НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
Навчально-науковий інститут прикладного системного аналізу

Кафедра штучного інтелекту

До захисту допущено:
В. о. завідувачки кафедри
_____Ірина ДЖИГИРЕЙ
«__» _____ 20__ р.

Дипломна робота

на здобуття ступеня бакалавра
за освітньо-професійною програмою «Системи і методи штучного інтелекту»
спеціальності 122 «Комп'ютерні науки»
на тему: «Інтелектуальна система прогнозування часових рядів продажів із урахуванням екзогенних факторів»

Виконала:
студентка IV курсу, групи КІ-21
Кузьмінська Віра Сергіївна _____

Керівник:
асистент кафедри ШІ,
Сидорський Володимир Сергійович _____

Консультант з нормоконтролю:
фахівець першої категорії кафедри ШІ
к. т. н., доцент,
Комариста Богдана Миколаївна _____

Рецензент:
асистент кафедри ММСА,
доктор філософії з системного аналізу,
Канцедал Георгій Олегович _____

Засвідчую, що у цій дипломній роботі
немає запозичень з праць інших авторів
без відповідних посилань.
Студентка _____

Київ — 2026 року

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
Навчально-науковий інститут прикладного системного аналізу
Кафедра штучного інтелекту

Рівень вищої освіти — перший (бакалаврський)
Спеціальність — 122 «Комп'ютерні науки»
Освітньо-професійна програма «Системи і методи штучного інтелекту»

ЗАТВЕРДЖУЮ
В. о. завідувачки кафедри
_____Ірина ДЖИГИРЕЙ
«__» _____ 2026 р.

ЗАВДАННЯ
на дипломну роботу студентці
Кузьмінській Вірі Сергіївні

1. Тема роботи: «Інтелектуальна система прогнозування часових рядів продажів із урахуванням екзогенних факторів», керівник роботи Сидорський Володимир Сергійович, затверджені наказом по НН ІПСА від «27» травня 2026 р. № 1944-с.

2. Термін подання студенткою роботи: «05» червня 2026 року.

3. Вихідні дані до роботи: дані щоденних продажів мережі продовольчих магазинів Еквадору (54 магазини, 33 товарні категорії, період з 1 січня 2013 року по 15 серпня 2017 року); наукові публікації та навчальні матеріали з тематики прогнозування часових рядів, машинного навчання, градієнтного бустингу та конструювання ознак; програмні засоби для розробки моделей машинного навчання (LightGBM, scikit-learn, pandas); результати експериментальної перевірки роботи побудованих моделей.

4. Зміст роботи: характеристика предметної області прогнозування часових рядів продажів; огляд статистичних методів, методів машинного навчання та

глибокого навчання для часових рядів; розвідувальний аналіз реального набору даних продажів та виявлення ключових закономірностей; розробка стратегії передобробки даних та конструювання ознакового простору з урахуванням класифікації екзогенних коваріатів; побудова та навчання прогнозних моделей різних класів (Ridge-регресія, градієнтний бустинг, гібридні та ансамблеві моделі) із застосуванням коректної стратегії валідації для часових рядів; порівняльний аналіз моделей за обраними метриками якості та дослідження впливу окремих груп екзогенних факторів на точність прогнозу.

5. Перелік ілюстративного матеріалу: використані метрики якості прогнозування, результати крос-валідації та оцінки на цільовому прогнозному вікні, абляційний аналіз внеску груп ознак, важливість ознак фінальної моделі, аналіз помилок за товарними категоріями, електронна презентація.

6. Дата видачі завдання: «02» лютого 2026 року.

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Примітка
1	Огляд літератури за темою	24.04.2026	Виконано
2	Підготовка першого розділу	01.05.2026	Виконано
3	Підготовка другого розділу	08.05.2026	Виконано
4	Розробка програмного продукту	22.05.2026	Виконано
5	Підготовка третього розділу	29.05.2026	Виконано
6	Оформлення розділів відповідно до нормоконтролю	03.06.2026	Виконано
7	Оформлення дипломної роботи	05.06.2026	Виконано
8	Підготовка презентації доповіді	10.06.2026	Виконано

Студентка Віра КУЗЬМІНСЬКА

Керівник Володимир СИДОРСЬКИЙ

РЕФЕРАТ

Дипломна робота: 92 с., 5 рис., 26 табл., 23 посилання, додаток.

ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ, МАШИННЕ НАВЧАННЯ, ГРАДІЄНТНИЙ БУСТИНГ, LIGHTGBM, ЕКЗОГЕННІ ФАКТОРИ, ОЗНАКОВИЙ ПРОСТІР, RMSLE, АНСАМБЛЕВЕ МОДЕЛЮВАННЯ.

Об'єктом дослідження є процес прогнозування обсягів продажів у мережі роздрібної торгівлі з урахуванням впливу зовнішніх факторів.

Предметом дослідження є методи машинного навчання для прогнозування багатовимірних часових рядів продажів із використанням екзогенних коваріатів.

Метою роботи є розробка інтелектуальної системи прогнозування часових рядів продажів, яка ефективно інтегрує екзогенні фактори різної природи для підвищення точності прогнозів порівняно з базовими підходами.

У роботі проведено огляд методів прогнозування часових рядів від класичних статистичних підходів до методів машинного та глибокого навчання, запропоновано класифікацію екзогенних факторів. Виконано розвідувальний аналіз реального набору даних мережі продовольчих магазинів Еквадору (1 782 часових ряди, період 2013–2017), розроблено стратегії передобробки та сконструйовано ознаковий простір з 27 ознак у семи групах. Запропоновано валідаційну схему розширювального вікна з трьома непересічними фолдами.

Реалізовано інтелектуальну систему прогнозування у вигляді інтегрованого конвеєра з шести компонент: попередньої обробки даних, конструювання ознак, data-driven відбору ознак свят, ансамблю чотирьох моделей LightGBM, автоматичного підбору ваг та валідаційної підсистеми. Зважений ансамбль досягнув RMSLE $= 0{,}425$ на цільовому прогнозному вікні — скорочення помилки на 59% відносно базової моделі. Виявлено два систематичні режими помилок, що характеризують фундаментальне обмеження класу історико-залежних моделей.

ABSTRACT

Bachelor's thesis: 92 pages, 5 figures, 26 tables, 23 references, appendix.

TIME SERIES FORECASTING, MACHINE LEARNING, GRADIENT BOOSTING, LIGHTGBM, EXOGENOUS FACTORS, FEATURE SPACE, RMSLE, ENSEMBLE MODELING.

The object of the study is the process of forecasting sales volumes in a retail network with consideration of the impact of external factors.

The subject of the study is machine learning methods for forecasting multivariate sales time series using exogenous covariates.

The aim of the work is to develop an intelligent system for forecasting sales time series that effectively integrates exogenous factors of different nature to improve forecast accuracy compared to baseline approaches.

The work reviews time series forecasting methods from classical statistical approaches to machine learning and deep learning, and proposes a classification of exogenous factors. An exploratory analysis of a real dataset from an Ecuadorian grocery store chain (1 782 time series, period 2013–2017) was performed, preprocessing strategies were developed, and a feature space of 27 features in seven groups was constructed. An expanding window validation scheme with three non-overlapping folds was proposed.

An intelligent forecasting system was implemented as an integrated pipeline of six components: data preprocessing, feature engineering, data-driven holiday feature selection, an ensemble of four LightGBM models, automatic weight optimization, and a validation subsystem. The weighted ensemble achieved RMSLE $= 0.425$ on the target forecast window — a 59% error reduction compared to the baseline model. Two systematic error regimes characterizing a fundamental limitation of history-dependent model classes were identified.

ЗМІСТ

ВСТУП

Прогнозування[1] попиту є однією з фундаментальних задач управління у сфері роздрібної торгівлі. Здатність точно передбачити обсяги продажів на короткостроковому горизонті безпосередньо впливає на якість рішень щодо управління запасами, планування ресурсів та фінансового прогнозування. Однак складність цієї задачі зумовлена множинною сезонністю, високою розмірністю даних та впливом численних зовнішніх факторів — від макроекономічних показників до непередбачуваних подій. Розвиток обчислювальних потужностей та доступність великих обсягів даних відкрили можливості для застосування методів машинного навчання, які демонструють суттєво вищу точність на реальних задачах прогнозування продажів.

Ця робота спрямована на дослідження та порівняння сучасних підходів до прогнозування продажів з акцентом на врахування зовнішніх факторів різної природи. Дослідження виконано на реальних даних мережі роздрібної торгівлі та охоплює повний цикл — від аналізу даних та конструювання ознак до побудови, навчання та порівняння прогнозних моделей. Отримані результати мають практичну цінність для управління запасами, планування закупівель та маркетингових кампаній у мережах роздрібної торгівлі.

[1] Тут і нижче використано такий інструмент штучного інтелекту, як чат-бот з генеративним штучним інтелектом Claude, виключно для корегування та редагування тексту, створеного автором цієї дипломної роботи, на основі автоматизованої перевірки граматики, структури та стилю, що відповідає Політиці використання штучного інтелекту для академічної діяльності в КПІ ім. Ігоря Сікорського (протокол №11 Вченої ради КПІ ім. Ігоря Сікорського від 11 грудня 2023 р.).

РОЗДІЛ 1
АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ОГЛЯД МЕТОДІВ ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ

1.1 Актуальність задачі прогнозування часових рядів продажів

Роздрібна торгівля є однією з найбільших галузей світової економіки. За даними National Retail Federation, прогнозований обсяг роздрібних продажів лише у Сполучених Штатах на 2026 рік складає 5,6 трильйона доларів із темпом зростання 4,4% відносно попереднього року [1]. Галузь забезпечує значну частку робочих місць у більшості країн світу та формує суттєву долю валового внутрішнього продукту. В умовах такого масштабу навіть незначне підвищення точності прогнозування попиту здатне принести відчутний економічний ефект.

Прогнозування продажів є ключовим елементом операційного управління в ритейлі. Від точності прогнозу безпосередньо залежать рішення щодо управління запасами, планування закупівель, логістики та маркетингових кампаній [2]. Неточні прогнози призводять до двох основних проблем: дефіциту товарів (stockout) та надлишкових запасів (overstock). Сукупні втрати від дисбалансу запасів у глобальному ритейлі вимірюються трильйонами доларів щорічно. Дефіцит товарів на полицях спричиняє не лише пряму втрату виручки, але й суттєве зниження лояльності покупців: споживач, який не знаходить потрібного товару, психологічно схильний звернутись до конкурента, і повернути його довіру значно складніше, ніж втратити. З іншого боку, надлишкові запаси заморожують оборотний капітал, збільшують витрати на зберігання та призводять до знецінення або повного списання товарів з обмеженим терміном придатності. Таким чином, точне прогнозування попиту є критичним фактором балансу між рівнем обслуговування клієнтів та ефективністю використання ресурсів.

Традиційно задача прогнозування продажів вирішувалась за допомогою експертних оцінок та класичних статистичних методів, таких як ARIMA та експоненціальне згладжування [4]. Однак сучасний ритейл характеризується зна-

чною складністю: десятки тисяч товарних позицій, сотні торгових точок, різноманітні промоакції, свята, макроекономічні фактори та непередбачувані події — природні катастрофи, пандемії, воєнні дії — створюють багатовимірний простір факторів впливу. Розвиток обчислювальних потужностей та доступність великих обсягів даних відкрили можливості для застосування методів машинного навчання до цієї задачі. За даними McKinsey, впровадження прогнозування на основі штучного інтелекту дозволяє скоротити помилки прогнозування на 20–50% та зменшити втрати від дефіциту товарів на 65% [5].

Таким чином, задача прогнозування часових рядів продажів із урахуванням екзогенних факторів є актуальною як з наукової точки зору — розвиток методів, здатних ефективно моделювати складні багатофакторні залежності, — так і з практичної — суттєвий економічний ефект від підвищення точності прогнозів у роздрібній торгівлі.

1.2 Особливості часових рядів продажів у роздрібній торгівлі

Часові ряди продажів у роздрібній торгівлі мають низку характерних особливостей, що відрізняють їх від часових рядів в інших предметних областях і створюють специфічні виклики для побудови прогнозних моделей [3].

Однією з найбільш виражених властивостей є множинна сезонність. Продажі демонструють періодичні коливання одночасно на кількох рівнях: тижневому (підвищений попит у вихідні дні), місячному (піки в дні виплати заробітних плат), річному (святковий сезон, шкільний сезон, сезонні товари). Ці сезонні патерни можуть накладатися один на одного, створюючи складну структуру, яку неможливо описати єдиною сезонною компонентою. Класичні методи декомпозиції часових рядів, розраховані на один сезонний цикл, виявляються недостатніми для адекватного опису такої поведінки [3].

Другою важливою характеристикою є висока розмірність задачі. Типовий ритейлер оперує тисячами товарних позицій у десятках торгових точок, що породжує тисячі взаємопов’язаних часових рядів [2]. Наприклад, мережа з 54 ма-

газинів і 33 товарними категоріями формує 1782 окремих часових ряди продажів, кожен із яких має власну динаміку, але водночас пов'язаний із іншими через спільні зовнішні фактори та поведінку споживачів. Це ставить питання вибору між локальними моделями (окрема модель на кожен ряд), глобальними моделями (єдина модель на всі ряди) та проміжними варіантами (модель на групу схожих рядів).

Наявність нульових та переривчастих рядів є ще однією характерною рисою ритейлу. Деякі товарні категорії можуть демонструвати тривалі періоди нульових продажів — через відсутність товару на складі, сезонне вилучення з асортименту або низький попит. Крім того, продажі дорівнюють нулю у дні, коли магазини закриті (наприклад, на Різдво чи Новий рік). Такі ряди погано піддаються моделюванню стандартними методами, які припускають неперервну природу цільової змінної.

Часові ряди продажів також характеризуються нестаціонарністю: їхні статистичні властивості змінюються з часом [4]. Довгострокові тренди (зростання або спад мережі), зміна структури асортименту, поява нових магазинів, інфляція — все це призводить до того, що закономірності, виявлені на історичних даних, можуть втрачати актуальність. Окремим випадком нестаціонарності є структурні зломи — різкі зміни у динаміці рядів, спричинені непередбачуваними подіями: природними катастрофами, пандеміями, воєнними діями або різкими змінами в економічній політиці.

Окрім перелічених вище особливостей, важливою рисою часових рядів продажів є залежність від екзогенних факторів — зовнішніх змінних, що впливають на динаміку цільової величини, але не залежать від неї [3]. До таких факторів належать макроекономічні індикатори (ціни на енергоносії, рівень інфляції, курс валют), календарні події (свята, вихідні, дні зарплат), промоакції, погодні умови та інші. Ефективне врахування цих факторів вимагає використання моделей, здатних працювати з багатовимірним вхідним простором, що виходить за рамки класичного одновимірного аналізу часових рядів.

Сукупність перелічених особливостей — множинна сезонність, висока розмірність, наявність нульових значень, нестаціонарність та залежність від екзогенних факторів — визначає вимоги до методів прогнозування та обґрунто-

вує необхідність застосування підходів машинного навчання, здатних ефективно працювати з табличними багатовимірними даними.

1.3 Роль екзогенних факторів у прогнозуванні продажів

Як зазначено у попередньому підрозділі, часові ряди продажів суттєво залежать від зовнішніх змінних, які не є частиною самого ряду, але впливають на його динаміку. У літературі з аналізу часових рядів такі змінні називають екзогенними факторами, або коваріатами [3]. Їх коректне врахування є одним із ключових чинників підвищення якості прогнозів, оскільки модель, побудована виключно на історичних значеннях цільової змінної, неспроможна пояснити значну частину варіації продажів [6].

Екзогенні фактори, що впливають на продажі у ритейлі, можна класифікувати за кількома ознаками.

За природою впливу виділяють макроекономічні, календарні, комерційні та ситуативні фактори [2]. Макроекономічні фактори включають ціни на енергоносії та сировину, рівень інфляції, курс національної валюти та загальний стан економіки. Коливання цих показників впливають на купівельну спроможність населення і, відповідно, на обсяги роздрібних продажів. Ступінь цього впливу залежить від структури економіки конкретної країни та її чутливості до зовнішніх шоків. Календарні фактори охоплюють свята, вихідні дні, початок навчального року, періоди виплати заробітних плат та інші регулярні події, що формують передбачувані піки та спади попиту. Комерційні фактори — це промоакції, знижки, рекламні кампанії та зміни в асортименті, які безпосередньо стимулюють або стримують продажі конкретних товарних категорій. Ситуативні фактори включають непередбачувані події: стихійні лиха, епідемії, воєнні конфлікти, — які здатні різко змінити структуру попиту на тривалий період.

За доступністю на момент прогнозування екзогенні фактори поділяють на минулі (past covariates), майбутні (future covariates) та статичні (static covariates) [7]. Минулі коваріати — це змінні, значення яких відомі лише за минулі

періоди. Прикладом є фактична кількість покупців у магазині: на момент прогнозування відомі лише історичні дані. Майбутні коваріати — це змінні, значення яких відомі або можуть бути оцінені заздалегідь на період прогнозного горизонту. До них належать календарні ознаки (день тижня, місяць, свята), заплановані промоакції та, за наявності відповідних прогнозів, макроекономічні індикатори. Статичні коваріати не змінюються з часом і характеризують об'єкт спостереження: місто та регіон розташування магазину, його тип, товарна категорія. Правильна класифікація коваріатів є принципово важливою для коректної побудови моделі, оскільки використання майбутньої інформації, яка не буде доступна в момент реального прогнозування, призводить до витоку даних і завищення оцінки якості моделі [7].

За характером впливу на часовий ряд фактори можуть бути адитивними (додають фіксований ефект до базового рівня продажів), мультиплікативними (масштабують продажі у певну кількість разів) або мати нелінійну взаємодію з іншими факторами [3]. Наприклад, ефект промоакції може залежати від дня тижня: знижка на вихідних дає більший приріст продажів, ніж та сама знижка в будній день, оскільки у вихідні більше потенційних покупців відвідують магазин. Подібні взаємодії складно моделювати лінійними методами, що є однією з причин переваги ансамблевих методів машинного навчання, здатних автоматично виявляти такі нелінійні залежності [8].

Окремо варто зазначити, що вплив екзогенних факторів часто проявляється не миттєво, а з певною затримкою [3]. Макроекономічні зміни поступово транслюються у споживчі витрати, а наслідки непередбачуваних подій можуть відчуватися протягом тижнів або місяців після їх настання. Це означає, що прогнозна модель повинна бути здатна враховувати не лише поточні значення екзогенних змінних, а й їхню історичну динаміку.

