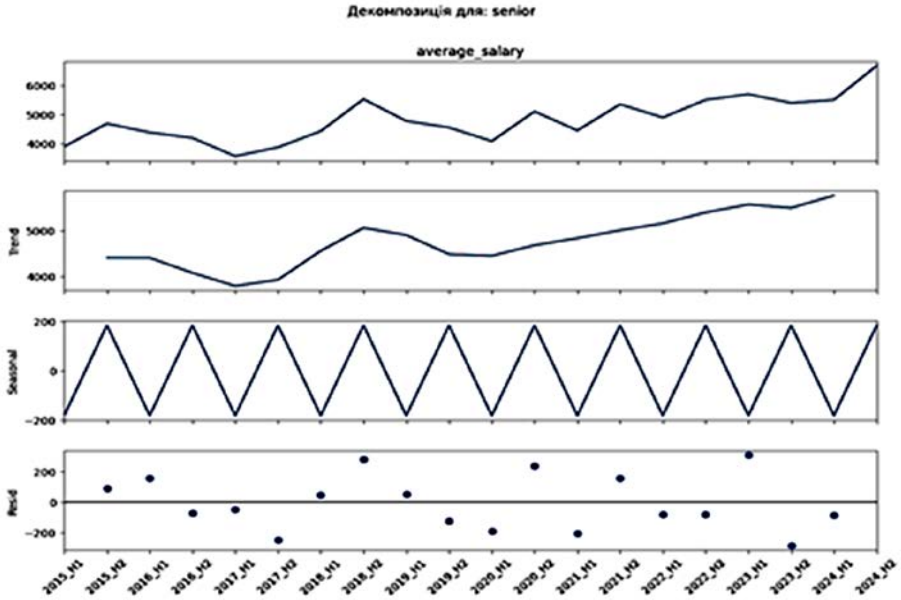

Рис. 11. Декомпозиція часового ряду для senior

Тренд для middle спеціалістів (Рис. 12) зростає до 2022 року, але потім починає знижуватися. Сезонна складова має менші коливання (±70 одиниць). Водночас залишкова складова демонструє високу мінливість, особливо у 2022-2023 роках.

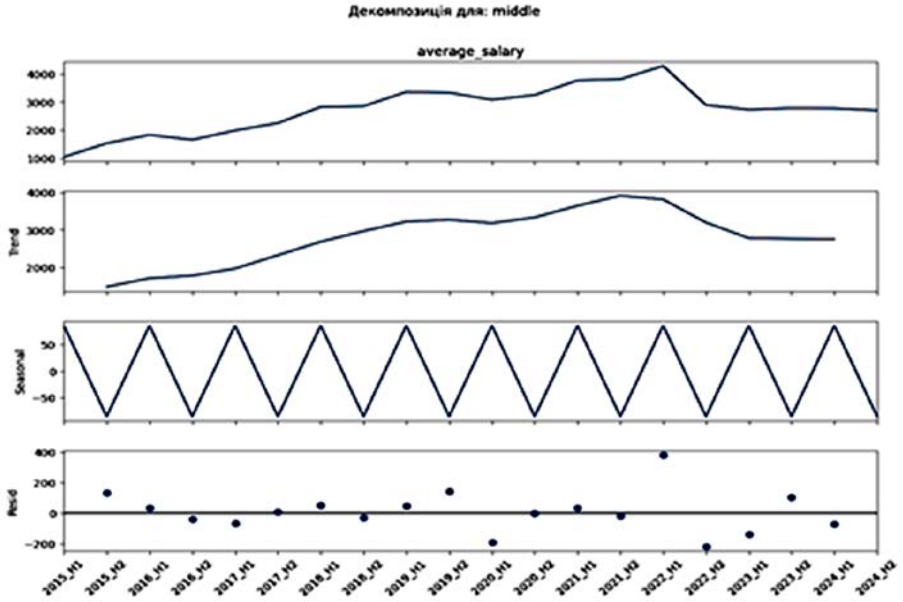

Рис. 12. Декомпозиція часового ряду для middle

Тренд для junior фахівців (Рис.13) помірно зростає до 2021 року з подальшою стабілізацією. Сезонна складова має найменші коливання (±40 одиниць). Залишкова складова є найстабільнішою серед усіх рівнів.