Таким чином, врахування екзогенних факторів є не опціональним доповненням, а необхідною умовою побудови якісної прогнозної моделі для часових рядів продажів.

1.4 Статистичні методи прогнозування часових рядів

Статистичні методи становлять фундамент прогнозування часових рядів та залишаються важливими як базові моделі (baselines) для порівняння з більш складними підходами [3]. У цьому підрозділі коротко розглянуто три класи статистичних методів: моделі сімейства ARIMA, методи експоненціального згладжування (ETS) та модель Prophet.

1.4.1 Моделі сімейства ARIMA

Модель ARIMA(p, d, q) (AutoRegressive Integrated Moving Average) поєднує авторегресійну компоненту порядку p, диференціювання порядку d для забезпечення стаціонарності та компоненту ковзного середнього порядку q [4]. У компактній формі модель записується як:

$$\phi(B)(1 - B)^d \, y_t = c + \theta(B) \, \varepsilon_t, \tag{1.1}$$

де B — оператор зворотного зсуву ($By_t = y_{t-1}$), $\phi(B)$ та $\theta(B)$ — характеристичні поліноми авторегресійної та ковзного середнього компонент відповідно, ε_t — білий шум [9].

Для врахування сезонності модель розширюється до

$$\text{SARIMA}(p, d, q)(P, D, Q)_m$$

додаванням сезонних компонент із періодом m [4]:

$$\phi(B) \, \Phi(B^m)(1 - B)^d (1 - B^m)^D y_t = c + \theta(B) \, \Theta(B^m) \, \varepsilon_t. \tag{1.2}$$

Варіант ARIMAX додає вектор екзогенних регресорів x_t із коефіцієнтами β, проте їхній вплив залишається лінійним [3].

1.4.2 Методи експоненціального згладжування (ETS)

Сімейство моделей ETS (Error, Trend, Seasonality) будує прогноз як зважену комбінацію минулих спостережень із вагами, що експоненціально спадають [3]. Модель описується трьома компонентами стану: рівень ℓ_t, тренд b_t та сезонність s_t, кожна з яких може бути адитивною, мультиплікативною або відсутньою. Наприклад, рівняння оновлення для адитивної моделі ETS(A,A,A):

$$y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t, \tag{1.3}$$

$$\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t, \tag{1.4}$$

$$b_t = b_{t-1} + \beta\varepsilon_t, \tag{1.5}$$

$$s_t = s_{t-m} + \gamma\varepsilon_t, \tag{1.6}$$

де α, β, γ — параметри згладжування. Модель Хольта–Вінтерса із загасаючим трендом є найбільш поширеним варіантом на практиці [9].

1.4.3 Prophet

Prophet — модель, розроблена Meta, що декомпозує часовий ряд на тренд, сезонність та ефект подій [10]:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t. \tag{1.7}$$

Окремим важливим аспектом трансформації є конструювання календарних ознак, що описують сезонність і структуру часу. До типових ознак належать день тижня, день місяця, місяць, номер тижня року та індекс часу від початку спостережень [14]. Спосіб кодування цих ознак (категоріальне, числове, тригонометричне) обирається з огляду на тип моделі: для лінійних моделей зручне one-hot або тригонометричне представлення, для дерев рішень — безпосереднє

числове або категоріальне кодування з вбудованою підтримкою категоріальних змінних.

1.4.4 Обмеження статистичних методів

Статистичні методи мають спільні переваги: теоретичну обґрунтованість, інтерпретованість та обчислювальну ефективність. Водночас для задачі прогнозування продажів вони стикаються з суттєвими обмеженнями [3, 6]: лінійність припущень не дозволяє моделювати складні взаємодії між факторами; підхід «одна модель — один ряд» є неефективним при тисячах рядів; інтеграція екзогенних змінних обмежена — ARIMAX підтримує лише лінійні ефекти, ETS не працює з коваріатами, а Prophet обмежений календарними подіями. Ці обмеження мотивують застосування методів машинного навчання, які розглядаються у поданому нижче підрозділі.

1.5 Методи машинного навчання для прогнозування часових рядів

Методи машинного навчання (ML) набули широкого застосування у задачах прогнозування часових рядів завдяки здатності автоматично виявляти складні нелінійні залежності у даних без явного задання функціональної форми моделі [6]. На відміну від статистичних методів, ML-підходи природним чином працюють із великою кількістю ознак різної природи, що робить їх особливо привабливими для задач із множиною екзогенних факторів та переносить акцент із вибору моделі на конструювання ознакового простору.

1.5.1 Трансформація задачі прогнозування у задачу навчання з учителем

Застосування методів машинного навчання до часових рядів вимагає попередньої трансформації задачі у формат табличної регресії [14]. Кожне спостереження y_t перетворюється на навчальний приклад, де ознаками слугують лагові значення цільової змінної ($y_{t-1}, y_{t-2}, \ldots, y_{t-k}$), ковзні статистики (середнє, стандартне відхилення за вікно заданого розміру), календарні ознаки (день тижня, місяць, тиждень року) та значення екзогенних змінних:

$$\hat{y}_t = f(y_{t-1}, \ldots, y_{t-k},\ \bar{y}_{t,w},\ \sigma_{t,w},\ x_t,\ c)\,, \tag{1.8}$$

де $\bar{y}_{t,w}$ та $\sigma_{t,w}$ — ковзне середнє та стандартне відхилення за вікно розміру w, x_t — вектор екзогенних змінних, c — вектор статичних коваріатів, а f — довільна нелінійна функція, яку апроксимує модель.

Такий підхід дозволяє будувати глобальні моделі — єдину модель, що навчається на даних усіх часових рядів одночасно, використовуючи статичні коваріати (ідентифікатор магазину, товарна категорія, регіон) як додаткові ознаки [14]. Глобальні моделі здатні виявляти спільні закономірності між рядами та ефективніше використовувати обмежені дані, що є суттєвою перевагою над локальними статистичними моделями.

Окремим важливим аспектом трансформації є конструювання ознак для опису сезонності. Окрім бінарних індикаторів (one-hot encoding дня тижня, місяця), ефективним підходом є використання рядів Фур'є [3]: для кожного сезонного періоду P генеруються пари ознак

$$\cos\left(\frac{2\pi n t}{P}\right), \quad \sin\left(\frac{2\pi n t}{P}\right), \quad n = 1, \ldots, N, \tag{1.9}$$

де N — кількість гармонік, що контролює гладкість сезонного патерну. Такий підхід дозволяє описувати множинну сезонність (тижневу, річну) компактним набором неперервних ознак.

1.5.2 Лінійні моделі та регуляризація

Найпростішим класом моделей у машинному навчанні є лінійні моделі. Лінійна регресія прогнозує значення цільової змінної як зважену суму ознак:

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j \, x_{ij}, \tag{1.10}$$

де β_0 — вільний коефіцієнт, β_j — вагові коефіцієнти при ознаках, p — кількість ознак. Параметри моделі визначаються мінімізацією суми квадратів відхилень прогнозу від спостереження (метод найменших квадратів, OLS).

Незважаючи на простоту, базова лінійна регресія має суттєві обмеження при роботі з реальними даними [19]. При великій кількості ознак, особливо за наявності корельованих змінних, оцінки коефіцієнтів стають нестабільними та чутливими до невеликих змін у вибірці. Крім того, модель схильна до перенавчання, коли кількість ознак стає порівнянною з кількістю спостережень. Для подолання цих проблем застосовують регуляризацію — додавання штрафу за величину коефіцієнтів до функції втрат.

Найпоширенішою формою регуляризації лінійних моделей є L_2-регуляризація, відома як Ridge-регресія [19]. Функція втрат набуває вигляду:

$$\mathcal{L}_{\text{Ridge}}(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \, x_{ij} \right)^2 + \alpha \sum_{j=1}^{p} \beta_j^2, \tag{1.11}$$

де $\alpha \geq 0$ — параметр регуляризації, що контролює силу штрафу. При $\alpha = 0$ Ridge зводиться до OLS; зі зростанням α коефіцієнти стискаються до нуля, проте, на відміну від L_1-регуляризації (Lasso), жоден з них не дорівнює нулю строго. Це робить Ridge придатною для задач, у яких усі ознаки потенційно інформативні, а основною проблемою є мультиколінеарність, а не відбір ознак.

Параметри Ridge-регресії мають закриту аналітичну форму:

$$\hat{\beta} = \left(X^\top X + \alpha I \right)^{-1} X^\top y, \tag{1.12}$$

де X — матриця ознак, y — вектор цільових значень, I — одинична матриця. Додавання αI забезпечує оборотність матриці навіть при колінеарності ознак, що є основним джерелом числової стабільності методу.

На практиці перед навчанням Ridge-регресії ознаки стандартизують, оскільки штраф L_2 є чутливим до масштабу. Параметр α обирають за крос-валідацією. У контексті прогнозування часових рядів Ridge є природним кандидатом на роль базової моделі: вона швидко навчається, має невелику кількість гіперпараметрів, інтерпретується через ваги ознак та слугує нижньою межею якості, з якою порівнюються складніші моделі. Лінійна структура також робить Ridge корисною як детермінована компонента в гібридних схемах, що розглядаються у підрозділі 1.7.

1.5.3 Дерево рішень

Дерево рішень (Decision Tree) — базовий метод машинного навчання, що рекурсивно розбиває простір ознак на прямокутні області, мінімізуючи функцію втрат у кожній із них [6]. Для задачі регресії прогноз у кожному листку дорівнює середньому значенню цільової змінної для спостережень, що потрапили до цього листка. Розбиття обирається жадібно: на кожному кроці шукається ознака та порогове значення, що максимально зменшують дисперсію цільової змінної. Формально, для вузла з множиною спостережень S обирається розбиття (j, s) за ознакою j та порогом s:

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in S_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in S_2(j,s)} (y_i - c_2)^2 \right], \tag{1.13}$$

де $S_1(j, s) = \{x \mid x_j \leq s\}$, $S_2(j, s) = \{x \mid x_j > s\}$, а c_1, c_2 — оптимальні константи (середні значення) у кожній підобласті. Окреме дерево рішень є слабким класифікатором із високою дисперсією, проте слугує будівельним блоком для ансамблевих методів, зокрема градієнтного бустинга.

1.5.4 Градієнтний бустинг

Градієнтний бустинг (Gradient Boosting Machine, GBM) — ансамблевий метод, що послідовно будує композицію слабких моделей (зазвичай дерев рішень обмеженої глибини), кожна з яких виправляє помилки попередніх [12].

Нехай $F_0(x)$ — початкове наближення. На кожному кроці $m = 1, \ldots, M$ обчислюються псевдозалишки — від'ємні градієнти функції втрат:

$$r_{im} = - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F=F_{m-1}}, \tag{1.14}$$

де L — функція втрат. На ці псевдозалишки навчається нове дерево $h_m(x)$, і модель оновлюється:

$$F_m(x) = F_{m-1}(x) + \eta \, h_m(x), \tag{1.15}$$

де $\eta \in (0, 1]$ — темп навчання (learning rate) [12]. Зменшення η потребує більшої кількості дерев, але зазвичай покращує якість прогнозу та зменшує ризик перенавчання.

1.5.5 LightGBM

LightGBM (Light Gradient Boosting Machine) — реалізація градієнтного бустинга, розроблена Microsoft Research у 2017 році з акцентом на ефективність при великих обсягах даних. Алгоритм вводить дві ключові інновації.

Gradient-based One-Side Sampling (GOSS): замість використання всіх спостережень для оцінки розбиття, алгоритм зберігає усі приклади з великим градієнтом та випадково відбирає підмножину прикладів із малим градієнтом, що суттєво зменшує обсяг обчислень.

Exclusive Feature Bundling (EFB): ознаки, що рідко приймають ненульо-

ві значення одночасно, об'єднуються у пучки (bundles), зменшуючи ефективну розмірність без втрати інформації.

LightGBM використовує стратегію зростання дерев leaf-wise замість level-wise: на кожному кроці розщеплюється листок із найбільшим зменшенням функції втрат, незалежно від глибини. Це дозволяє будувати точніші дерева за меншу кількість ітерацій, хоча потребує контролю максимальної глибини для запобігання перенавчанню [13].

Важливою практичною перевагою LightGBM є вбудована підтримка категоріальних ознак без необхідності попереднього one-hot кодування, що є суттєвим при роботі зі змінними на кшталт типу магазину, міста чи товарної категорії. Класифікація коваріатів на минулі, майбутні та статичні (підрозділ 1.3) безпосередньо відображається у структурі ознакового простору, що подається на вхід моделі. Поєднання гнучкого формату вхідних даних, високої обчислювальної ефективності та конкурентної якості зумовило широке використання LightGBM як у промислових системах, так і у вітчизняних наукових дослідженнях; зокрема, у роботі [20], виконаній у Навчально-науковому інституті прикладного системного аналізу КПІ ім. Ігоря Сікорського, LightGBM застосовано до задачі побудови прогнозних моделей з пропущеними даними та продемонстровано його перевагу над іншими класичними класифікаторами.

1.5.6 Особливості застосування до часових рядів

Методи на основі дерев рішень мають низку переваг для прогнозування продажів [6]: здатність моделювати нелінійні залежності та взаємодії між ознаками; природна робота з ознаками різних типів та масштабів; вбудована оцінка важливості ознак; можливість побудови глобальних моделей на тисячах рядів одночасно. Усі ці властивості переносять центр ваги моделювання з вибору алгоритму на ретельне конструювання ознак: якість прогнозу істотно залежить від того, наскільки повно вхідні ознаки відображають структуру задачі [14].

Водночас дерева рішень не здатні екстраполювати за межі діапазону на-

вчальних даних, що ускладнює прогнозування трендів [14]. Для подолання цього часто використовують попереднє видалення тренду або конструювання ознак, що його описують (часові індекси, ковзні середні). При багатокроковому прогнозуванні виникає проблема накопичення помилок у рекурсивних стратегіях або необхідність навчання окремих моделей для кожного кроку горизонту (direct strategy) [7]. Одним із підходів до вирішення проблеми екстраполяції є гібридні моделі, що поєднують лінійні методи для детермінованих компонент із бустингом для нелінійних залежностей; цей підхід детальніше розглядається у підрозділі 1.7.

Результати великомасштабних змагань із прогнозування, зокрема M5 Competition, підтвердили конкурентоспроможність градієнтного бустинга: моделі на основі LightGBM посіли провідні позиції у змаганні з прогнозування продажів мережі роздрібної торгівлі з понад 30 тисячами часових рядів та множиною екзогенних факторів [15].

1.6 Методи глибокого навчання для прогнозування часових рядів

Глибоке навчання (Deep Learning) становить окремий клас підходів до прогнозування часових рядів, що ґрунтується на нейронних мережах із множиною прихованих шарів. На відміну від методів на основі дерев рішень, нейромережеві архітектури здатні навчатися безпосередньо на послідовних даних без попередньої трансформації у табличний формат [6].

Рекурентні нейронні мережі (RNN) обробляють послідовність елемент за елементом, передаючи прихований стан між кроками. Класичні RNN страждають від проблеми зникаючого градієнта, що обмежує їхню здатність моделювати довгострокові залежності. Архітектура LSTM (Long Short-Term Memory) [16] вирішує цю проблему завдяки механізму вентилів (gates), що контролюють потік інформації: вентиль забування визначає, яку частину попереднього стану зберегти, вхідний вентиль — яку нову інформацію додати, а вихідний — що передавати далі. LSTM набули широкого застосування у прогнозуванні часо-

вих рядів, проте їхня послідовна природа обмежує можливості паралелізації та збільшує час навчання.

Архітектура Transformer [17] використовує механізм самоуваги (self-attention) для паралельної обробки всіх елементів послідовності, що усуває обмеження послідовної обробки RNN. На основі Transformer розроблено спеціалізовані моделі для часових рядів, зокрема Temporal Fusion Transformer (TFT) [18], який поєднує механізм уваги з явною обробкою коваріатів різних типів (статичних, минулих та майбутніх) та забезпечує інтерпретованість прогнозів через механізм змінної селекції.

Незважаючи на теоретичну потужність, нейромережеві підходи мають суттєві практичні обмеження. Вони потребують значно більших обсягів даних та обчислювальних ресурсів для навчання порівняно з градієнтним бустингом, а також є чутливими до вибору гіперпараметрів. Результати великомасштабних змагань з прогнозування, зокрема M5 Competition, показали, що для табличних даних із помірною кількістю спостережень методи градієнтного бустинга часто демонструють порівнянну або вищу точність при значно нижчій обчислювальній складності [15].

1.7 Гібридні та ансамблеві підходи

У попередніх підрозділах розглянуто статистичні методи та методи машинного навчання як окремі класи підходів до прогнозування часових рядів. На практиці ці підходи часто комбінуються для компенсації взаємних обмежень [6].

1.7.1 Гібридні моделі

Гібридний підхід передбачає послідовне застосування моделей різних класів, де кожна нова модель працює із залишками попередньої [14]. Типова

схема полягає у декомпозиції часового ряду на детерміновану та стохастичну складові: лінійна модель описує передбачувані компоненти (тренд, сезонність, ефекти свят), а нелінійна модель навчається на залишках, що містять складніші залежності.

Лінійна регресія є природним кандидатом для моделювання детермінованих компонент [6]. У контексті часових рядів вона набуває вигляду:

$$\hat{y}_t^{\text{lin}} = \beta_0 + \gamma^\top x_t^{\text{det}}, \tag{1.16}$$

де x_t^{det} — вектор детермінованих ознак моменту часу t, що включає календарні маркери (день тижня, місяць, рік), індекс часу для опису тренду, статичні характеристики об'єкту прогнозування (категорія, локація) та зовнішні фактори, відомі заздалегідь (свята, промоакції). Параметри оптимізуються методом найменших квадратів, можливо із додаванням регуляризації, описаної у підрозділі 1.5.2.

Після навчання лінійної моделі обчислюються залишки:

$$e_t = y_t - \hat{y}_t^{\text{lin}}, \tag{1.17}$$

на яких навчається нелінійна модель (наприклад, градієнтний бустинг) $\hat{e}_t = g(z_t)$, де z_t — вектор ознак, що може включати лагові значення залишків, ковзні статистики та екзогенні змінні. Фінальний прогноз є сумою двох компонент:

$$\hat{y}_t = \hat{y}_t^{\text{lin}} + \hat{e}_t. \tag{1.18}$$

Такий підхід має кілька переваг [14]. По-перше, він вирішує проблему нездатності дерев рішень до екстраполяції: лінійна модель забезпечує коректне продовження тренду за межі навчальних даних. По-друге, розділення задачі на дві простіші підзадачі полегшує навчання нелінійної моделі, оскільки їй не потрібно «витрачати ємність» на опис регулярних компонент.

Варіантом гібридного підходу є stacked hybrid, де нелінійна модель отримує на вхід не лише залишки, а й прогноз лінійної моделі як додаткову ознаку, що дозволяє їй коригувати лінійний прогноз у обох напрямках [7].

1.7.2 Ансамблювання прогнозів

Ансамблювання прогнозів — це комбінація кількох самостійних моделей з метою отримання більш точного та стабільного результату. На відміну від гібридного підходу, де моделі застосовуються послідовно, в ансамблі моделі будують прогнози незалежно, після чого їх результати агрегуються [3].

Найпростішою формою ансамблювання є усереднення прогнозів:

$$\hat{y}_t = \frac{1}{K} \sum_{k=1}^{K} \hat{y}_t^{(k)}, \tag{1.19}$$

де $\hat{y}_t^{(k)}$ — прогноз k-ї моделі. Більш гнучким варіантом є зважене усереднення:

$$\hat{y}_t = \sum_{k=1}^{K} w_k \, \hat{y}_t^{(k)}, \quad \sum_{k=1}^{K} w_k = 1, \tag{1.20}$$

де ваги w_k можуть визначатись на основі якості моделей на валідаційній вибірці [3].

Ефективність ансамблювання ґрунтується на принципі диверсифікації: - якщо моделі допускають різні помилки, їх агрегація зменшує загальну дисперсію прогнозу [6]. Диверсифікація може досягатися кількома способами: використанням різних алгоритмів, різних наборів ознак, різних підвибірок навчальних даних або різних гіперпараметрів однієї й тієї самої моделі.

Останній варіант — ансамбль моделей одного алгоритму з різними конфігураціями — є особливо поширеним у задачах прогнозування часових рядів. Наприклад, моделі з різними горизонтами лагів (короткостроковий та довгостроковий контекст) фокусуються на різних аспектах динаміки ряду: модель із малими лагами краще відстежує останні зміни, тоді як модель із великими лагами захоплює річну сезонність та довгострокові тренди [7]. Їх комбінація дозволяє отримати прогноз, що враховує закономірності на різних часових масштабах.

Результати змагань M5 Competition продемонстрували, що переважна більшість найкращих рішень використовувала ту чи іншу форму ансамблюван-

ня [15]. Комбінація моделей виявилась одним із найбільш стабільних способів підвищення якості прогнозу в умовах великої кількості гетерогенних часових рядів.

1.8 Метрики якості прогнозування часових рядів

Вибір метрики якості суттєво впливає на процес побудови та оцінки прогнозних моделей, оскільки різні метрики по-різному зважують помилки та мають різну чутливість до характеристик даних [3]. У цьому підрозділі розглянуто основні метрики, що застосовуються у задачах прогнозування часових рядів продажів.

Середня абсолютна помилка (MAE)

MAE (Mean Absolute Error) обчислює середнє абсолютне відхилення прогнозу від фактичного значення:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{1.21}$$

MAE є інтуїтивно зрозумілою метрикою, що виражається у тих самих одиницях, що й цільова змінна, та є стійкою до викидів порівняно з квадратичними метриками [3]. Водночас MAE однаково зважує помилки незалежно від масштабу продажів, що може бути небажаним: помилка в 100 одиниць має різне значення для товару з продажами 10 000 та товару з продажами 200.

Корінь із середньоквадратичної помилки ($RMSE$)

$RMSE$ (Root Mean Squared Error) надає більшу вагу великим помилкам завдяки квадратичній функції:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{1.22}$$

$RMSE$ також виражається в одиницях цільової змінної, проте сильніше

штрафує за великі відхилення [3]. Це може бути як перевагою (якщо великі помилки є критичними для бізнесу), так і недоліком (якщо дані містять аномалії, що непропорційно впливають на метрику).

Коефіцієнт детермінації (R^2)

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{1.23}$$

де $\bar{y}$ — середнє значення цільової змінної. Значення $R^2 = 1$ відповідає ідеальному прогнозу, $R^2 = 0$ — прогнозу на рівні середнього, а від'ємні значення свідчать про те, що модель прогнозує гірше за константу [3]. Перевагою R^2 є масштабонезалежність та інтуїтивна інтерпретація, проте метрика чутлива до викидів та не враховує специфіку часових рядів (наприклад, наявність нульових значень).

Середня абсолютна відсоткова помилка ($MAPE$)

$MAPE$ (Mean Absolute Percentage Error) нормалізує помилку відносно фактичного значення:

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|. \tag{1.24}$$

$MAPE$ дозволяє порівнювати якість прогнозів для рядів різного масштабу. Однак метрика має суттєвий недолік: вона не визначена при $y_i = 0$ та стає нестабільною при малих значеннях y_i [3]. У контексті ритейлу, де нульові продажі є поширеним явищем (закриті магазини, відсутність товару), це робить $MAPE$ непридатною без додаткової обробки. Крім того, $MAPE$ є асиметричною: вона сильніше штрафує за завищення прогнозу, ніж за заниження тієї самої величини.

Корінь із середньоквадратичної логарифмічної помилки ($RMSLE$)

$RMSLE$ (Root Mean Squared Logarithmic Error) обчислюється як $RMSE$

у логарифмічному просторі:

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\ln(1+y_i) - \ln(1+\hat{y}_i)\right)^2}. \tag{1.25}$$

Додавання одиниці під логарифмом забезпечує коректну роботу при нульових значеннях продажів. $RMSLE$ має кілька важливих властивостей, що роблять її особливо придатною для прогнозування продажів [11].

По-перше, метрика оцінює відносну помилку: логарифмічне перетворення робить помилку пропорційною масштабу продажів. Відхилення прогнозу на 50 одиниць при фактичних продажах 100 штрафується значно сильніше, ніж те саме відхилення при продажах 10 000. Це є природним для ритейлу, де важлива саме відносна точність.

По-друге, $RMSLE$ сильніше штрафує за заниження прогнозу (underprediction), ніж за завищення тієї самої величини. Ця асиметрія відповідає бізнес-логіці: дефіцит товарів (наслідок заниженого прогнозу) зазвичай коштує дорожче, ніж помірний надлишок запасів.

По-третє, метрика є стійкою до викидів порівняно з $RMSE$, оскільки логарифмічне перетворення стискає діапазон значень.

$RMSLE$ широко застосовується як цільова метрика у змаганнях з прогнозування продажів та попиту [15]. Оптимізація моделі під $RMSLE$ може здійснюватись безпосередньо (використовуючи відповідну функцію втрат) або шляхом логарифмічного перетворення цільової змінної з подальшою оптимізацією MSE у логарифмічному просторі.

Вибір метрики

Вибір метрики залежить від специфіки задачі та бізнес-вимог [3]. Для задач із часовими рядами різного масштабу та наявністю нульових значень перевагу мають метрики, що працюють у відносному або логарифмічному просторі. У таблиці 1.1 наведено порівняння розглянутих метрик за ключовими властивостями.

Таблиця 1.1 – Порівняння метрик якості прогнозування

Властивість	*MAE*	*RMSE*	R^2	*MAPE*	*RMSLE*
Стійкість до викидів	+	–	–	+	+
Робота з $y_i = 0$	+	+	+	–	+
Масштабонезалежність	–	–	+	+	+
Штраф за великі помилки	–	+	+	–	+

1.9 Стратегії валідації для часових рядів

Оцінка якості прогнозних моделей для часових рядів вимагає спеціальних підходів до валідації, що враховують темпоральну структуру даних. Стандартна випадкова крос-валідація, широко застосовувана в задачах машинного навчання, є некоректною для часових рядів, оскільки порушує хронологічний порядок спостережень і призводить до витоку інформації з майбутнього у навчальну вибірку [3].

Проста розбивка (Hold-out)

Найпростішою стратегією є одноразова розбивка даних на навчальну та тестову частини за часовою межею t^* [3]:

$$\mathcal{D}_{\text{train}} = \{(y_t, x_t) : t \leq t^*\}, \quad \mathcal{D}_{\text{test}} = \{(y_t, x_t) : t > t^*\}. \tag{1.26}$$

Модель навчається виключно на даних до моменту t^* та оцінюється на даних після нього. Перевагою є простота та обчислювальна ефективність, проте оцінка якості базується лише на одному часовому відрізку й може бути нестабільною [7].

Ковзна крос-валідація (Time Series Cross-Validation)

Для отримання більш надійної оцінки використовується ковзна крос-валідація, також відома як expanding window або rolling origin [3]. На кожному кроці $k = 1, \ldots, K$ навчальне вікно розширюється, а модель оцінюється на відрізку фіксованої довжини h (горизонт прогнозування), що йде безпосередньо за ме-

жею вікна:

$$\mathcal{D}_{\text{train}}^{(k)} = \{y_t : t \leq t_0 + k\}, \quad \mathcal{D}_{\text{test}}^{(k)} = \{y_t : t_0 + k < t \leq t_0 + k + h\}, \tag{1.27}$$

де t_0 — мінімальний розмір початкового навчального вікна. Фінальна оцінка обчислюється як середнє метрики якості по всіх фолдах.

Варіантом є ковзне вікно фіксованого розміру (sliding window), де навчальна вибірка має сталу довжину W:

$$\mathcal{D}_{\text{train}}^{(k)} = \{y_t : t_0 + k - W < t \leq t_0 + k\}. \tag{1.28}$$

Фіксоване вікно може бути доцільним за наявності суттєвої нестаціонарності, коли старі дані втрачають релевантність [7].

Витік даних у часових рядах

Окремою проблемою є витік даних (data leakage) — ситуація, коли модель під час навчання отримує доступ до інформації, що не буде доступна в момент реального прогнозування [7]. У контексті часових рядів типовими джерелами витоку є:

Використання майбутніх значень при конструюванні ознак — наприклад, обчислення ковзного середнього з вікном, що включає майбутні спостереження, або застосування нормалізації на основі статистик усього набору даних замість лише навчальної частини [14].

Некоректна класифікація коваріатів — використання змінних, значення яких не будуть відомі на момент прогнозування, як майбутніх коваріатів (future covariates). Наприклад, фактична кількість покупців у магазині є минулим коваріатом (past covariate), оскільки вона невідома наперед, і її не можна подавати на вхід моделі для майбутніх періодів [7].

Обчислення лагових ознак без урахування горизонту прогнозування — при багатокроковому прогнозуванні на h кроків уперед мінімально допустимий лаг дорівнює h, а не 1, оскільки значення $y_{t-1}, \ldots, y_{t-h+1}$ ще не будуть спостережені на момент прогнозу [14].

Забезпечення коректності валідації є критичним для отримання реалістичної оцінки якості моделі. Завищена оцінка внаслідок витоку даних може при-

звести до вибору моделі, що погано працює в реальних умовах.

1.10 Постановка задачі дипломної роботи

На основі проведеного аналізу предметної області та огляду методів прогнозування можна сформулювати задачу дипломної роботи.

Об’єктом дослідження є процес прогнозування обсягів продажів у мережі роздрібної торгівлі з урахуванням впливу зовнішніх факторів.

Предметом дослідження є методи машинного навчання для прогнозування багатовимірних часових рядів продажів із використанням екзогенних коваріатів.

Мета роботи — розробка інтелектуальної системи прогнозування часових рядів продажів, яка ефективно інтегрує екзогенні фактори різної природи (макроекономічні, календарні, комерційні) для підвищення точності прогнозів порівняно з базовими підходами.

У рамках дипломної роботи передбачається виконання таких задач: аналіз існуючих підходів до прогнозування часових рядів продажів, що охоплює статистичні методи, методи машинного навчання та глибокого навчання; розвідувальний аналіз даних реального набору продажів мережі роздрібної торгівлі з дослідженням трендів, сезонності, аномалій та взаємозв’язків між змінними; розробка стратегії передобробки даних та конструювання ознак з урахуванням специфіки часових рядів та класифікації екзогенних факторів; побудова та навчання прогнозних моделей різних класів із забезпеченням коректної стратегії валідації для часових рядів; порівняльний аналіз моделей за обраними метриками якості та дослідження впливу окремих груп екзогенних факторів на точність прогнозу.

Огляд, поданий вище, охоплює широке коло методів — від класичних статистичних моделей до нейромережевих архітектур, — однак не всі вони однаково придатні для задачі з тисячами часових рядів та багатовимірним простором екзогенних коваріатів. Тому в експериментальній частині фокус зосереджено на тих класах моделей, які за результатами наукової літератури демонструють

найкращий баланс точності, обчислювальної складності та гнучкості в роботі з ознаками різних типів. Особлива увага приділена розробці ознакового простору та дослідженню внеску окремих груп ознак, оскільки за специфіки задачі саме структура ознак, а не вибір конкретного алгоритму, найбільшою мірою визначає якість прогнозу.

Практична цінність роботи полягає у тому, що розроблені підходи та отримані результати можуть бути застосовані для прогнозування обсягів продажів у мережах роздрібної торгівлі, що дозволяє покращити управління запасами, планування закупівель та операційне планування мережі.

Висновки до розділу 1

У першому розділі досліджено предметну область прогнозування часових рядів продажів та проведено огляд методів, що застосовуються до цієї задачі.

Обґрунтовано актуальність задачі: роздрібна торгівля є однією з найбільших галузей світової економіки, і навіть незначне підвищення точності прогнозування попиту здатне суттєво зменшити втрати від дисбалансу запасів та підвищити ефективність операційного управління.

Проаналізовано ключові особливості часових рядів продажів у роздрібній торгівлі: множинна сезонність, висока розмірність, наявність нульових та переривчастих рядів, нестаціонарність та залежність від зовнішніх факторів. Ці особливості визначають вимоги до прогнозних моделей та обґрунтовують необхідність застосування підходів, здатних працювати з багатовимірними табличними даними.

Розглянуто роль екзогенних факторів у прогнозуванні продажів та запропоновано їх класифікацію за природою впливу, доступністю на момент прогнозування та характером взаємодії з цільовою змінною. Показано, що врахування екзогенних факторів є необхідною умовою побудови якісної прогнозної моделі.

Розглянуто статистичні методи — ARIMA, ETS та Prophet, — які забезпечують теоретичну обґрунтованість та інтерпретованість, проте мають спільні

обмеження: лінійність припущень, підхід «одна модель — один ряд» та обмежені можливості інтеграції екзогенних факторів.

Проаналізовано методи машинного навчання — лінійні моделі з регуляризацією (Ridge-регресія), дерева рішень та градієнтний бустинг (LightGBM), — які здатні моделювати нелінійні залежності, працювати з великою кількістю ознак різних типів та будувати глобальні моделі на тисячах часових рядів одночасно. Описано ключові алгоритмічні інновації LightGBM: GOSS, EFB та leaf-wise стратегію зростання дерев.

Розглянуто методи глибокого навчання — LSTM та Transformer-based моделі, — які здатні працювати безпосередньо з послідовними даними, проте потребують значно більших обчислювальних ресурсів і на табличних даних часто поступаються градієнтному бустингу.

Розглянуто гібридні та ансамблеві підходи: гібридні моделі, що поєднують лінійну компоненту для детермінованих складових ряду з нелінійною моделлю для залишків, та ансамблювання прогнозів кількох моделей із різними конфігураціями для підвищення стабільності та точності результату.

Проаналізовано метрики якості прогнозування — MAE, $RMSE$, R^2, $MAPE$ та $RMSLE$. Показано, що $RMSLE$ є особливо придатною для задач прогнозування продажів завдяки масштабонезалежності, коректній роботі з нульовими значеннями та асиметричному штрафу, що відповідає бізнес-логіці.

Описано стратегії валідації для часових рядів — hold-out та ковзну крос-валідацію, — а також типові джерела витоку даних, що можуть призводити до завищення оцінки якості моделі.

На основі проведеного аналізу сформульовано мету та задачі дипломної роботи, визначено об’єкт, предмет та практичну цінність дослідження. Проведений огляд показав, що для задач прогнозування з тисячами часових рядів та багатовимірним простором екзогенних факторів методи градієнтного бустингу демонструють найкращий баланс точності, обчислювальної складності та гнучкості у роботі з ознаками різних типів. Цей висновок визначає вектор подальшого дослідження, у якому центральну роль відіграватиме стратегія конструювання ознак та емпіричний аналіз внеску окремих груп екзогенних коваріатів у якість прогнозу.

РОЗДІЛ 2
АНАЛІЗ ДАНИХ ТА КОНСТРУЮВАННЯ ОЗНАК

2.1 Опис набору даних та його структури

Для дослідження вибрано набір даних великої мережі продовольчих магазинів Еквадору [21], що містить щоденні продажі 54 магазинів за 33 товарними категоріями протягом періоду з 1 січня 2013 року по 15 серпня 2017 року. Набір охоплює 1684 дні спостережень та формує 1782 часових ряди (54×33), що в сумі складає понад 3 мільйони записів. Горизонт прогнозування — 16 днів (з 16 по 31 серпня 2017 року).

Географічний контекст є суттєвим для розуміння даних: Еквадор належить до країн, економіка яких значною мірою залежить від видобутку та експорту нафти — у 2017 році нафта забезпечувала близько третини доходів державного бюджету та 32% експортних надходжень [22]. Це робить макроекономічну ситуацію — а відтак і споживчий попит — вразливою до коливань світових цін на енергоносії.

Вибір саме цього набору зумовлений кількома факторами. По-перше, дані містять багатий набір екзогенних факторів різної природи та класифікації: макроекономічні (ціна нафти), календарні (свята національного, регіонального та локального масштабу), комерційні (промоакції) та операційні (кількість покупців). Ця різноманітність дозволяє дослідити вплив факторів різних типів на якість прогнозу. По-друге, набір демонструє типові особливості роздрібної торгівлі, описані у розділі 1: множинну сезонність, високу розмірність, нульові та переривчасті ряди, структурні зломи (зокрема, наслідки землетрусу у квітні 2016 року). По-третє, наявність метаданих магазинів (місто, штат, тип, кластер) забезпечує статичні коваріати, а розділення екзогенних змінних на минулі (транзакції), майбутні (свята, промоакції) та статичні коваріати відповідає класифікації, введеній у підрозділі 1.3.

Дані організовані у п’ять взаємопов’язаних таблиць.