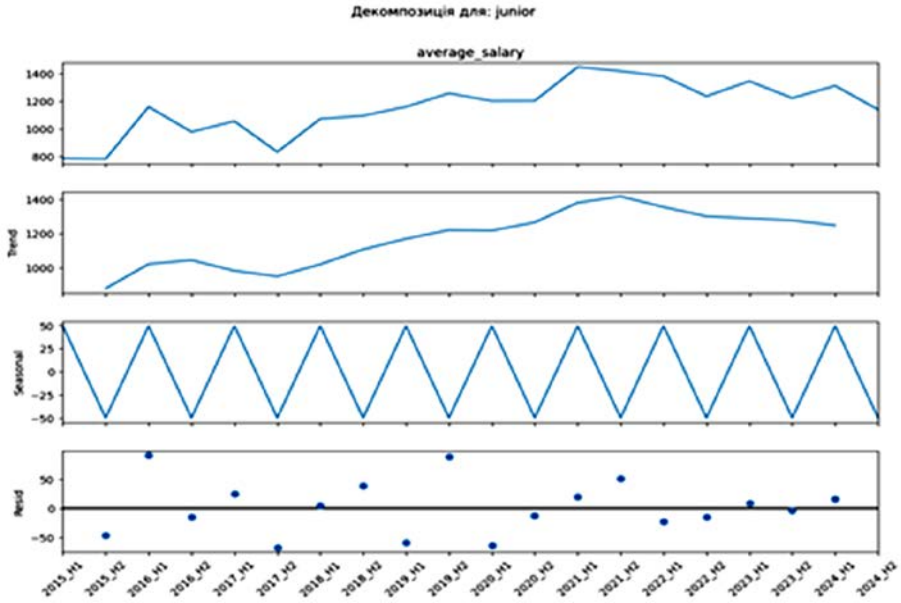

Рис. 13. Декомпозиція часового ряду для junior

Декомпозиція часових рядів середньої заробітної плати за рівнями кваліфікації дає змогу проаналізувати динаміку кожної складової (тренд, сезонність, залишки) та виявити структурні відмінності між професійними категоріями.

Періодограма (рис. 14) показує, що середня зарплата змінюється не хаотично, а з певною регулярністю. Найбільші коливання значень показника у часі характеризуються повільною динамікою, що може бути зумовлено сезонними або циклічними економічними процесами. Менші коливання значень показника у часі, які відбуваються частіше, майже не впливають на загальну картину. Отже, головні зміни у зарплаті відбуваються поступово, а не різко. Така інформація корисна при прогнозуванні значення періодів, які впливають на зміни в зарплатах.

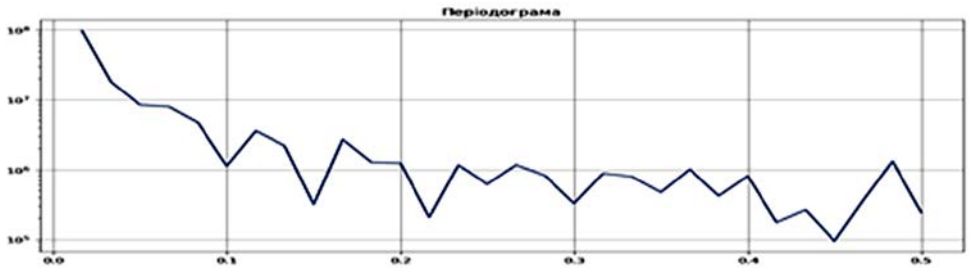

Рис. 14. Періодограма

При заповненні пропущених значень враховано відсутність значень в колонці обмінний курс гривні до долара США за допомогою лінійної інтерполяції. Система знаходить відсутні значення, обчислюючи їх як середнє між сусідніми точками. Такий спосіб підходить, коли курс змінюється поступово й без різких стрибків. Далі перевіряється, чи залишилися пропуски, які могли бути на початку або в кінці ряду, де інтерполяція неможлива через недостатність даних з обох боків. Після цього розраховуються відсутні значення ВВП у доларах США, його зміна та відсоткова зміна