Основна таблиця продажів

Центральним елементом набору є таблиця щоденних продажів, що містить 3 000 888 записів з такими полями: дата спостереження (`date`), ідентифікатор магазину (`store_nbr`), товарна категорія (`family`), обсяг продажів у одиницях товару (`sales`) та кількість товарних позицій, що перебували на промоакції (`onpromotion`). Обсяг продажів є цільовою змінною для прогнозування; значення може бути дробовим, оскільки деякі товари продаються на вагу. Поле `onpromotion` вказує кількість одиниць складського обліку (Stock Keeping Unit, SKU) відповідної категорії, на які діяла промоакція у певному магазині в певний день. 33 товарні категорії охоплюють широкий спектр — від продовольчих товарів першої необхідності (GROCERY, BEVERAGES, PRODUCE, DAIRY) до непродовольчих (AUTOMOTIVE, HOME APPLIANCES, BOOKS). Приклад записів наведено у таблиці 2.1.

Таблиця 2.1 – Перші записи таблиці продажів

id	*date*	*store_nbr*	*family*	*sales*	*onpromotion*
0	2013-01-01	1	AUTOMOTIVE	0.0	0
1	2013-01-01	1	BABY CARE	0.0	0
2	2013-01-01	1	BEAUTY	0.0	0
3	2013-01-01	1	BEVERAGES	0.0	0
4	2013-01-01	1	BOOKS	0.0	0

Метадані магазинів

Таблиця магазинів містить статичні характеристики кожного з 54 магазинів: місто (`city`), штат (`state`), тип магазину (`type`) та кластер (`cluster`). Магазини розташовані у 22 містах 16 штатів, із найбільшою концентрацією у Кіто (18 магазинів) та Гуаякілі (8 магазинів). Тип магазину приймає одне з п’яти значень (A–E), а кластер — одне з 17 значень; кластер є групуванням магазинів за подібністю характеристик. Приклад записів наведено у таблиці 2.2.

Ціна нафти

Набір містить щоденні ціни нафти марки WTI (West Texas Intermediate) за весь період спостережень, включаючи як навчальний, так і тестовий періоди. Дані подано лише за робочі дні біржі, що зумовлює наявність пропусків у

Таблиця 2.2 – Перші записи таблиці магазинів

store_nbr	*city*	*state*	*type*	*cluster*
1	Quito	Pichincha	D	13
2	Quito	Pichincha	D	13
3	Quito	Pichincha	D	8
4	Quito	Pichincha	D	9
5	Santo Domingo	Santo Domingo de los Tsachilas	D	4

вихідні та святкові дні (3,5% записів). Економіка Еквадору є нафтозалежною та вразливою до шоків цін на нафту: коливання нафтових цін безпосередньо впливають на купівельну спроможність населення та обсяги роздрібних продажів. Приклад записів наведено у таблиці 2.3.

Таблиця 2.3 – Перші записи таблиці цін на нафту

date	*dcoilwtico*
2013-01-02	93.14
2013-01-03	92.97
2013-01-04	93.12
2013-01-07	93.20

Свята та події

Таблиця свят та подій містить 350 записів із зазначенням дати, типу, масштабу (`locale`: National, Regional, Local) та текстового опису (таблиця 2.4). Національні свята (174 записи) впливають на всі магазини, регіональні (24) — на магазини відповідного штату, а локальні (152) — на магазини конкретного міста.

Таблиця 2.4 – Перші записи таблиці свят та подій

date	*type*	*locale*	*locale_name*	*description*	*transferred*
2012-03-02	Holiday	Local	Manta	Fundacion de Manta	False
2012-04-01	Holiday	Regional	Cotopaxi	Provincializacion de Cotopaxi	False
2012-04-12	Holiday	Local	Cuenca	Fundacion de Cuenca	False
2012-04-14	Holiday	Local	Libertad	Cantonizacion de Libertad	False
2012-04-21	Holiday	Local	Riobamba	Cantonizacion de Riobamba	False

Поле `type` набуває значення Holiday, Event, Transfer, Additional, Bridge,

Work Day і дозволяє розрізняти регулярні свята, разові події, перенесені вихідні та компенсаційні робочі дні. Перенесення офіційного свята на інший день означає, що оригінальна дата стає звичайним робочим днем, тоді як день переносу набуває святкового статусу.

Транзакції

Таблиця транзакцій містить 83 488 записів щоденної кількості транзакцій (чеків) для кожного магазину. Ця змінна є минулим коваріатом (past covariate): на момент прогнозування відомі лише її історичні значення, оскільки кількість транзакцій у майбутніх періодах невідома. Приклад записів наведено у таблиці 2.5.

Таблиця 2.5 – Перші записи таблиці транзакцій

date	*store_nbr*	*transactions*
2013-01-01	25	770
2013-01-02	1	2111
2013-01-02	2	2358
2013-01-02	3	3487
2013-01-02	4	1922

Додаткові особливості даних

Окрім структурованих таблиць, набір даних характеризується кількома важливими контекстуальними факторами. Заробітні плати у державному секторі виплачуються двічі на місяць — 15-го числа та в останній день місяця, — що може створювати регулярні піки продажів у ці дати. Крім того, 16 квітня 2016 року стався землетрус магнітудою 7,8, після якого населення масово закуповувало воду та товари першої необхідності, що суттєво вплинуло на динаміку продажів протягом кількох тижнів після події. Ця аномалія є прикладом структурного злому, описаного у підрозділі 1.2.

Зв'язки між таблицями

Зв'язки між таблицями організовано за допомогою спільних ключів: поле `store_nbr` пов'язує таблицю продажів із метаданими магазинів та транзакціями, а поле `date` — із цінами нафти та таблицею свят. Для побудови єдиного навчального набору необхідне об'єднання (join) усіх таблиць, що дозволяє для

кожного спостереження (date, store_nbr, family) отримати повний набір ознак: характеристики магазину, ціну нафти, інформацію про свята та кількість транзакцій. Структура зв'язків наведена на рисунку 2.1.

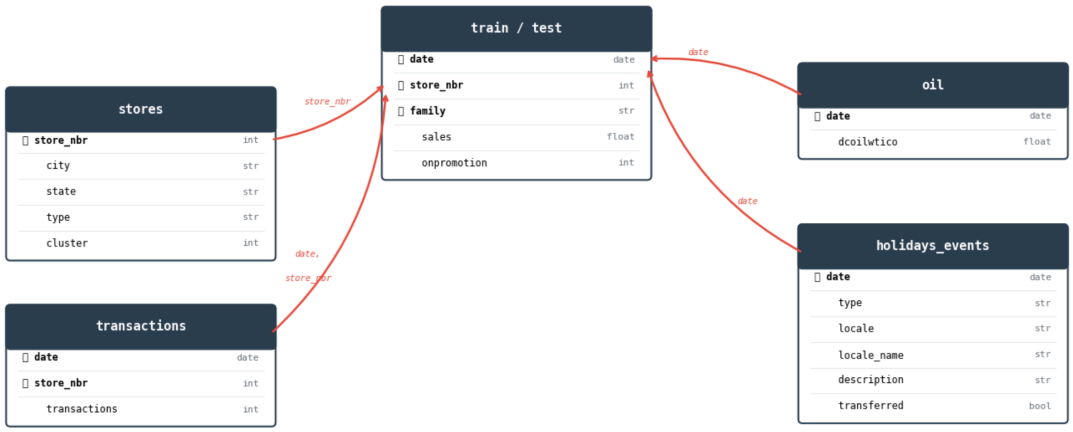

Рисунок 2.1 – Структура набору даних та зв'язки між таблицями

2.2 Розвідувальний аналіз даних

Розвідувальний аналіз даних має на меті виявити основні закономірності та особливості, що визначатимуть стратегію передобробки, конструювання ознак та вибору моделей прогнозування.

Розподіл цільової змінної

Розподіл обсягу продажів є сильно правоскошеним (рис. 2.2): медіана складає лише 11 одиниць при середньому значенні 358, коефіцієнт асиметрії дорівнює 7,4, а максимум перевищує 125 тисяч. Значна частина спостережень (31,3%) має нульові продажі. Нулі виникають з кількох причин: закриття магазинів у святкові дні (зокрема, на Різдво дані повністю відсутні за всі роки), відсутність окремих товарних категорій у деяких магазинах (53 часових ряди є повністю нульовими), а також наявність так званих leading zeros — початкових періодів без продажів у рядах, де категорія з'явилась пізніше за дату початку спостережень. Логарифмічне перетворення $\ln(1 + y)$ суттєво покращує розпо-

діл, наближаючи його до нормального.

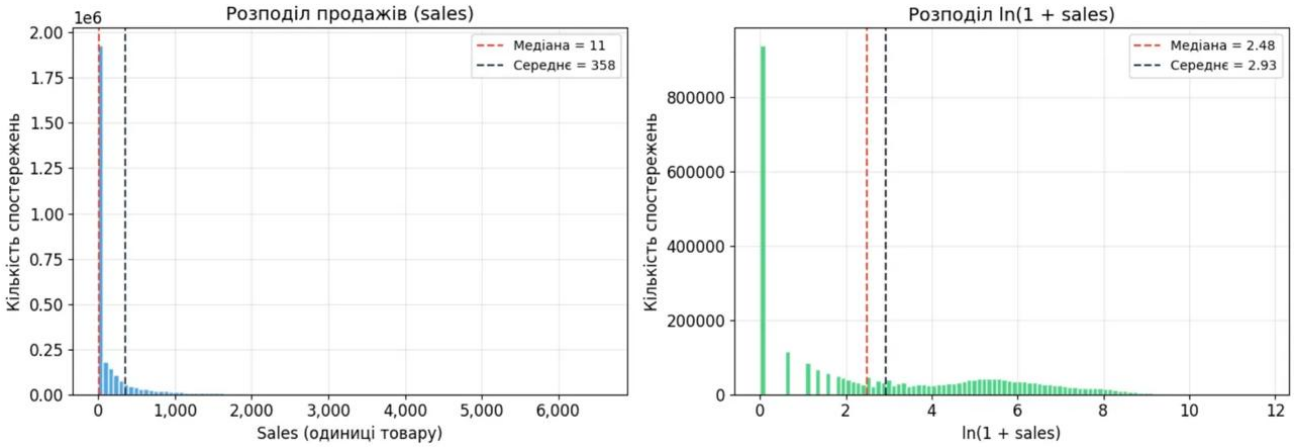

Рисунок 2.2 – Розподіл продажів: оригінальний (ліворуч) та після логарифмічного перетворення (праворуч)

Аналіз продажів за категоріями

Масштаб продажів суттєво відрізняється між 33 товарними категоріями (рис. 2.3). Домінуючою є GROCERY I (32% сумарних продажів), за нею — BEVERAGES (20%) та PRODUCE (11%). На іншому полюсі — BOOKS, BABY CARE та HOME APPLIANCES із часткою менше 0,01% та нульовими продажами у 73–97% спостережень. Така гетерогенність означає, що глобальна модель повинна одночасно працювати з рядами, що відрізняються за масштабом на кілька порядків, а ознакова структура має включати інформацію про товарну категорію як статичний коваріат.

Тренди

Загальна динаміка продажів мережі демонструє виражений висхідний тренд (рис. 2.4): середні щоденні продажі зросли з 0,39M у 2013 до 0,86M у 2017 році, тобто більш ніж удвічі. Це зростання пов’язане як із відкриттям нових магазинів, так і зі збільшенням обсягів продажів існуючих. Наявність такого вираженого тренду має враховуватись при побудові моделі: алгоритми на основі дерев рішень не здатні екстраполювати значення цільової змінної за межі діапазону навчальних даних. Це робить необхідним явне включення в ознаковий простір змінних, що описують часову динаміку, — зокрема індексів часу та ковзних статистик за різні вікна.

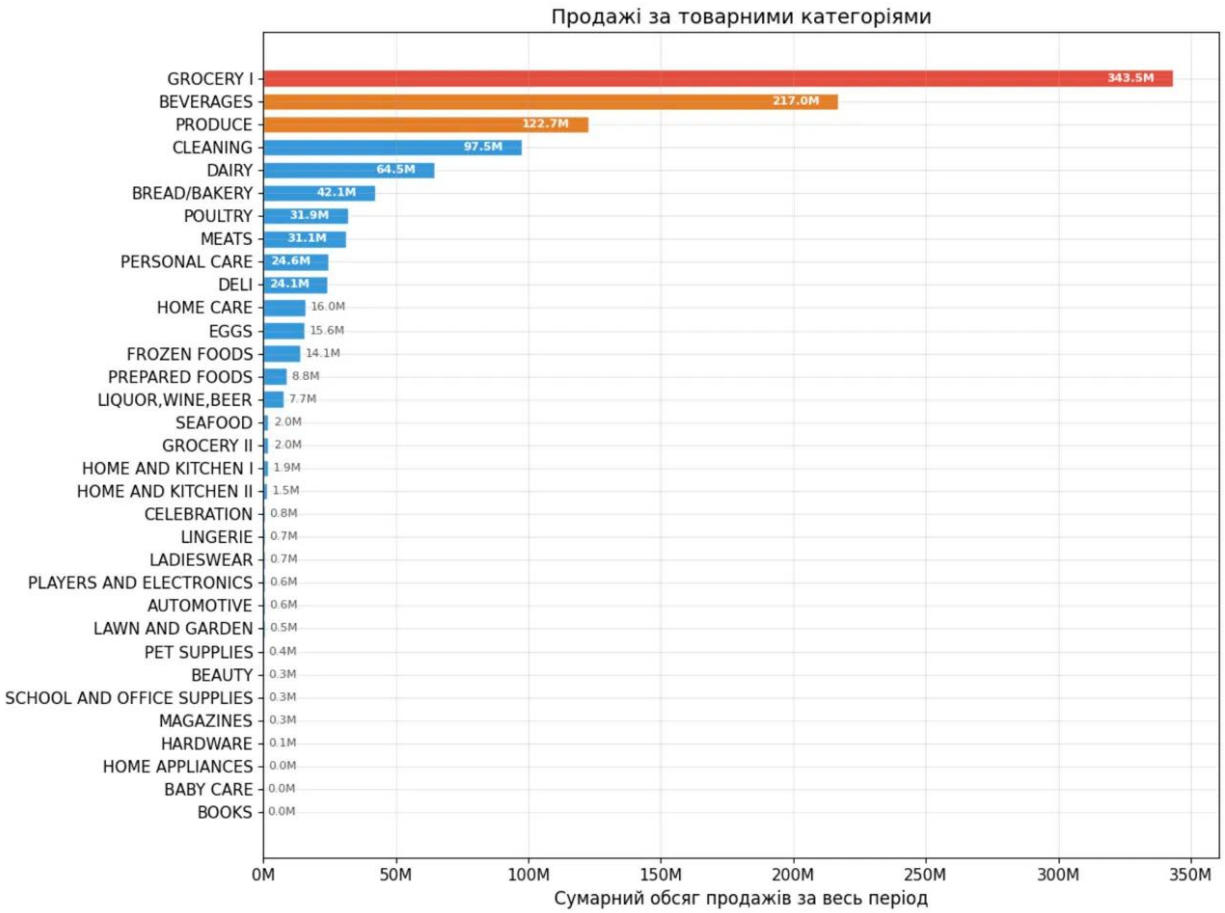

Рисунок 2.3 – Сумарний обсяг продажів за товарними категоріями за весь період спостережень

Сезонність

Дані демонструють множинну сезонність на кількох рівнях. Тижнева сезонність є найбільш вираженою: максимум продажів припадає на неділю (0,83M) та суботу (0,77M), а мінімум — на четвер (0,51M). Річна сезонність проявляється у грудневому піку (0,81M, що на 25% вище за середньорічний рівень), пов’язаному зі святковим сезоном. Крім того, спостерігається ефект дня місяця: продажі у перші дні місяця є систематично вищими, що узгоджується з графіком виплати заробітних плат у державному секторі (15-го числа та в останній день місяця). Наявність множинної сезонності визначає вимоги до календарних ознак: модель повинна мати доступ до інформації про день тижня, день місяця, місяць та інші маркери, що дозволяють розрізняти періоди різної інтенсивності продажів.

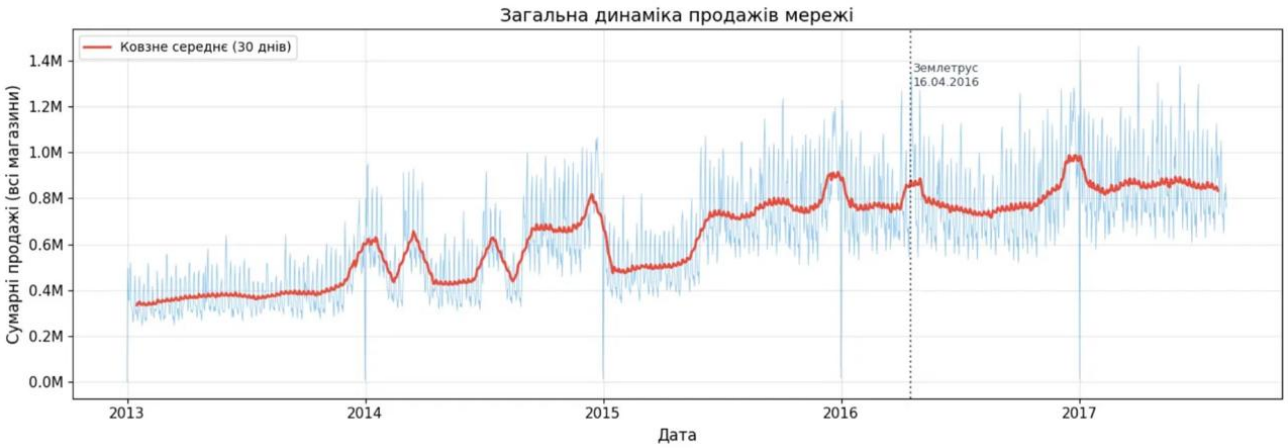

Рисунок 2.4 – Загальна динаміка продажів мережі з ковзним середнім за 30 днів

Екзогенні фактори

Аналіз впливу екзогенних факторів виявив низку неоднозначних залежностей, що визначають стратегію їх використання при моделюванні.

Кількість транзакцій (чеків) демонструє сильну кореляцію з обсягом продажів на рівні окремого магазину ($r = 0{,}84$). Це водночас і перевага, і обмеження: транзакції є минулим коваріатом, що вимагає лагування на горизонт прогнозування, а висока кореляція з продажами означає, що історична кількість транзакцій несе значною мірою ту саму інформацію, що й історичні продажі — з ризиком дублювання сигналу. Питання, чи додає ця змінна нову інформацію порівняно з лаговими значеннями цільової змінної, потребує окремого емпіричного аналізу.

Ціна нафти WTI на перший погляд демонструє помітну від'ємну кореляцію з продажами (від $-0{,}41$ до $-0{,}65$ залежно від категорії). Однак після видалення тренду (detrending) кореляція зникає повністю ($r \approx 0$ для всіх категорій), що свідчить про спурійний характер лінійного зв'язку: протягом періоду спостережень нафта дешевшала, а продажі зростали через незалежні причини. Теоретична значущість нафтових цін для нафтозалежної економіки залишається обґрунтованою, проте відсутність вимірюваної кореляції після усунення тренду ставить під сумнів їх корисність як прямого предиктора.

Промоакції мають суттєвий та диференційований вплив на продажі: ефект

коливається від +43% для категорії CLEANING до +208% для PRODUCE. Національні свята підвищують продажі у будні дні в середньому на 28% (медіана 0,75M проти 0,58M), тоді як вплив регіональних та локальних свят є значно меншим.

Аномалії та структурні зломи

Найбільш помітною аномалією є землетрус магнітудою 7,8, що стався 16 квітня 2016 року на узбережжі Еквадору [23] (рис. 2.5). Після нього спостерігається різке зростання продажів, особливо товарів першої необхідності: PERSONAL CARE (+34%), GROCERY I (+32%), BEVERAGES (+19%). Ефект тривав приблизно 3–4 тижні з поступовим згасанням. Ця аномалія є прикладом структурного злому, описаного у підрозділі 1.2, і потребує спеціальної обробки при конструюванні ознак, що детально розглядається у підрозділі 2.4.5.

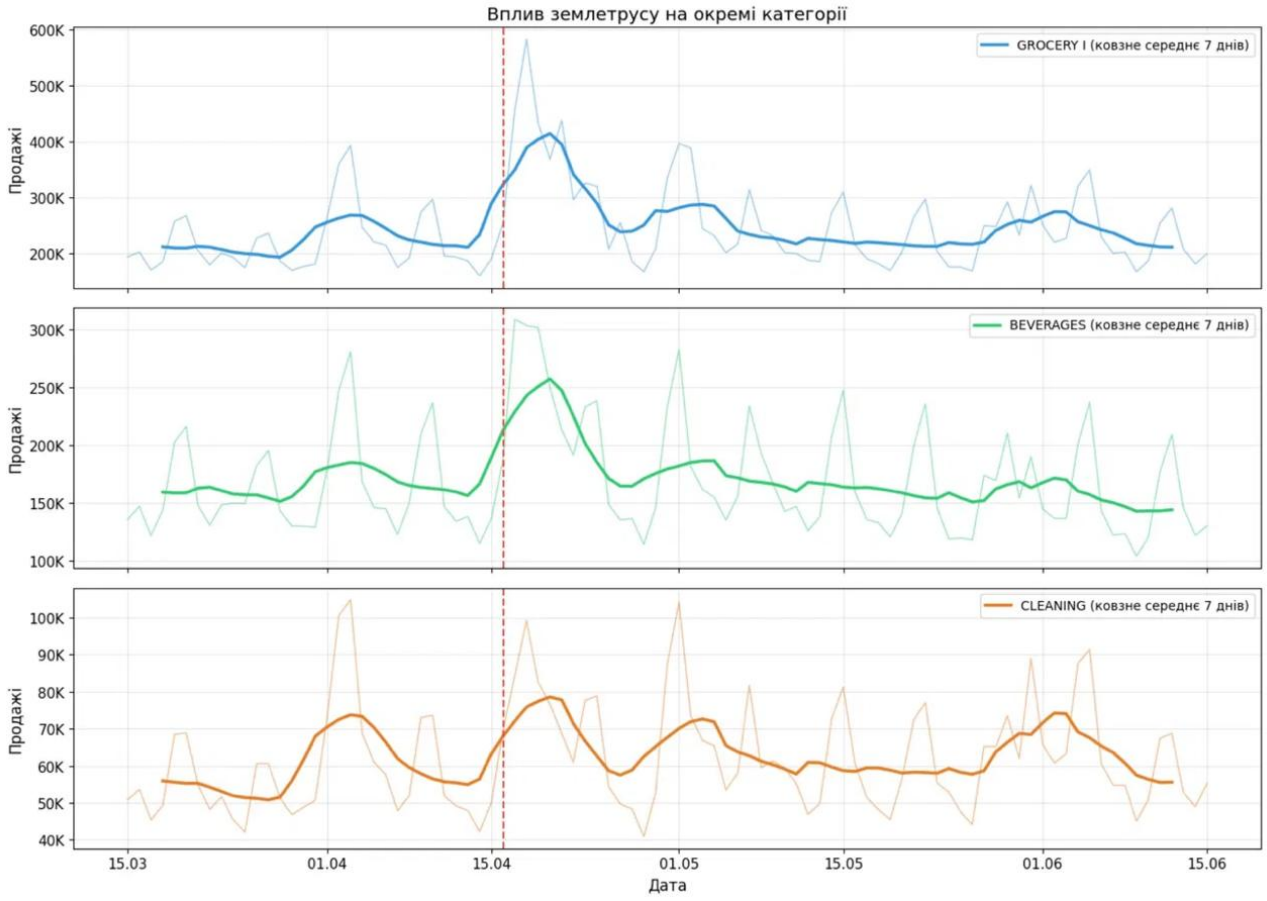

Рисунок 2.5 – Вплив землетрусу 16 квітня 2016 року на продажі трьох найбільших категорій

Іншою регулярною аномалією є Різдво (25 грудня): дані за ці дати повністю відсутні у навчальному наборі, що свідчить про закриття всіх магазинів. Новий рік також характеризується суттєво зниженими продажами.

Пропущені значення

Пропуски у даних мають різну природу. Ціна нафти містить 3,5% пропусків, що відповідають вихідним та святковим дням біржі. Таблиця транзакцій має 8,4% відсутніх записів — частково через закриття магазинів, частково через відсутність даних. Чотири дати (Різдво кожного року) повністю відсутні у навчальних даних. Стратегія обробки цих пропусків докладно описана у підрозділі 2.3.1.

2.3 Передобробка даних

Виявлені у підрозділі 2.2 особливості сирих даних — пропуски різного походження, сильна правоскошеність розподілу цільової змінної та фрагментована структура даних, розподілена між п’ятьма таблицями, — потребують серії підготовчих процедур перед побудовою прогнозних моделей. У цьому підрозділі описано три такі процедури: обробку пропусків (підрозділ 2.3.1), логарифмічне перетворення цільової змінної (підрозділ 2.3.2) та об’єднання таблиць у єдиний навчальний набір (підрозділ 2.3.3).

2.3.1 Обробка пропусків

Пропуски у наборі даних мають три різні джерела, кожне з яких потребує власної стратегії обробки.

Пропуски в основній таблиці продажів

Як було показано у підрозділі 2.2, у навчальному наборі повністю відсу-

тні чотири дати — Різдво (25 грудня) у кожному з 2013–2016 років, — що відповідає закриттю всіх магазинів. Залишення цих пропусків у вихідних даних ускладнило б подальші операції з часовими рядами (зокрема, обчислення лагових ознак та ковзних статистик), оскільки нерегулярність часової сітки порушує припущення стандартних реалізацій таких ознак. Тому таблицю продажів переіндексовано на повний декартів добуток (`date` $\times$ `store_nbr` $\times$ `family`) за повним діапазоном дат, а нові рядки заповнено значеннями `sales` $= 0$ та `onpromotion` $= 0$. Така інтерпретація є коректною: магазини були зачинені, фактичні продажі дорівнювали нулю. Після переіндексації набір містить 1 688 днів та 3 008 016 рядків ($1\,688 \times 54 \times 33$).

Пропуски у ціні нафти

Таблиця нафти містить ціни лише за робочі дні біржі (близько 71% повного календаря) — WTI є фінансовим контрактом, що не торгується у вихідні та біржові свята. Відсутність значень у такі дні не означає, що ціна нафти невідома: умовно її поточне значення дорівнює ціні останньої торгової сесії. Тому таблицю переіндексовано на повний денний календар, що охоплює як навчальний, так і тестовий періоди (1 704 дні), та застосовано пряму імпутацію (forward-fill). Для першого рядка (1 січня 2013 року, святковий день, дані за який недоступні з самого початку) застосовано зворотну імпутацію (back-fill) із найближчого подальшого спостереження. Така стратегія є стандартною для повільнозмінних макроекономічних показників.

Пропуски в таблиці транзакцій

У таблиці транзакцій бракує 8,4% рядків (7 664 із 91 152 можливих комбінацій магазин $\times$ день). Аналіз показав, що пропуски не є випадковими, а мають дві систематичні причини: частина магазинів відкривалась пізніше за 1 січня 2013 року (тому за відповідний період не існувало даних взагалі), а ще близько 10 днів на магазин — це дні закриття, що збігаються зі святами. Таблицю транзакцій переіндексовано на регулярну сітку (`date` $\times$ `store_nbr`), однак пропущені значення *не імпутовано*, а залишено як `NaN`. Це рішення обґрунтоване двома аспектами. По-перше, обрана модель LightGBM має вбудовану підтримку відсутніх значень: при пошуку розбиття дерева кожен `NaN` може бути спрямований у ліву або праву гілку залежно від того, що зменшує функцію втрат [13].

По-друге, сама *відсутність* значення несе інформативний сигнал (магазин ще не відкрився або зачинений), і його заповнення штучним значенням спотворило б цю інформацію.

Підсумок процедур обробки пропусків наведено у таблиці 2.6.

Таблиця 2.6 – Підсумок обробки пропусків

Таблиця	*До обробки*	*Після обробки*	*Метод обробки*
train	3 000 888	3 008 016	Reindex; `sales`=0, `onpromotion`=0
oil	1 218	1 704	Reindex + forward-fill (+ back-fill)
transactions	83 488	91 152	Reindex; `NaN` збережено для LightGBM

2.3.2 Логарифмічне перетворення цільової змінної

Розвідувальний аналіз показав, що розподіл цільової змінної `sales` є сильно правоскошеним: при медіані 11 одиниць середнє значення складає 358, коефіцієнт асиметрії дорівнює 7,4, а 31,3% спостережень дорівнює нулю. Пряме навчання моделі на необробленій величині продажів призвело б до того, що функція втрат була б повністю домінована поодинокими великими спостереженнями, тоді як точність прогнозу на типових малих магазинах залишалася б другорядною. Для подолання цього ефекту цільову змінну перетворено за формулою:

$$y' = \ln(1 + y), \tag{2.1}$$

де y — спостережувана величина продажів, а y' — величина після перетворення. Обернене перетворення:

$$y = \exp(y') - 1. \tag{2.2}$$

Формула (2.1) має три важливі властивості. По-перше, вона коректно визначена для $y = 0$: $\ln(1+0) = 0$, що дозволяє без додаткової обробки працювати з нульовими продажами. По-друге, логарифмічне перетворення суттєво зменшує асиметрію розподілу: для нашого набору коефіцієнт асиметрії після перетворення зменшується з 7,4 до близько 0,5, наближаючи розподіл до нормально-

го (рисунок 2.2, права частина). По-третє, перетворення переводить мультиплікативні впливи в адитивні. Багато факторів, що визначають продажі — зокрема, промоакції та свята, — діють приблизно мультиплікативно: масштабують продажі у певну кількість разів, а не додають фіксовану кількість одиниць. У логарифмічному просторі такі ефекти перетворюються на адитивні константи, що відповідає природній структурі регресійних моделей.

Окреме значення має зв'язок між логарифмічним перетворенням та метрикою якості RMSLE, визначеною у підрозділі 1.8. Навчання моделі зі стандартною функцією втрат MSE на перетворених значеннях y' є точно еквівалентним прямій оптимізації RMSLE на оригінальних значеннях y. Дійсно, позначивши $y'_i = \ln(1 + y_i)$ та $\hat{y}'_i = \ln(1 + \hat{y}_i)$, отримуємо:

$$\mathrm{RMSLE}^2(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\big(\ln(1 + y_i) - \ln(1 + \hat{y}_i)\big)^2 = \frac{1}{n}\sum_{i=1}^{n}(y'_i - \hat{y}'_i)^2 = \mathrm{MSE}(y', \hat{y}'). \tag{2.3}$$

Тобто $\mathrm{RMSLE}(y, \hat{y}) = \sqrt{\mathrm{MSE}(y', \hat{y}')}$. На практиці це означає, що достатньо подати на вхід моделі перетворені значення y', використовувати стандартну функцію втрат `l2`, і отриманий під час навчання оптимум відповідатиме оптимуму за метрикою RMSLE на оригінальних значеннях. На етапі прогнозу зворотне перетворення $\hat{y} = \exp(\hat{y}') - 1$ повертає результат в одиниці оригінальної шкали; отриманий прогноз додатково обмежується знизу нулем, оскільки реальні продажі не можуть бути від'ємними.

2.3.3 Об'єднання таблиць

Кінцевим кроком передобробки є об'єднання очищених таблиць у єдиний навчальний набір. Бажаний формат — довга (long-format) таблиця, у якій кожен рядок є унікальною комбінацією (`date`, `store_nbr`, `family`), а стовпці містять усі доступні ознаки.

Структура об'єднання визначається такими ключами зв'язку. Поле `sto-`

re_nbr зв'язує основну таблицю з метаданими магазинів за принципом «один-до-багатьох»; поле date — із цінами нафти (також «один-до-багатьох»); пара (date, store_nbr) — із таблицею транзакцій. Об'єднання основної таблиці з трьома допоміжними виконується операцією LEFT JOIN, що зберігає всі рядки основного набору. Об'єднання з таблицею свят є нетривіальним і потребує власної логіки, описаної нижче.

Очищення таблиці свят

Аналіз вмісту поля type показав, що 12 рядків мають transferred=True. Семантично ці рядки відповідають оригінальним датам перенесених свят: офіційно свято припадало на цю дату, проте практично воно було перенесене на інший день, а на оригінальну дату магазини працювали як зазвичай. Фактичне свято (з закриттям магазинів) фіксується окремим рядком зі значенням type=Transfer. Залишення обох рядків призвело б до подвійного маркування святкового статусу для двох різних дат, тому всі 12 рядків з transferred=True вилучено. Після очищення таблиця свят містить 338 записів.

Логіка зв'язку за масштабом свята

Поле locale визначає, на які магазини поширюється святковий статус: національні свята застосовуються до всіх 54 магазинів незалежно від географічного розташування; регіональні свята — лише до магазинів, штат яких збігається зі штатом, зазначеним у locale_name; локальні свята — лише до магазинів відповідного міста.

Оскільки декілька свят можуть припадати на одну й ту саму дату (31 дата у наборі містить два або більше свят, а максимально — чотири), єдиний бінарний прапорець is_holiday призвів би до втрати інформації про природу святкового статусу. Натомість для кожного спостереження (date, store_nbr) обчислюються три незалежні бінарні прапорці: holiday_national, holiday_regional та holiday_local, кожен з яких набуває значення 1, якщо на цю дату припадає свято відповідного масштабу. Така декомпозиція дозволяє моделі окремо навчитися величині ефекту для кожного типу свята — національні свята демонструють набагато більший вплив на продажі, ніж регіональні чи локальні.

Більш складна обробка свят — зокрема моделювання вікон впливу навколо святкових дат та data-driven редизайн ознак свят — є частиною конструювання ознак і розглядається у підрозділі 2.4.5.

Після об'єднання отримуємо таблицю розміром 3 008 016 рядків зі складеним ключем ідентифікації (`date`, `store_nbr`, `family`), цільовою змінною `sales`, комерційною ознакою `onpromotion`, статичними характеристиками магазину (`city`, `state`, `type`, `cluster`), ціною нафти `dcoilwtico`, минулим коваріатом `transactions` (може містити `NaN`) та трьома прапорцями свят за масштабом. Ця таблиця є відправною точкою для конструювання ознак у підрозділі 2.4.

2.4 Конструювання ознак

Об'єднана таблиця, отримана в результаті процедур, описаних у підрозділі 2.3.3, містить лише ті змінні, що походять безпосередньо із сирих джерел. Для побудови якісного прогнозу цього недостатньо: модель потребує явного представлення часової структури, історії продажів, святкових режимів та інших похідних сигналів, які у сирому вигляді відсутні. У цьому підрозділі описано процес конструювання ознак — формування набору атрибутів, на основі якого глобальна модель прогнозування буде навчатися та робити передбачення.

Конструювання ознак підпорядковане двом принципам, що випливають з природи задачі. Перший — принцип доступності на момент прогнозування. Для кожного передбачуваного дня моделі доступні лише ті значення, які можна обчислити, маючи інформацію станом на дату прогнозу. Цей принцип безпосередньо обмежує лагові та ковзні ознаки: оскільки горизонт прогнозування становить 16 днів, мінімальний дозволений лаг будь-якої ознаки, побудованої з історії цільової змінної або з минулих коваріатів, дорівнює 16 дням. Менші лаги формально містили б витік даних — модель «знала» б майбутні значення, недоступні у виробничому середовищі.

Другий принцип — узгодженість обчислення ознак між навчальним і те-

стовим періодами. Лагові ознаки та ковзні статистики мають обчислюватися на безперервній часовій сітці без розривів між кінцем навчального та початком тестового вікон. Для цього навчальну й тестову таблиці об'єднано в єдиний кадр, на якому обчислено всі похідні ознаки, а розділення рядків за маркером `is_test` виконано вже після завершення обчислень.

Сконструйований набір ознак поділяється на сім груп за природою сигналу: календарні ознаки (підрозділ 2.4.1), статичні категоріальні ознаки магазинів (2.4.2), ознаки історії продажів (2.4.3), ознаки промоакцій (2.4.4), ознаки свят і структурних подій (2.4.5), ознаки ціни нафти (2.4.6) та ознаки транзакцій (2.4.7).

2.4.1 Календарні ознаки

Календарні ознаки — це функції, що залежать виключно від дати спостереження і не потребують жодних інших джерел даних. Їх перевага полягає в тому, що вони доступні для будь-якої майбутньої дати без потреби в прогнозуванні, а тому не порушують принцип доступності, сформульований вище. У роботі сконструйовано вісім таких ознак.

День тижня `dow` (значення від 0 = понеділок до 6 = неділя) представляє тижневу сезонність, яку розвідувальний аналіз визначив як найбільш виражений сезонний компонент. День місяця `day` (1–31) кодує локальний паттерн всередині місячного циклу, зокрема, дозволяє моделі виявити ефект днів виплати заробітної плати. Місяць `month` (1–12) виражає річну сезонність із грудневим піком. Календарний рік `year` та номер тижня року `week_of_year` (стандарт ISO 8601) надають додаткові часові маркери для розрізнення схожих за іншими атрибутами періодів. Бінарна ознака `is_weekend` приймає значення 1 для суботи та неділі і 0 для решти днів; вона дублює інформацію з `dow`, але виділяє вихідні явно, що корисно для алгоритмів дерев рішень.