шляхом перетворення даних з гривень у долари з урахуванням курсу. Відсоткова зміна визначається як відношення зміни до значення попереднього періоду, що є стандартним підходом в економіці.

На завершення створено новий датафрейм df_usd, у якому залишаються лише економічні показники, виражені в доларах США або пов’язані з ними, зокрема середня зарплата, ВВП, індекс споживчих цін, курс обміну та інші ключові змінні, що використовуються для подальшого аналізу у валюті, зручній для міжнародного порівняння. Перевірка кількості пропущених значень у фінальних даних дозволяє впевнитися, що підготовка була виконана якісно.

Далі здійснено опрацювання індексів датафрейму df_usd, де кожен запис має формат типу "YYYY_H1" або "YYYY_H2", що відповідає першому або другому півріччю певного року. Із цього текстового індексу виділяється рік і півріччя. Рік перетворюється на ціле число і записується в колонку датафрейму, а півріччя позначається як 'H1' або 'H2'. Створюється дата, яка репрезентує кінець відповідного півріччя, для 'H1' − це 30 червня, для 'H2' − 31 грудня. Після цього текстові значення 'H1' і 'H2' замінюються на числові маркери 1 і 2, які також записуються у датафрейм. Дані сортуються за створеною датою, після чого ця дата встановлюється як новий індекс датафрейму, що дозволяє ефективно працювати з часовим рядом.

Функція статистичної декомпозиції часових рядів на три компоненти − тренд, сезонність і залишок (stl_decompose_by_seniority), із використанням локальної регресії, була застосована для кожної групи працівників окремо − наприклад, junior, middle, senior. Для цього дані розподілено на підгрупи за значенням змінної 'final_seniority'. Для кожної підгрупи виконується розкладання часового ряду на три складові: тренд (довгострокова зміна), сезонність (регулярні повторювані коливання) та залишки (шум або випадкові відхилення), з використанням методу сезонно-трендової декомпозиції з використанням локальної регресії.

У результаті обчислена залишкова компонента, яка зберігається в новій колонці 'cleaned'. Вона містить значення заробітної плати після вилучення сезонних і трендових впливів, тобто показує лише ті зміни, які не пояснюються звичайною структурою часового ряду. Це дозволило виявляти нетипові коливання або аномалії. Усі опрацьовані підгрупи об'єднано в один датафрейм, який і повертається як результат.

Метод plot_acf_pacf_by_seniority візуалізував автокореляційну (ACF) та часткову автокореляційну (PACF) функції для часових рядів середньої заробітної плати, очищених від тренду та сезонності, окремо для кожного рівня сеньйорності. Це дало змогу оцінити часову залежність і глибину автокореляційних ефектів у межах кожної групи.

Вхідним аргументом є датафрейм, який уже пройшов STL-декомпозицію і містить колонку 'cleaned', що відображає залишкову компоненту (тобто ряд без сезонності та тренду). Для кожного рівня 'final_seniority' відфільтровувалась відповідна підгрупа даних, після чого до очищеного ряду застосовувався аналіз автокореляції. Графік ACF дозволив оцінити наявність залежностей між поточними та попередніми значеннями часового ряду, тоді як PACF дозволив виявити прямі зв’язки між значеннями з урахуванням проміжних лагів. Кількість лагів, які аналізувалися, визначалися параметром lags. Візуалізація для кожної категорії сеньйорності (див. рис. 15, 16, 17) виконана окремо на двох панелях: одна для ACF, інша для PACF. Це дає змогу дослідити структуру залишкових залежностей і визначити, чи є автокореляція суттєвою після очищення ряду, що важливо для побудови моделей часових рядів.