Окреме значення має ознака `t_idx` — лінійний індекс часу, обчислений як кількість днів від першої дати у наборі. Деревні алгоритми за своєю природою не здатні екстраполювати значення цільової змінної за межі діапазону

навчальних даних: розщеплення дерева оперує лише тими порогами, що спостерігалися у тренувальному наборі. Введення ознаки `t_idx` частково компенсує цю властивість, дозволяючи моделі явно враховувати глобальну часову позицію спостереження та формувати окремі гілки для пізніших періодів.

Восьма ознака — `days_to_payday` — кодує мінімальну відстань у днях до найближчої дати виплати заробітної плати у державному секторі Еквадору (15 число або останній день місяця). Обчислення виконується як мінімум серед чотирьох кандидатів: 15 число поточного місяця, останній день поточного місяця, 15 число наступного місяця та останній день попереднього місяця. Така формулювання дає симетричний сигнал як перед, так і після дня виплати, що відповідає емпіричному паттерну: підвищення продажів триває кілька днів навколо дати виплати. Підсумок усіх восьми календарних ознак наведено у таблиці 2.7.

Таблиця 2.7 – Сконструйовані календарні ознаки

Ознака	*Тип*	*Опис*
`dow`	ціле число	День тижня (0 — понеділок)
`day`	ціле число	День місяця (1–31)
`month`	ціле число	Місяць (1–12)
`year`	ціле число	Календарний рік
`week_of_year`	ціле число	Номер тижня року за ISO 8601
`is_weekend`	бінарна	1 для суботи та неділі, 0 в інші дні
`t_idx`	ціле число	Кількість днів від першої дати набору
`days_to_payday`	ціле число	Відстань у днях до найближчої дати виплати

Усі вісім ознак є детермінованими функціями дати, тому повністю обчислені як для навчального, так і для тестового періодів без потреби в імпутації.

2.4.2 Статичні категоріальні ознаки магазинів

Статичні ознаки магазинів — це характеристики, що не змінюються в часі: вони описують магазин як суб’єкт, а не його стан у конкретний день. У на-

борі даних такими є місто розташування (`city`), штат (`state`), тип магазину (`type`, значення A–E) та кластер (`cluster`, значення 1–17). Усі чотири змінні приєднано до основної таблиці на етапі передобробки (підрозділ 2.3.3) шляхом об'єднання за полем `store_nbr`.

Окрім перерахованих, до групи статичних ознак включено два додаткові поля: ідентифікатор товарної категорії `family` (33 значення) та сам ідентифікатор магазину `store_nbr` (54 значення). Хоча формально `store_nbr` є числовою змінною, його значення є довільними мітками, а не вимірюваннями; ту саму властивість має й `cluster`. Тому обидві ці змінні явно перетворено на категоріальний тип. Загалом до групи статичних категоріальних ознак віднесено шість змінних, перелічених у таблиці 2.8.

Таблиця 2.8 – Статичні категоріальні ознаки

Ознака	*К-сть значень*	*Опис*
`family`	33	Товарна категорія
`store_nbr`	54	Ідентифікатор магазину
`city`	22	Місто розташування магазину
`state`	16	Штат розташування магазину
`type`	5	Тип магазину (A–E)
`cluster`	17	Кластер магазинів за подібністю

Перетворення цих ознак на категоріальний тип має принципове значення для роботи з LightGBM. Алгоритм має вбудовану підтримку категоріальних змінних [13]: при пошуку розбиття для категоріального стовпця він розділяє множину значень на дві групи у спосіб, що максимізує зменшення функції втрат, минаючи стандартні впорядковані порогові розщеплення. Це усуває потребу у one-hot кодуванні, яке для змінних з великою кількістю унікальних значень суттєво розширило б ознаковий простір та сповільнило навчання.

З аналітичної точки зору, статичні ознаки відіграють роль умовних змінних (conditioning variables) у глобальній моделі. Глобальна модель навчається одночасно на всіх 1 782 часових рядах, об'єднаних в єдину навчальну вибірку, і має певним чином розрізняти ряди при формуванні прогнозу. Саме категоріальні ознаки виконують цю функцію: розщеплення за `family` спрямовує спосте-

реження GROCERY I та BOOKS у різні гілки дерева, де ефекти інших ознак — зокрема, лагових — можуть мати різний масштаб та форму. Аналогічну роль для географічного розрізнення відіграють `city`, `state` та `cluster`: магазини в одному кластері навчаються спільно як близькі за поведінкою, але можуть бути відокремлені від магазинів інших кластерів.

2.4.3 Ознаки історії продажів

Ознаки історії продажів — лагові значення цільової змінної та ковзні статистики, обчислені за історичними спостереженнями — у задачах прогнозування часто є найбільш інформативними. Розвідувальний аналіз показав сильну тижневу та річну сезонність продажів, а також виражений тренд; усі ці закономірності найкраще кодуються через історичні значення самого ряду.

При конструюванні ознак історії продажів дотримано двох ключових проєктних рішень. Перше — мінімальний лаг 16 днів, що дорівнює горизонту прогнозування. Для будь-якої дати прогнозу глибиною від 1 до 16 днів значення продажів 16-денної давнини є найсвіжішою інформацією, гарантовано доступною на момент прогнозування. Використання менших лагів призвело б до витоку даних: при прогнозуванні 16-го дня горизонту значення, отримане 15 днів тому, припадало б на 1-й день горизонту й було б недоступним.

Друге проєктне рішення — обчислення ознак на логарифмічно перетвореному цільовому стовпці $\ln(1 + \texttt{sales})$, а не на оригінальній шкалі. Підстава для цього вибору лежить у гетерогенності масштабу продажів між товарними категоріями: середньодобові продажі коливаються від 0,07 одиниць (BOOKS) до близько 3 800 одиниць (GROCERY I) — різниця у п’ять порядків. У логарифмічному просторі ця гетерогенність стискається, а пороги розщеплення, які модель навчиться застосовувати до лагових ознак, узагальнюються між рядами з різним базовим рівнем продажів.

Сконструйовано п’ять ознак історії продажів:

- `sales_lag16` — значення $\ln(1 + \texttt{sales})$ за 16 днів до дати спостереже-

ння. Це найсвіжіша гарантовано доступна точка історії;

- `sales_lag28` — значення $\ln(1 + \texttt{sales})$ за 28 днів до дати спостереження. Зсув кратний семи дням зберігає фазу тижневої сезонності;
- `sales_rmean7_lag16` — ковзне середнє $\ln(1+\texttt{sales})$ за вікном 7 днів зі зсувом 16 днів, що кодує короткостроковий рівень продажів;
- `sales_rmean28_lag16` — ковзне середнє за вікном 28 днів зі зсувом 16 днів, що кодує середньостроковий рівень;
- `sales_rmean56_lag16` — ковзне середнє за вікном 56 днів зі зсувом 16 днів, що кодує довгостроковий рівень.

Усі п'ять ознак обчислено окремо для кожної комбінації (`store_nbr`, `family`) — тобто строго в межах одного часового ряду — через групування за цими полями. За відсутності групування ковзне середнє могло б «перекинутися» між різними рядами, що порушило б семантику ознаки.

Лагові ознаки за своєю природою мають пропущені значення на початку кожного ряду — для перших 16, 28 або 56 днів кожної пари (`store_nbr`, `family`) історичних спостережень потрібного діапазону ще не існує. Ці NaN зберігаються у навчальній вибірці; LightGBM коректно опрацьовує їх через вбудований механізм роботи з відсутніми значеннями. У тестовій частині набору всі ознаки історії доступні без пропусків.

2.4.4 Ознаки промоакцій

Поле `onpromotion`, описане в підрозділі 2.1, кодує кількість унікальних товарних позицій (SKU) відповідної товарної категорії, що перебували на промоакції в магазині в день спостереження. На відміну від ознак історії продажів, ця змінна доступна також для майбутніх дат — набір даних містить її і для тестового періоду — оскільки рішення щодо промоакцій приймаються заздалегідь.

За класифікацією коваріатів з підрозділу 1.3, `onpromotion` є майбутнім відомим коваріатом (known future covariate). На відміну від минулих коваріатів (як-от транзакції), вона не потребує лагування або агрегації для зведення до доступної на момент прогнозування форми, і тому використовується безпосередньо у сирому вигляді.

Розвідувальний аналіз показав, що промоакції мають суттєвий і диференційований за товарними категоріями вплив на продажі: ефект на категорію CLEANING складає $+43\%$, тоді як для PRODUCE — $+208\%$. Цей сильний і нелінійний за категоріями сигнал є важливою причиною для збереження ознаки у моделі без додаткових перетворень. Алгоритм LightGBM здатний автоматично виявити взаємодію між `onpromotion` та `family`, формуючи розщеплення, що залежать від обох ознак одночасно.

2.4.5 Ознаки свят та структурних подій

Свята є неоднозначним за впливом сигналом: одні підвищують продажі (передсвятковий шопінг), інші — знижують (закриття магазинів у саме свято), а ефект свят регіонального або локального масштабу проявляється лише для частини магазинів. Це робить кодування святкової інформації нетривіальним і вимагає окремої уваги.

У базовому наборі ознак, побудованому на етапі передобробки, вже сформовано три бінарні прапорці: `is_national_holiday`, `is_regional_holiday`, `is_local_holiday`. Кожен прапорець набуває значення 1, якщо для конкретного спостереження (`date`, `store_nbr`) відповідне свято має ефект, і 0 в іншому випадку. Логіка зв'язку між магазином і регіональним або локальним святом враховує штат і місто магазину, як описано у підрозділі 2.3.3. Ці три прапорці представляють грубе (coarse) кодування святкового статусу: вони відрізняють святковий день від звичайного, але не розрізняють різні свята між собою.

До трьох грубих прапорців додано одну ознаку структурної події —

`hol_earthquake_window`. Це бінарний прапорець, що набуває значення 1 для всіх дат з 16 квітня по 16 травня 2016 року (31-денне вікно) і 0 в інші дні. Введення цієї ознаки мотивоване результатами розвідувального аналізу: після землетрусу 16 квітня 2016 року продажі товарів першої необхідності зросли на 19–34% і поверталися до базового рівня протягом приблизно чотирьох тижнів. Цей ефект не може бути закодований через регулярні святкові ознаки, оскільки землетрус є унікальною одноразовою подією.

Описаний у цьому пункті набір святкових ознак — три грубі прапорці плюс прапорець структурної події — є початковим. Він має кілька свідомо прийнятих обмежень. По-перше, грубі прапорці не розрізняють окремі свята між собою: різдвяний день, день незалежності Гуаякіля та локальне свято Кіто отримують однакове кодування, хоча їх ефект на продажі може суттєво відрізнятися. По-друге, у поточній формі вони не моделюють передсвятковий шопінговий буфер — передріздвяні дні, які за даними розвідувального аналізу демонструють найвищі продажі у мережі, кодуються так само, як звичайний понеділок. По-третє, ефект свята не масштабується ані за його категорією, ані за тривалістю. Чи виявляться ці обмеження істотними для якості прогнозу, та чи виправдане ускладнення кодування, — питання, на яке відповідатиме емпіричний експеримент у розділі 3.

2.4.6 Ознаки ціни нафти

Денна ціна нафти WTI `dcoilwtico`, отримана після імпутації пропусків у підрозділі 2.3.1, є потенційним макроекономічним предиктором, обґрунтованим залежністю еквадорської економіки від нафтових доходів (підрозділ 2.1). Денна ціна, однак, має високу часову мінливість, тому використання її у сирому вигляді як предиктора з горизонтом 16 днів недоцільне. Замість цього сконструйовано згладжене представлення — `oil_rmean30`, 30-денне ковзне середнє ціни WTI на безперервному денному календарі. Згладжене представлення зберігає трендову складову нафтової динаміки та знижує чутливість моделі до

денних викидів біржових сесій. Ознака обчислюється спільно для всіх магазинів, оскільки ціна нафти є глобальним фактором, не залежним від географічного розташування. Для перших днів навчального періоду, де 30-денне вікно ще не сформоване, значення ознаки залишається `NaN`.

2.4.7 Ознаки транзакцій

Кількість транзакцій `transactions`, отримана після переіндексації таблиці у підрозділі 2.3.1, є минулим коваріатом за класифікацією з підрозділу 1.3: її значення відомі лише за історичні періоди, тому пряме використання вимагає лагування на горизонт прогнозування. Сконструйовано дві ознаки:

- `transactions_lag16` — значення кількості транзакцій за 16 днів до дати спостереження; найсвіжіше гарантовано доступне значення для горизонту прогнозування;
- `transactions_rmean28` — 28-денне ковзне середнє ознаки `transactions_lag16`, що пом'якшує денний шум та підвищує стабільність сигналу.

Обидві ознаки обчислено окремо для кожного магазину (групування за `store_nbr`). Пропущені значення `NaN`, що виникають у початкові дні (до того, як магазин відкрився, або у дні до того, як набралось вікно для ковзного середнього), залишаються незаповненими — LightGBM коректно опрацьовує їх через вбудований механізм роботи з відсутніми значеннями.

2.5 Стратегія валідації для багатовимірних часових рядів

Як показано у підрозділі 1.9, оцінка прогнозних моделей для часових рядів вимагає валідаційних схем, що зберігають хронологічний порядок спосте-

режень. Стандартна k-fold крос-валідація з випадковим перемішуванням рядків призводить до витоку інформації з майбутнього у навчальну вибірку та неприпустима. Для нашої задачі схема валідації має додатково враховувати дві специфічні особливості: фіксований 16-денний горизонт прогнозування та багатовимірну структуру даних — одночасну присутність 1 782 часових рядів у єдиному наборі.

Реалізована стратегія — розширювальне вікно (expanding window) з трьома непересічними валідаційними фолдами по 16 днів кожен, розташованими безпосередньо перед цільовим прогнозним вікном (2017-08-16 — 2017-08-31). На кожному фолді модель навчається на всіх історичних спостереженнях до межі дня початку валідаційного вікна та оцінюється на спостереженнях усередині цього вікна. Визначення фолдів наведено у таблиці 2.9.

Таблиця 2.9 – Конфігурація фолдів крос-валідації

Фолд	*Кінець навчального вікна*	*Валідаційне вікно*
1	2017-06-28	2017-06-29 — 2017-07-14
2	2017-07-14	2017-07-15 — 2017-07-30
3	2017-07-30	2017-07-31 — 2017-08-15

Конфігурація сформована на основі декількох проєктних рішень. Кожне валідаційне вікно охоплює рівно 16 днів — стільки ж, скільки тестове. Це гарантує, що крос-валідаційна оцінка моделює точно той самий статистичний режим, який буде у фактичному прогнозуванні: модель робить ті самі 16 прогнозів на той самий горизонт. Будь-яке скорочення або подовження валідаційного вікна змінило б розподіл глибин прогнозу у вибірці та призвело б до зміщеної оцінки.

Вибір розширювального, а не ковзного вікна, зумовлений характером нестаціонарності у наборі. Альтернативою є sliding window з фіксованою довжиною навчальної вибірки. Перевага розширювального вікна — використання максимально можливої кількості історичних даних на кожному фолді; перевага ковзного — стійкість до нестаціонарності та зменшення впливу старих даних. Основними режимами нестаціонарності у нашому наборі є загальний висхідний тренд та структурний злом, спричинений землетрусом 16 квітня 2016 року. Обидва ці явища представлені експліцитно через ознаки `t_idx` та

`hol_earthquake_window` (підрозділ 2.4), що зменшує мотивацію відкидати старі дані. Тому збереження повної історії, прийняте розширювальним вікном, є більш виправданим.

Окремої уваги заслуговує розташування валідаційних вікон безпосередньо перед тестовим. Три валідаційні вікна охоплюють період з 29 червня по 15 серпня 2017 року. Тестове вікно припадає на 16–31 серпня того ж року. Усі чотири періоди розташовані в одному сезоні (середина-кінець літа 2017) та в одному макроекономічному режимі, що мінімізує сезонний зсув між крос-валідаційною оцінкою та фактичною якістю на тестовому вікні. Альтернативний підхід — рівномірний розподіл валідаційних фолдів за всіма роками — усереднював би оцінку за зимовими, весняними та літніми вікнами, що мало пов'язано з якістю на конкретному (літньому) тестовому вікні.

Валідаційні вікна не перекриваються і не мають проміжків між собою: фолд 1 закінчується 14 липня, фолд 2 починається 15 липня. Жодне спостереження не входить до двох валідаційних вибірок, що забезпечує статистичну незалежність помилок між фолдами та коректність усереднення RMSLE.

Кожне валідаційне вікно містить рівно 28 512 рядків — 16 днів $\times$ 1 782 серії. Метрика RMSLE обчислюється агреговано за всіма цими рядками, а не окремо для кожної серії з подальшим усередненням. Така схема узгоджена з фінальним оцінюванням моделі: на цільовому прогнозному вікні RMSLE також обчислюється агреговано за всіма позиціями одночасно. Вона також природна для глобальної моделі: оскільки одна модель прогнозує всі 1 782 серії, доцільно оцінювати її якість сумарно.

Звітною крос-валідаційною оцінкою слугує середнє значення RMSLE по трьох фолдах, що супроводжується пофолдовою розкладкою. Розкид між фолдами характеризує стабільність моделі: за наявності помітної різниці оцінка є шумною, і відмінності між моделями, менші за цей розкид, не варто інтерпретувати як значущі. Для діагностичних процедур, що стосуються внутрішньої структури моделей, як референсний обрано фолд 3 (валідаційне вікно 31 липня — 15 серпня 2017 року) — найближчий за датою до цільового прогнозного вікна; триразова валідація для подібних процедур додаткової точності не дала б.

Метрикою якості у всіх експериментах слугує RMSLE (1.25). Як показано у підрозділі 2.3.2, оптимізація стандартної функції втрат MSE на логарифмічно перетвореній цільовій змінній є точно еквівалентною оптимізації RMSLE на оригінальних значеннях, що знімає потребу у реалізації власної функції втрат та гарантує точну відповідність між тим, що оптимізує модель, і тим, як вона оцінюється.

Висновки до розділу 2

У другому розділі здійснено аналіз сирого набору даних, виконано його передобробку та сформовано робочий ознаковий простір, на якому будуть будуватися прогнозні моделі.

Описано набір даних мережі продовольчих магазинів Еквадору [21]: 54 магазини, 33 товарні категорії, 1 782 часові ряди продажів за чотири з половиною роки. Обґрунтовано доцільність вибору саме цього набору з огляду на багатий склад екзогенних факторів різної природи, наявність метаданих магазинів як статичних коваріатів та присутність типових для роздрібної торгівлі особливостей: множинної сезонності, переривчастих рядів, структурних зломів.

Проведено розвідувальний аналіз даних, який виявив низку ключових властивостей. Цільова змінна має сильно правоскошений розподіл (медіана $= 11$, середнє $= 358$, асиметрія $= 7{,}4$) із суттєвою часткою нульових спостережень (31,3%). Виявлено множинну сезонність — тижневу, річну та внутрішньомісячну, — виражений висхідний тренд (зростання середніх продажів удвічі за період), сильний вплив промоакцій ($+43\%$ до $+208\%$ залежно від категорії), помірний вплив національних свят та спурійний характер кореляції з ціною нафти (зникає після видалення тренду). Окремою аномалією є землетрус 16 квітня 2016 року, що спричинив 19–34%-вий приріст продажів товарів першої необхідності протягом приблизно місяця.

Розроблено стратегію передобробки даних з окремими підходами до трьох категорій пропусків: пропущені дати Різдва заповнено нульовими прода-

жами; пропуски у ціні нафти усунено прямою імпутацією; пропуски в таблиці транзакцій збережено як NaN з огляду на здатність LightGBM коректно опрацьовувати відсутні значення. Обґрунтовано та застосовано логарифмічне перетворення цільової змінної; показано, що при цьому стандартна функція втрат MSE на перетворених значеннях є точно еквівалентною оптимізації RMSLE на оригінальних, що знімає потребу у власній реалізації функції втрат. Виконано об'єднання п'яти таблиць у єдиний навчальний набір розміром 3 008 016 рядків.

Сформовано початковий ознаковий простір у сім груп за природою сигналу: вісім календарних ознак, шість статичних категоріальних ознак магазинів, п'ять ознак історії продажів, одну ознаку промоакцій, чотири ознаки свят і структурної події, одну ознаку ціни нафти та дві ознаки транзакцій (разом 27 ознак). Ключові проектні рішення — мінімальний лаг 16 днів для ознак, побудованих на історії цільової змінної та минулих коваріатах, обчислення лагових ознак історії продажів у логарифмічному просторі, ізоляція обчислень у межах однієї пари (магазин, категорія) — обґрунтовано з боку відсутності витоку даних та узагальнюваності між рядами різного масштабу.

Запропоновано стратегію валідації, що відповідає характеру задачі: розширювальне вікно з трьома непересічними валідаційними фолдами по 16 днів кожен, розташованими безпосередньо перед цільовим прогнозним вікном. Така конфігурація узгоджує крос-валідаційну оцінку з реальним статистичним режимом прогнозування (та сама довжина горизонту, той самий сезон) та забезпечує статистичну незалежність помилок між фолдами. Метрикою якості обрано RMSLE.

Сформований у цьому розділі набір даних та ознаковий простір є відправною точкою для експериментальної частини дослідження.

РОЗДІЛ 3
ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ

3.1 Базова модель: лінійна регресія на екзогенних факторах

Перед побудовою складних моделей, що використовують усю інформацію зі сформованого у розділі 2 ознакового простору, доцільно встановити нижню межу якості — значення метрики, яке досягається моделлю мінімальної складності, що не використовує історію цільової змінної. Такий орієнтир виконує дві функції: характеризує величину сигналу, що міститься у самих лише екзогенних факторах, та задає поріг, який повинна перевершити будь-яка складніша модель, щоб виправдати своє ускладнення.

Базовою моделлю обрано Ridge-регресію, описану у підрозділі 1.5.2. Цей вибір обґрунтований трьома міркуваннями. Лінійна структура моделі робить базову оцінку інтерпретованою через ваги ознак та забезпечує її обчислювальну ефективність; L_2-регуляризація стабілізує оцінки коефіцієнтів за умови наявності корельованих ознак; функція втрат MSE, яку мінімізує Ridge, на логарифмічно перетвореній цільовій змінній є точно еквівалентною RMSLE (підрозділ 2.3.2), що знімає невідповідність між цілями оптимізації моделі та оцінювання.

Ознаковий простір базової моделі формують лише екзогенні фактори: вісім календарних ознак, шість статичних категоріальних ознак магазинів, бінарна змінна `onpromotion`, три грубі прапорці свят та згладжена ціна нафти `oil_rmean30` — разом 18 ознак. Свідомо виключено всі лагові ознаки та ковзні статистики цільової змінної: модель має характеризувати прогностичний потенціал самих лише екзогенних факторів, без участі історії цільового ряду.

Передобробка виконана відповідно до вимог Ridge-регресії. Категоріальні ознаки (`family`, `city`, `state`, `type`, `cluster`) перетворено через one-hot кодування у розріджену матрицю; числові ознаки стандартизовано за статистиками навчальної частини кожного фолду окремо. Пропущені значення

`oil_rmean30` у перших днях навчального періоду заповнено середнім значенням ознаки за навчальною частиною. Цільову змінну перетворено за формулою $y' = \ln(1 + y)$.

Параметр регуляризації α обрано шляхом перебору серед значень $\{1, 10, 100\}$. Результати наведено у таблиці 3.1.

Таблиця 3.1 – Підбір параметра регуляризації Ridge

α	*CV-RMSLE*	*CV-MAE*	*CV-RMSE*
1	1,28 (нестабільно)	2 675	251 779
10	1,08	481	2 095
100	1,030	378	1 630

При $\alpha = 1$ модель є практично нерегуляризованою і дає неприпустимо нестабільні прогнози, що відображається у CV-RMSE на рівні $2{,}5 \cdot 10^5$. Збільшення α до 10 пригнічує цю нестабільність, але регресія все ще генерує локальні викиди. При $\alpha = 100$ оцінка стабілізується: пофолдове RMSLE становить $1{,}01/1{,}03/1{,}04$ — тісне групування, що свідчить про узгодженість поведінки моделі між валідаційними вікнами. Це значення зафіксовано як остаточне для базової моделі.

Конвеєр прогнозування реалізовано таким чином. Модель навчається на матриці X_{train} та логарифмічно перетворених цільових значеннях, мінімізуючи функцію втрат (1.11). На етапі прогнозу логарифмічні передбачення обрізаються до діапазону навчальної частини, потім застосовується обернене перетворення $\hat{y} = \exp(\hat{y}') - 1$ та обрізання знизу нулем. Кінцевий прогноз оцінюється за RMSLE на сирій шкалі продажів.

Базова модель досягає середнього CV-RMSLE $= 1{,}030$. Прогноз тестового вікна, отриманий шляхом повторного навчання моделі на повному навчальному наборі, дає оцінку $1{,}042$ — різниця з крос-валідаційною оцінкою становить $+0{,}012$, що підтверджує адекватність обраної валідаційної процедури.

Значення CV-RMSLE 1,030 показує, що з одних лише екзогенних факторів вдається стиснути логарифмічну помилку прогнозу до значення близько одиниці — календарна структура, статичні характеристики магазинів та інформація про промоакції містять реальний прогностичний сигнал. Водночас цей сигнал

є далеко не вичерпним: середня абсолютна помилка прогнозу базової моделі складає 378 одиниць продажів. Це значення є нижньою межею, з якою порівнюватиметься якість моделей, що використовують історію продажів.

3.2 Глобальна модель градієнтного бустингу

Установивши базовий орієнтир CV-RMSLE $=$ $1{,}030$, переходимо до основного класу моделей цього дослідження — глобального градієнтного бустингу на повному ознаковому просторі, сформованому в розділі 2.

3.2.1 Архітектура та конфігурація моделі LightGBM

Як основну архітектуру обрано LightGBM — реалізацію градієнтного бустингу на деревах рішень, описану в підрозділі 1.5.5. Вибір цієї бібліотеки обґрунтований кількома алгоритмічними властивостями.

По-перше, LightGBM має вбудовану підтримку категоріальних змінних [13]. Для шести статичних категоріальних ознак нашого набору це знімає потребу у one-hot кодуванні: алгоритм самостійно розбиває множину значень категоріальної ознаки на дві групи у спосіб, що максимізує зменшення функції втрат.

По-друге, алгоритм коректно опрацьовує пропущені значення без потреби у попередній імпутації [13]. Це особливо важливо для лагових ознак історії продажів, які за визначенням мають NaN у перших 16–56 днях кожного ряду, та для ознак транзакцій, де NaN кодує інформативну подію (закриття магазину або період до його відкриття).

По-третє, LightGBM застосовує leaf-wise стратегію зростання дерев та техніки GOSS і EFB, що забезпечують високу швидкість навчання на великих наборах даних. Для нашого набору з понад 3 млн рядків ця ефективність дозволяє

виконувати десятки повторних навчань за прийнятний обчислювальний час.

Гіперпараметри моделі обрано на основі загальноприйнятих рекомендацій для табличних задач середнього розміру [13] та зафіксовано на одних і тих самих значеннях для всіх експериментів цього розділу. Така фіксація забезпечує контрольованість порівнянь між моделями: будь-яка спостережувана у подальших порівняннях різниця у якості пов'язана з варіаціями ознакового простору. Конкретні значення наведено у таблиці 3.2.

Таблиця 3.2 – Гіперпараметри LightGBM, зафіксовані для всіх експериментів

Параметр	*Значення*	*Коментар*
`objective`	regression	MSE на $\ln(1+y)$
`metric`	rmse	RMSE на $\ln(1+y)$ = RMSLE на y
`learning_rate`	0,05	Темп навчання
`num_leaves`	63	Максимум листків у дереві
`feature_fraction`	0,8	Частка ознак для кожного дерева
`bagging_fraction`	0,8	Частка рядків для кожного дерева
`bagging_freq`	1	Підвибірку оновлюють щораз
`min_data_in_leaf`	100	Мін. кількість спостережень у листку
`seed`	42	Для відтворюваності результатів

Темп навчання 0,05 забезпечує помірне накопичення сигналу. Кількість листків 63 визначає максимальну глибину інтерактивних залежностей та відповідає балансу між гнучкістю моделі і її схильністю до перенавчання. Параметри `feature_fraction` і `bagging_fraction` разом із `bagging_freq` $= 1$ задають випадкове підвибирання 80% ознак та 80% рядків при побудові кожного дерева, що зменшує кореляцію між послідовно побудованими деревами та сприяє узагальнювальній здатності. Мінімальна кількість спостережень у листку 100 запобігає утворенню листків з низькою статистичною надійністю.

3.2.2 Опис набору ознак

Глобальна модель LightGBM використовує повний ознаковий простір, сформований у розділі 2. До нього входять 27 ознак, поділених на сім груп; склад наведено у таблиці 3.3.

Таблиця 3.3 – Склад ознакового простору глобальної моделі LightGBM

Група	*К-сть*	*Ознаки*
Календарні	8	`dow`, `day`, `month`, `year`, `week_of_year`, `is_weekend`, `t_idx`, `days_to_payday`
Статичні категоріальні	6	`family`, `store_nbr`, `city`, `state`, `type`, `cluster`
Промоакції	1	`onpromotion`
Свята та структурна подія	4	`is_national_holiday`, `is_regional_holiday`, `is_local_holiday`, `hol_earthquake_window`
Ціна нафти	1	`oil_rmean30`
Транзакції	2	`transactions_lag16`, `transactions_rmean28`
Історія продажів	5	`sales_lag16`, `sales_lag28`, `sales_rmean7_lag16`, `sales_rmean28_lag16`, `sales_rmean56_lag16`

Принципова відмінність від базової моделі полягає у способі обробки категоріальних ознак: LightGBM передає їх у вигляді категоріального типу та застосовує native categorical splits, що було неможливим для Ridge-регресії без one-hot кодування.

Комерційну змінну `onpromotion` (підрозділ 2.4.4) включено без додаткових перетворень як майбутній відомий коваріат. Святкова група об'єднує чотири бінарні ознаки (підрозділ 2.4.5): три грубі прапорці локальних, регіональних та національних свят разом із прапорцем 31-денного вікна землетрусу як структурної події. Згладжена 30-денна ціна нафти `oil_rmean30` є зовнішнім макроекономічним сигналом, поданим у пом'якшеному вигляді через ковзне середнє.

Транзакції є минулим коваріатом — їх значення відомі лише за історичні періоди, тому використання вимагає лагування на горизонт прогнозування. Ознака `transactions_lag16` дає найсвіжіше доступне значення, `transactions_rmean28` — 28-денне ковзне середнє для додаткової стабільності проти денного шуму. Ознаки історії продажів (підрозділ 2.4.3) складаються з двох лагових значень та трьох ковзних середніх з лагом 16, обчислених у логарифмічному просторі $\ln(1 + \texttt{sales})$.

3.2.3 Процедура навчання та результати на крос-валідації

На вхід LightGBM подаються матриця ознак X_{train} та цільові значення $y'_{\text{train}} = \ln(1 + y_{\text{train}})$. Модель навчається з функцією втрат MSE, що відповідно до (2.3) є еквівалентним оптимізації RMSLE на оригінальній шкалі. Кількість дерев визначається через ранню зупинку: початкове обмеження `num_boost_round` $= 3000$ задає максимально допустиму глибину навчання; реальна кількість дерев відповідає тій ітерації, на якій якість на валідаційному фолді перестала поліпшуватися протягом 100 послідовних раундів.

На етапі прогнозу логарифмічні передбачення перетворюються в одиниці сирої шкали через $\hat{y} = \exp(\hat{y}') - 1$ та обрізаються знизу нулем. На відміну від базової Ridge, обрізання логарифмічних прогнозів зверху не застосовується: LightGBM як алгоритм дерев рішень за визначенням не виходить за межі значень, спостережених у навчанні.

Результати трьох фолдів крос-валідації наведено у таблиці 3.4.

Таблиця 3.4 – Результати глобальної моделі LightGBM на крос-валідації

Фолд	*best_iter*	*RMSLE*
1	2 304	0,374
2	3 000	0,389
3	709	0,404
Середнє	—	**0,389**

Середнє CV-RMSLE становить 0,389, з відповідними значеннями MAE $= 63$ та RMSE $= 229$ в одиницях сирої шкали. Порівняно з базовою моделлю (1,030) це означає 62%-ве зниження логарифмічної помилки прогнозу. Така величина різниці свідчить про те, що дві ключові архітектурні відмінності — наявність ознак історії продажів та нелінійна функціональна форма дерев рішень — спільно вносять основну частку прогностичного сигналу.

Пофолдова розкладка демонструє помірну варіативність: RMSLE окремих фолдів коливається у діапазоні від 0,374 до 0,404, розкид становить 0,030. Ця величина задає шумовий поріг крос-валідаційної оцінки — відмінності між

моделями у межах кількох тисячних RMSLE не варто інтерпретувати як значущі без додаткової перевірки. Кількість ітерацій до спрацювання ранньої зупинки варіює між фолдами (2 304, 3 000 без спрацювання, 709), що відображає різну глибину сигналу, доступного моделі у різних валідаційних вікнах.

3.3 Декомпозиція внеску екзогенних факторів

Глобальна модель LightGBM демонструє CV-RMSLE $= 0{,}388$ — результат, що утворився від спільного внеску усіх семи груп ознак. Залишається відкритим питання, яка з цих груп критично необхідна, а яка може бути видалена без втрати якості. У цьому підрозділі це питання вирішується у три кроки: абляційний аналіз груп ознак (3.3.1), data-driven редизайн ознак свят (3.3.2) та фіксація підсумкового ознакового простору (3.3.3).

3.3.1 Абляційний аналіз груп ознак

Абляційний аналіз ґрунтується на простій ідеї: для оцінки внеску групи ознак її вилучають із повного набору, перенавчають модель і порівнюють отриману якість з референсною. У цьому пункті аналізуються чотири групи екзогенних коваріатів: промоакції, грубі прапорці свят, ціна нафти та транзакції. Календарні, статичні категоріальні ознаки, ознаки історії продажів та прапорець вікна землетрусу зберігаються у всіх варіантах.

Експериментальна процедура є точним повторенням навчання глобальної моделі з 3.2 для кожного абляційного варіанту. Єдиною зміною є збільшення верхньої межі кількості раундів градієнтного бустингу з 3 000 до 5 000 — мотивоване спостереженням з пункту 3.2.3 про різну глибину сигналу у різних валідаційних вікнах. Підвищена межа гарантує, що жоден з варіантів не буде штучно обмежений у кількості ітерацій.

Результати чотирьох абляційних експериментів наведено у таблиці 3.5.

Таблиця 3.5 – Результати абляційного аналізу груп ознак

Варіант	*CV-RMSLE*	Δ *RMSLE*	*Інтерпретація*
Повна модель (референс)	0,3884	—	—
Без промоакцій	0,4225	+0,0341	Поза шумовим порогом
Без грубих прапорців свят	0,3894	+0,0010	У межах шумового порогу
Без ціни нафти	0,3858	−0,0026	Невелика позитивна дельта
Без транзакцій	0,3895	+0,0011	У межах шумового порогу

Промоакції виявляються єдиним екзогенним сигналом, видалення якого помітно погіршує модель. Дельта +0,0341 помітно перевищує перевищує шумовий поріг 0,030 між фолдами і відповідає 8,8%-ному погіршенню моделі від одного-єдиного бінарного входу. Цей результат узгоджується з даними розвідувального аналізу (підрозділ 2.2), де ефект промоакцій становив від +43% до +208% продажів залежно від товарної категорії. Промо є категорично необхідним і безперечно зберігається у фінальній моделі.

Грубі прапорці свят і ознаки транзакцій демонструють зміни на рівні +0,001, що знаходиться глибоко в межах шумового порогу. Для транзакцій правдоподібним поясненням є інформаційна дублювальність: ознаки історії продажів вже несуть переважну частину сигналу за умови сильної кореляції з кількістю покупців ($r = 0{,}84$). Для грубих прапорців свят відсутність вимірюваного ефекту може свідчити або про реальну нерелевантність святкової інформації за наявності лагових ознак, або про те, що сам спосіб кодування є занадто грубим — альтернатива, що вирішується у пункті 3.3.2.

Найбільш несподіваним є результат для ціни нафти — невелика позитивна дельта при її вилученні. На початку дослідження (підрозділ 2.1) було підкреслено нафтозалежний характер економіки Еквадору [22], що становило теоретичне підґрунтя для очікування корисності цієї ознаки. Однак на етапі розвідувального аналізу було виявлено, що візуальна від'ємна кореляція між WTI та продажами повністю зникає після видалення тренду. Абляційний результат поглиблює цю інтерпретацію: навіть пом'якшене 30-денним ковзним середнім представлення нафти зберігає трендовий компонент, який модель може частково помилково

використовувати як прогностичний сигнал. Ця інформація ефективно дублюється індексом часу `t_idx` та календарними ознаками. Макроекономічна значущість нафтових цін залишається теоретично обґрунтованою, проте на горизонті 16 днів цей сигнал не виявляє самостійного позитивного емпіричного внеску.