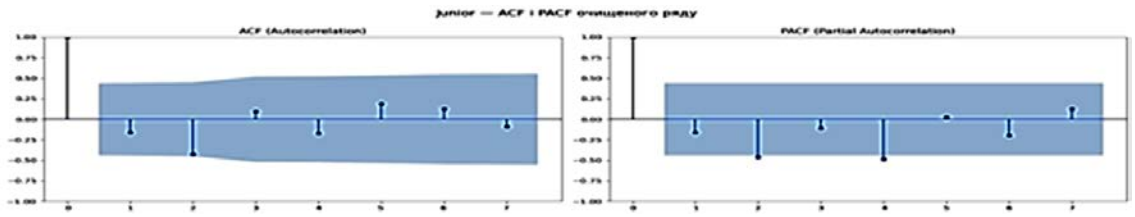

Рис. 15. Графік ACF і PACF для junoir-спеціалістів

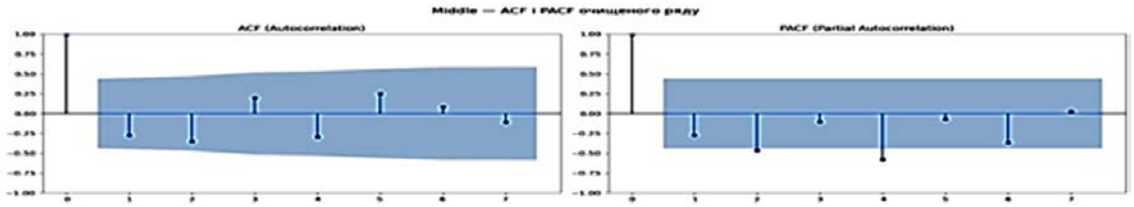

Рис.16. Графік ACF і PACF для middle-спеціалістів

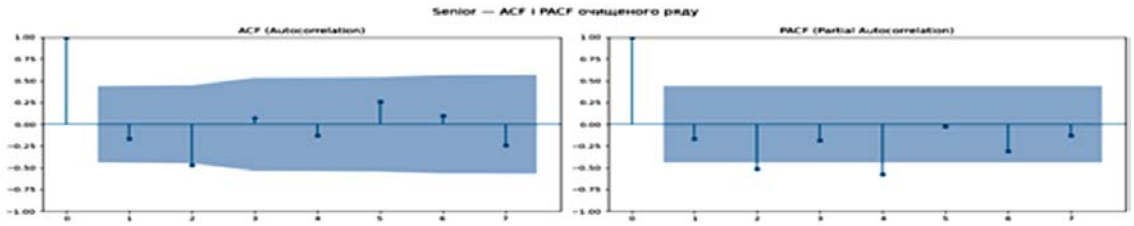

Рис. 17. Графік ACF і PACF для senior-спеціалістів

Для того, щоб модель могла враховувати зміни середньої зарплати у попередніх періодах, створено нові колонки з лагами. Наприклад, при заданні лагу 2, додалась нова колонка з даними про зарплату два півріччя назад. Такі колонки дали змогу моделі зрозуміти, як зарплата змінювалася з часом. Після створення лагів видаляються рядки з порожніми значеннями, оскільки на початку таблиці ще немає попередніх значень для деяких лагів. Отримана таблиця використовувалася для подальшого навчання моделі.

Окремо для кожного рівня сеньйорності відбулося навчання моделі XGBoost. Спочатку з основної таблиці відбиралися тільки ті записи, які відповідають конкретному рівню. Потім до них додавалися лагові ознаки за допомогою описаної вище функції. Далі формувався набір ознак, які подавалися на вхід моделі. Серед них макроекономічні показники (ВВП, курс долара, індекс споживчих цін тощо), похідні змінні (зміни в абсолютному й відсотковому вираженні), календарні ознаки (рік, півріччя), очищена зарплата та створені лаги. Дані розподілялися на тренувальну і тестову частини. Останні чотири періоди залишалися для перевірки, а решта – для навчання моделі. Після цього модель XGBRegressor навчалася на тренувальних даних і робила прогноз на тестовій частині.

Для кожного рівня сеньйорності − Junior, Middle та Senior − була створена окрема модель машинного навчання. Ми використовували алгоритм XGBRegressor, який добре підходить для задач регресії. Для навчання моделі бралися дані про економічні показники (наприклад, ВВП, курс долара, індекс споживчих цін), пошукові запити в Google, а також попередні значення зарплат. Додатково ми створили ознаки на основі затримки на 2 періоди (lag = 2), щоб модель могла враховувати динаміку. Щоб перевірити якість прогнозів, ми розділили дані. Більшу частину використали для навчання, а останні 4 періоди − для тестування. Результати роботи моделі оцінили за допомогою двох метрик: MAE, RSME. Підсумкові результати моделей подано на рис. 18, 19, 20. Для рівня Junior середнє абсолютне відхилення (MAE) становило 134 долари, а корінь з середньоквадратичної помилки (RMSE) − 165 доларів. Для Middle ці значення склали 610 і 698 доларів відповідно. У випадку з Senior − MAE дорівнює 360 доларів, а RMSE – 444 долари. Моделі навчались на даних до 2022 року, а для перевірки використовувалися значення після 2022-го. У цей період спостерігалося різке зниження зарплат, пов’язане з повномасштабною війною в Україні, що вплинуло на точність прогнозів. Щоб отримати кращі результати, зробили декілька покращень. По-перше, додали нові ознаки, які враховують події, що впливають на IT-сферу, або особливості конкретної країни. По-друге, провели налаштування моделі − підібрали оптимальні параметри, щоб вона краще підлаштовувалася під дані. Результати можуть покращитися, якщо розширити обсяг даних для навчання.

Рис. 18. Фактична та прогнозована заробітна плата для junior спеціалістів

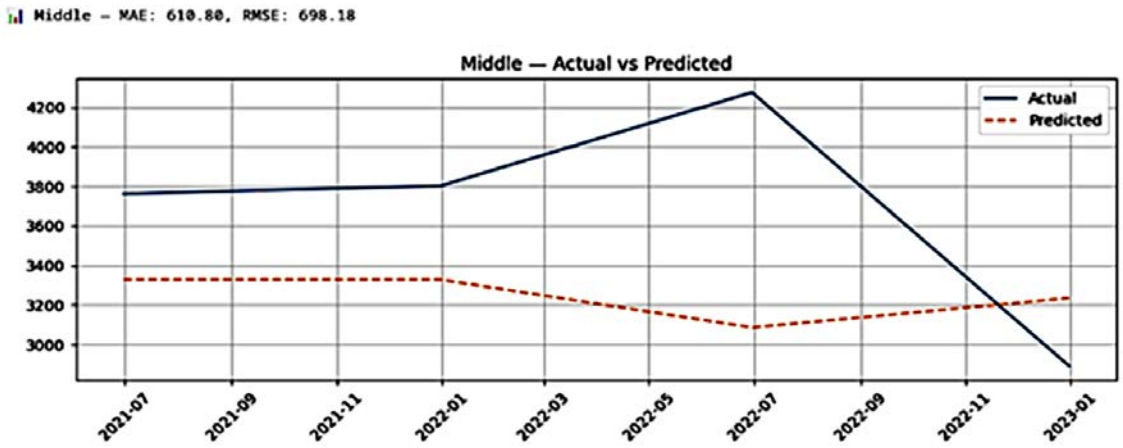

Рис. 19. Фактична та прогнозована заробітна плата для middle спеціалістів

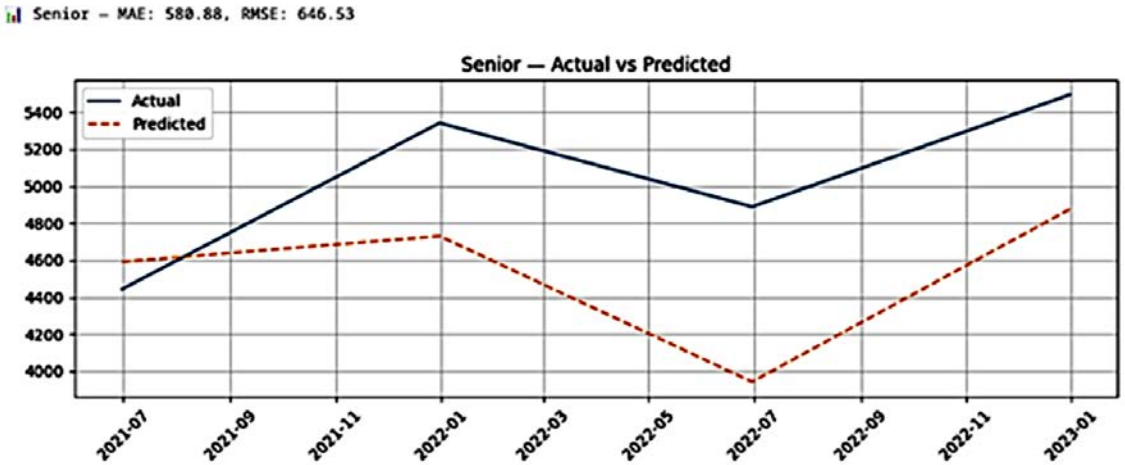

Рис. 20. Фактична та прогнозована заробітна плата для senoir спеціалістів

Для зручності користувача було створено графічний веб-інтерфейс (рис. 21) за допомогою бібліотеки Gradio. Це надало додаткові зручності при створенні прогнозу зарплати спеціаліста з Data Science, враховуючи різні економічні показники та рівень кваліфікації (junior, middle або senior).

Після запуску додатку користувач бачить просту форму, де потрібно ввести значення таких параметрів, як валовий внутрішній продукт (GDP), курс гривні до долара, кількість пошуків "data science" у Google, споживчі витрати, індекс споживчих цін, середня зарплата, приріст зарплати та інші. Також потрібно обрати період (рік і півріччя), на який необхідно отримати прогноз.

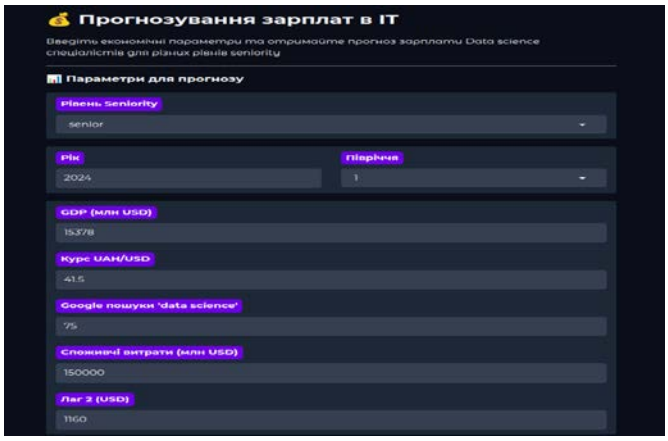

Рис. 21. Графічний веб-інтерфейс

Для кожного рівня seniority створена окрема модель, яка тренується на історичних даних. При натисканні кнопки "Передбачити зарплату" система проводить розрахунок та показує результат (див. рис. 22) – прогнозовану зарплату у доларах США.

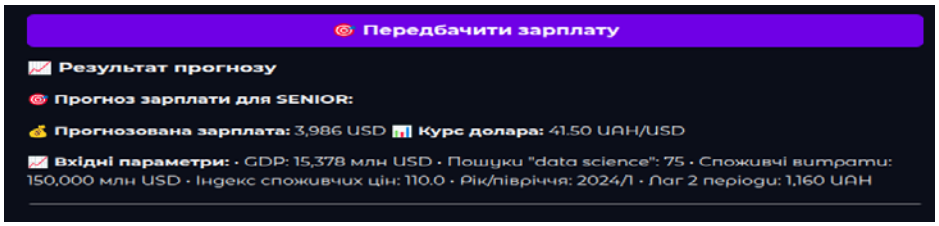

Рис. 22. Результати передбачення

Запропонована гібридна модель прогнозування IT-ринку підтверджує задекларовані новації: поєднання класичних і ML-методів у єдиному конвеєрі; урахування специфічних українських макроекономічних чинників; адаптивне ансамблювання, що підвищує стійкість прогнозів до структурних змін.

Висновки

Створена модель машинного навчання прогнозує, скільки можуть заробляти фахівці, залежно від економічної ситуації в країні та інших факторів. Така модель може бути корисною як для компаній, так і для працівників − щоб краще планувати витрати, зарплати та приймати обґрунтовані рішення. На першому етапі було зібрано дані про заробітні плати спеціалістів Data Science з різних відкритих джерел. Після цього дані були опрацьовані. Виправлено помилки, видалено зайве, додано економічні показники − наприклад, ВВП, курс долара, інфляцію тощо. Це дозволило створити повний і чистий набір даних для подальшого аналізу. Далі було проведено попередній аналіз даних, що включав побудову графіків, перегляд розподілу зарплат, аналіз змін у часі та інші дії, щоб краще зрозуміти, як поводяться дані. Також було перевірено, чи можна використовувати дані для побудови прогнозів − зокрема, проаналізовано тренди та сезонність у заробітних платах.

Для побудови прогнозу було обрано популярний алгоритм XGBoost. Цей алгоритм добре підходить для задач регресії, тобто коли треба передбачити числове значення − у нашому випадку, зарплату. Модель навчалася на історичних даних, а потім була перевірена на нових прикладах. Було використано дві основні метрики для оцінки точності MAE та RMSE. Результати показали, що модель коректно працює. Середня помилка для рівня Junior склала 134 долари, для Middle − 610 доларів, для Senior – 360 доларів.

Щоб зробити модель зручною у використанні, було створено простий веб-інтерфейс за допомогою бібліотеки Gradio. Користувачу надається можливість введення основних економічних параметрів (наприклад, рівень ВВП чи курс долара) − і одразу отримати прогноз зарплати для різних рівнів досвіду.

Доведено переваги машинного навчання у точності, стабільності та здатності працювати з динамічними й нелінійними структурами даних та практичну значущістю прогнозів для державних стратегій цифрового розвитку, бізнес-аналітики та формування освітніх політик у сфері ІТ.

У майбутньому модель буде покращуватись, додаватимуться нові характеристики, що враховують події у сфері ІТ або специфіку окремих країн. Також планується зібрати більше даних, щоб краще налаштувати модель для ще точніших результатів.

СПИСОК ЛІТЕРАТУРИ

Alwanin, R., Bchir, O., & Ben Ismail, M. M. (2025). Gradient boosting-based simultaneous classification and regression approach. IEEE Access, PP(1), 1–1. https://doi.org/10.1109/ACCESS.2025.3576086

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2016). Time series analysis: Forecasting and control (5th ed.). John Wiley & Sons. https://www.wiley.com/en-us/Time+Series+Analysis:+Forecasting+and+Control,+5th+Edition-p-9781118675021

Bustos, O., Pomares-Quimbaya, A., & Stellian, R. (2025). Machine learning, stock market forecasting, and market efficiency: A comparative study. International Journal of Data Science and Analytics, 20, 6815–6839.

CB Insights. (2025). State of the global tech market report 2025. CB Insights Research.https://www.cbinsights.com/research

Chollet, F., & Allaire, J. J. (2022). Deep learning with R (2nd ed.). Manning Publications. https://www.manning.com/books/deep-learning-with-r-second-edition

Elkan, C. (2010). Predictive analytics and data mining for business intelligence. Information Processing & Management, 59(3), 102901. https://www.researchgate.net/publication/228780185_Predictive_analytics_and_data_mining

Fatima, H., Tahir, B., & Hamid, K. (2025). Hybrid ARIMA and LSTM deep learning models empowering and enhancing forecast accuracy in sales. Spectrum of Engineering Sciences, 3(9), 117–133. https://www.sesjournal.org/index.php/1/article/view/96

Gartner. (2024). Top 10 Strategic Technology Trends for 2026.. Gartner Inc.https://www.gartner.com/en/research

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.https://www.deeplearningbook.org

Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. International Journal of Forecasting, 37(1), 388–427.

McKinsey & Company. (2025). AI-driven forecasting: How machine learning reshapes technology market planning. McKinsey Global Institute.https://www.mckinsey.com/mgi

Panchal, S., Ferdouse, L., & Sultana, A. (2024). Comparative analysis of ARIMA and LSTM models for stock price prediction. In Proceedings of the 2024 IEEE International Conference on Software, Networking and Parallel/Distributed Processing (SNPD) (pp. 240–244). IEEE. https://doi.org/10.1109/SNPD61259.2024.10673919.

Shah, R., Bhutia, T., Manga, D., Limboo, H., & Rai, M. (2025). Trends in ensemble learning and model optimization. Expert Systems with Applications, 228, 120586. https://doi.org/10.24874/QF.25.135

World Bank. (2024). Digital economy indicators database.https://databank.worldbank.org/source/digital-economy

FORECASTING THE DEVELOPMENT TRENDS OF THE IT MARKET USING MACHINE LEARNING METHODS

Ihor Pasemko, Oleksandr Lozytskyy

Lviv Polytechnic National University,
Information Systems and Networks Department, Lviv, Ukraine
E-mail: Ihor.s.Pasemko @lpnu.ua, ORCID: 0009-0000-2055-7170
E-mail: oleksandr.a.lozytskyy@lpnu.ua, ORCID: 0000-0001-8395-8385

The article explores approaches to forecasting the development trends of the IT market using machine learning methods. The relevance of the research is driven by the high dynamics of the digital economy, rapid technological changes, and the need for scientifically grounded analytical tools in the IT domain. The purpose of the study is to develop a forecasting model capable of identifying patterns in socio-economic, technological, and behavioral indicators that determine the state and prospects of IT market development.

The study employs time series analysis, correlation-regression modeling, and machine learning techniques − specifically, the Random Forest, Gradient Boosting, and Long Short-Term Memory algorithms. The dataset integrates macroeconomic indicators (GDP, exchange rate, consumer price index), IT market metrics (average salary, job demand, IT service export volume), and behavioral data (Google Trends search indices). Data preprocessing, feature normalization, and stationarity testing of time series were performed.

The results show that the Random Forest and LSTM models achieve the highest forecasting accuracy ($R^2 > 0.85$), effectively capturing both short- and medium-term trends in IT market development. Key predictors of market dynamics were identified, including the level of consumer spending, average salary, and IT query popularity index.

The practical significance of the research lies in the possibility of applying the proposed models to support strategic decision-making in the digital economy, educational and workforce planning, and forecasting technological trends in the national IT sector.

Keywords – machine learning; forecasting; IT market; time series; economic indicators; stationarity; seasonal-trend decomposition; Random Forest; Gradient Boosting; Long Short-Term Memory.