За результатами цього пункту: промоакції зберігаються; ціна нафти та ознаки транзакцій вилучаються з фінального набору; питання щодо свят вирішується у наступному пункті.

3.3.2 Data-driven редизайн ознак свят

Нульовий внесок грубих прапорців свят допускає дві інтерпретації. Перша: інформація про святковий статус нерелевантна за наявності ознак історії продажів, оскільки ковзні середні 7- та 28-денних вікон уже частково охоплюють підвищення або зниження продажів навколо святкових дат. Друга: інформація релевантна, але загрубений спосіб кодування (єдиний бінарний прапорець для всіх свят відповідного масштабу) приховує суттєву різницю між окремими святами — наприклад, між підготовкою до Різдва (різке підвищення) та самим Різдвом (нульові продажі через закриття магазинів). У цьому пункті описано емпіричну процедуру розрізнення цих альтернатив через заміну трьох загальних прапорців окремим індивідуальним прапорцем для кожного значущого свята, відібраного безпосередньо з даних.

Вимірювання ефекту окремих свят

Для оцінки величини ефекту кожного окремого свята застосовано підхід «відхилення від базової лінії». Для кожного спостереження $(date, store, family)$, що припадає на день, маркований у таблиці свят, обчислюється логарифмічне відхилення:

$$d_i = \ln(1 + y_i) - b_i, \tag{3.1}$$

де y_i — фактичні продажі, а b_i — 28-денне центроване ковзне середнє $\ln(1 + y)$

для тієї ж пари $(store, family)$ та того ж дня тижня. Така конструкція базової лінії одночасно контролює два чинники: тижневу сезонність (порівнюються лише п'ятниці з п'ятницями) та індивідуальний масштаб ряду.

Подальший аналіз агрегує відхилення по всіх спостереженнях, що поділяють одне й те саме текстове позначення свята у вихідній таблиці. Для кожного унікального позначення обчислюється середнє значення $\overline{d}$ та кількість дат n, на які воно припадає. Відсортувавши позначення за $|\overline{d}|$, отримуємо ранжування свят за сприйнятим моделлю ефектом.

З отриманого ранжування формуються кандидати для включення у модель за двома критеріями (трек A): свято повторюється не менше трьох разів ($n \geq 3$) і абсолютне значення середнього відхилення перевищує 0,15 у логарифмічному просторі ($|\overline{d}| \geq 0{,}15$). Цим фільтрам задовольняють 12 позначень свят. Топ ранжування переважно становить різдвяне вікно (`navidad`, `navidad-1`, `navidad-2`, `navidad-3`, `primer_dia_del_ano`, `primer_dia_del_ano-1`) разом із кількома регіональними та локальними святами.

Окремо стоїть структурна подія — землетрус 16 квітня 2016 року, представлена прапорцем `hol_earthquake_window`. Цей прапорець методологічно належить до іншого треку (трек B): він описує одноразову подію, а не повторюване свято, тому критерії частоти та порогу абсолютного відхилення до нього непридатні. Прапорець уже включено до ознакового простору у пункті 2.4.5 і зберігається у всіх подальших експериментах.

CV-вибір розміру набору ознак свят

Маючи 12 ранжованих кандидатів треку A, тестуються три варіанти $N \in \{5, 8, 12\}$, для кожного з яких будується модель з топ-N ознаками. Базовий ознаковий простір формується з повного 27-ознакового набору шляхом вилучення трьох грубих прапорців свят, ціни нафти та ознак транзакцій (відповідно до рішень з 3.3.1). Залишається 21 ознака, до яких додається топ-N ознак треку A та постійний прапорець вікна землетрусу. Результати наведено у таблиці 3.6.

Варіант топ-5 показує погіршення відносно референсу: п'яти ознак недостатньо для покриття значущих свят. Варіант топ-12 повертається до якості повної моделі: додавання шостого, сьомого і подальших гранулярних свят понад

Таблиця 3.6 – Вибір розміру набору гранулярних ознак свят за крос-валідацією

Варіант	*К-сть ознак*	*CV-RMSLE*	Δ *vs. референс*
Повна модель (референс)	27	0,3884	—
Без грубих свят	24	0,3894	+0,0010
Топ-5 трек А + землетрус	27	0,3899	+0,0015
Топ-8 трек А + землетрус	**30**	**0,3867**	**−0,0017**
Топ-12 трек А + землетрус	34	0,3882	−0,0002

топ-5 додає ознаки, чий індивідуальний внесок є малим. Варіант топ-8 є оптимальним і єдиним, що демонструє вимірюване поліпшення відносно повної моделі — хоча й невелике ($-0{,}0017$), воно стабільно повторюється на всіх трьох фолдах. Це число $N = 8$ фіксується для фінального набору ознак.

3.3.3 Фінальний набір ознак та результати

З початкового 27-ознакового набору повної моделі видалено три грубі прапорці свят, ознаку ціни нафти та дві ознаки транзакцій (загалом шість ознак); натомість додано вісім ознак гранулярних свят треку А. Прапорець вікна землетрусу (трек В) зберігається. Загальна кількість ознак у фінальному наборі становить 29; склад наведено у таблиці 3.7.

Таблиця 3.7 – Склад фінального ознакового простору

Група	*К-сть*
Календарні ознаки	8
Статичні категоріальні ознаки магазинів	6
Промоакції	1
Гранулярні ознаки свят (топ-8 треку А)	8
Структурна подія (землетрус)	1
Ознаки історії продажів	5
Усього	**29**

Фінальна модель досягає CV-RMSLE $= 0{,}3867$ — стабільне поліпшення

відносно повної моделі на 0,0017, що відтворюється на всіх трьох фолдах. Це поліпшення отримане не за рахунок збільшення ємності моделі, а за рахунок цілеспрямованого перерозподілу ознакового простору. Підсумок еволюції моделей цього розділу наведено у таблиці 3.8.

Таблиця 3.8 – Еволюція якості моделі за рахунок зміни ознакового простору

Модель	*К-сть ознак*	*CV-RMSLE*
Базова Ridge (підрозділ 3.1)	18	1,030
Повна LightGBM (підрозділ 3.2)	27	0,388
Фінальна LightGBM (3.3)	29	0,387

3.4 Розширення системи: альтернативні архітектури та ансамблювання

Фінальна модель з пункту 3.3.3 досягає CV-RMSLE $= 0{,}387$. Подальше зниження помилки на одному й тому самому ознаковому просторі вимагає переходу до ансамблевих підходів. Теоретична основа ансамблювання (підрозділ 1.7.2) ґрунтується на принципі диверсифікації помилок: за умови, що окремі моделі помиляються по-різному, агрегований прогноз має меншу варіацію та краще узагальнюється. У цьому підрозділі досліджуються дві стратегії побудови різноманітних моделей — архітектурна (гібридна модель Ridge + LightGBM, 3.4.1) та ознакова (варіанти з модифікованим лаговим горизонтом, 3.4.2) — після чого емпірично перевіряється диверсифікованість помилок (3.4.3) та формується підсумковий ансамбль (3.4.4).

3.4.1 Гібридна модель Ridge + LightGBM

Гібридна модель поєднує лінійну та нелінійну компоненти у послідовній двоетапній архітектурі (підрозділ 1.7.1). На першому етапі Ridge-регресія прогнозує цільову змінну на основі підмножини детермінованих ознак: календарних, статичних категоріальних, промоакцій та святкових. На другому етапі LightGBM навчається не на самій цільовій змінній, а на залишках Ridge:

$$e_i = y'_i - \hat{y}'_{\text{Ridge},i}, \tag{3.2}$$

де $y'_i = \ln(1+y_i)$, а $\hat{y}'_{\text{Ridge},i}$ — лінійний прогноз. Фінальний прогноз обчислюється як сума двох компонент:

$$\hat{y}_i = \exp\bigl(\hat{y}'_{\text{Ridge},i} + \hat{e}_{\text{LGBM},i}\bigr) - 1. \tag{3.3}$$

Принципово важливим аспектом конфігурації є виключення ознак історії продажів з лінійної частини. Якби Ridge мав доступ до лагових ознак, він би поглинув переважну частину сигналу, і залишок e_i містив би лише шум. Передбачуваною є ситуація, коли Ridge захоплює детерміновану складову (тренд, сезонність, святкові режими), а LightGBM моделює нелінійні залежності.

Результати наведено у таблиці 3.9.

Таблиця 3.9 – Порівняння однорівневої та гібридної моделі на крос-валідації

Модель	*Фолд 1*	*Фолд 2*	*CV-RMSLE*
Фінальна LightGBM (3.3.3)	0,3717	0,3868	0,3867
Гібрид Ridge + LightGBM	—	—	0,3908

Гібридна архітектура, попри теоретичну обґрунтованість, не покращила якість прогнозу: CV-RMSLE $=$ $0{,}3908$ є гіршим від однорівневої моделі на 0,0041. Гібридний підхід є найбільш ефективним, коли цільовий ряд декомпозується на чітку глобальну детерміновану складову та невеликі нелінійні зали-

шки. Наш набір не має такої структури: основний прогностичний сигнал зосереджений у власній історії кожного ряду, а не у глобальному тренді. Прогноз Ridge на детермінованих ознаках має якість приблизно 1,03 RMSLE — залишок, який LightGBM повинен моделювати, є практично таким самим за масштабом, як і сама цільова змінна. LightGBM не отримує переваги від попередньо віднятої лінійної компоненти; навпаки, прогноз додатково навантажується помилкою Ridge.

Гібридна модель попри від'ємний індивідуальний результат зберігається як один із чотирьох кандидатів для ансамблю, оскільки її архітектурна відмінність може давати корисну для ансамблю диверсифікацію помилок.

3.4.2 Варіанти з модифікованим лаговим горизонтом

Альтернативний механізм диверсифікації — варіація лагового горизонту: побудова моделей, що використовують різні часові масштаби історії продажів. Моделі з різними лагами будуть по-різному реагувати на типові ситуації — зокрема, на тижні з нетиповою динамікою у близькому минулому або, навпаки, на місячно-квартальному горизонті.

Побудовано два варіанти. Короткий-лаговий варіант використовує лише ознаки історії продажів з вікнами до місяця: `sales_lag16`, `sales_rmean7_lag16` та `sales_rmean14_lag16` (три ознаки). Довгий-лаговий варіант використовує ознаки з більшими часовими вікнами: `sales_lag28`, `sales_lag56`, `sales_rmean28_lag16` та `sales_rmean91_lag16` (чотири ознаки). Для обох варіантів усі інші ознаки зберігаються незмінними; гіперпараметри LightGBM та конфігурація крос-валідації — ті ж, що й для фінальної моделі.

Довгий-лаговий варіант досягає якості, статистично невідмінної від референсної однорівневої моделі (різниця 0,0007 у межах шумового порогу). Короткий-лаговий поступається референсу на 0,0043. Обидва варіанти зберігаються як кандидати для ансамблю: їх роль не полягає у самостійному покра-

Таблиця 3.10 – Варіанти моделі з модифікованим лаговим горизонтом

Модель	*К-сть ознак історії*	*CV-RMSLE*
Фінальна LightGBM (3.3.3) — референс	5	0,3867
Короткий-лаговий варіант	3	0,3910
Довгий-лаговий варіант	4	0,3874

щенні якості, а в тому, що вони можуть давати помилки, частково ортогональні до помилок основної моделі.

3.4.3 Кореляційний аналіз помилок між моделями

Перш ніж об'єднувати чотири моделі в ансамбль, доцільно емпірично перевірити, наскільки їхні помилки справді різняться. Чотири моделі-кандидати (однорівнева з 3.3.3, гібрид із 3.4.1, короткий-лаговий та довгий-лаговий варіанти з 3.4.2) навчено на фолді 3 крос-валідації; для кожної моделі зібрано прогнози на валідаційному вікні фолду 3. Відповідно до домовленості з підрозділу 2.5, цей діагностичний експеримент виконано на одному фолді.

З прогнозів обчислено логарифмічні помилки $e_i^{(m)} = \ln(1 + y_i) - \ln(1 + \hat{y}_i^{(m)})$, а потім — кореляційну матрицю Пірсона по всіх $n = 28\,512$ спостереженнях. Результати наведено у таблиці 3.11.

Таблиця 3.11 – Кореляційна матриця логарифмічних помилок моделей на фолді 3

	Однорівнева	*Довгий-лаг*	*Короткий-лаг*	*Гібрид*
Однорівнева	1,000	0,976	0,972	0,968
Довгий-лаг	0,976	1,000	0,965	0,966
Короткий-лаг	0,972	0,965	1,000	0,964
Гібрид	0,968	0,966	0,964	1,000

Середнє абсолютне значення позадіагональних коефіцієнтів становить 0,969 — стійко висока кореляція, що відображає консистентність моделей у виявленні спільного прогностичного сигналу. Така узгодженість має зрозумі-

ле структурне пояснення: усі чотири моделі ґрунтуються на спільному ядрі ознак — лагових значеннях та ковзних середніх історії продажів, статичних категоріальних ознаках та календарних змінних. Архітектурні модифікації змінюють спосіб обробки цього ядра, але не його склад.

Висока кореляція, попри інтуїтивне сприйняття як обмеження, є радше позитивною характеристикою системи: вона свідчить, що всі чотири моделі коректно та послідовно вирішують одну й ту саму задачу. Водночас ненульова частина дисперсії помилок, не пояснена спільним сигналом, залишає простір для часткової диверсифікації при ансамблюванні. Найменшу кореляцію (0,964) демонструє пара гібрид — короткий-лаговий варіант.

3.4.4 Підбір ваг та результати ансамблю

Ансамбль чотирьох моделей формується як зважене середнє їхніх прогнозів у просторі сирих продажів:

$$\hat{y}_i^{\text{ens}} = \sum_{m=1}^{4} w_m \, \hat{y}_i^{(m)}, \qquad \sum_{m=1}^{4} w_m = 1, \quad w_m \geq 0. \tag{3.4}$$

Об'єднання у просторі сирих продажів, а не у логарифмічному, обрано тому, що цільова метрика RMSLE обчислюється саме на сирих значеннях. Ваги w_m обираються методом обмеженої оптимізації SLSQP на фолді 3, з обмеженнями неперевищення одиниці у сумі та невід'ємності. Цільовою функцією є RMSLE на валідаційному вікні фолду 3. Для порівняння обчислюється також RMSLE при рівних вагах. Результати наведено у таблиці 3.12.

Оптимальний ансамбль покращує RMSLE на фолді 3 з 0,4016 до 0,3987 — різниця 0,0029, що є невеликим, але реальним поліпшенням. Структура оптимальних ваг є інформативною. Однорівнева модель та довгий-лаговий варіант отримали практично однакові ваги $\approx 0{,}41$, що відображає їхню статистичну невідмінність. Короткий-лаговий варіант отримав вагу 0,149, гібридна модель —

Таблиця 3.12 – Підбір ваг ансамблю та результати на фолді 3

Конфігурація	$w_{одно}$	$w_{довг}$	$w_{коротк}$	$w_{гібр}$	*RMSLE*
Краща окрема модель (однорівнева)	—	—	—	—	0,4016
Рівні ваги	0,250	0,250	0,250	0,250	0,3996
Оптимальні ваги (SLSQP)	0,405	0,422	0,149	0,024	0,3987
Округлені ваги	0,40	0,40	0,15	0,05	—

мінімальний внесок 0,024.

Для побудови фінального ансамблю ваги округлено до більш стабільної конфігурації $(0{,}40,\ 0{,}40,\ 0{,}15,\ 0{,}05)$, що зберігає сумарне нормування та робастніше до випадкових варіацій оптимуму на одному фолді. Гібридна модель залишається у ансамблі з малою, але ненульовою вагою, що зберігає компонент архітектурної різноманітності. З огляду на пропорційне співвідношення між помилкою на фолді 3 і середньою CV-помилкою окремих моделей, очікувана CV-RMSLE ансамблю становить близько 0,385.

3.5 Аналіз помилок та обмежень системи

Маючи фінальну однорівневу модель з CV-RMSLE $= 0{,}387$ та зважений ансамбль з очікуваним поліпшенням до $\approx 0{,}385$, доцільно зосередити увагу на структурі залишкової помилки: де саме і чому модель помиляється. Така діагностика характеризує, які компоненти ознакового простору насправді несуть прогностичний сигнал (3.5.1), виявляє товарні категорії, на яких модель працює стабільно та нестабільно (3.5.2), та ідентифікує систематичні режими помилок — ситуації, у яких модель припускається помилок не випадково, а внаслідок структурних обмежень обраного підходу (3.5.3). Усі аналізи виконано на одному фолді (фолд 3) відповідно до домовленості з підрозділу 2.5.

3.5.1 Важливість ознак фінальної моделі

Алгоритм LightGBM надає два взаємодоповнюючі показники важливості. `Gain` — сумарне зменшення функції втрат, що його приписують усім розщепленням за певною ознакою; `split` — кількість разів використання ознаки для розщеплення. Ознака з високим `gain` та низьким `split` виконує концентровану важливу роботу через кілька критичних розщеплень; ознака з низьким `gain` та високим `split` виконує розподілене тонке налаштування через багато дрібних розщеплень.

Найважливіші 15 ознак фінальної моделі за `gain` наведено у таблиці 3.13.

Таблиця 3.13 – Важливість ознак фінальної моделі за метриками `gain` і `split`

Ознака	*Gain, %*	*Split, %*
`sales_rmean7_lag16`	56,17	5,71
`sales_rmean28_lag16`	23,21	4,77
`sales_rmean56_lag16`	8,28	6,97
`sales_lag16`	4,27	3,25
`sales_lag28`	1,75	4,29
`family`	1,70	15,54
`t_idx`	0,88	10,31
`store_nbr`	0,56	19,17
`hol_navidad`	0,54	0,71
`onpromotion`	0,51	2,08
`day`	0,40	4,54
`week_of_year`	0,40	6,50
`hol_primer_dia_del_ano`	0,39	0,65
`month`	0,37	3,87
`dow`	0,26	4,37

Першою і найбільш помітною характеристикою є домінування ознак історії продажів. П’ять ознак цієї групи разом дають 93,7% сумарного `gain` — майже весь прогностичний сигнал, який модель навчилася використовувати, концентрований у власній історії кожного ряду. Особливо виділяється семиденне

ковзне середнє sales_rmean7_lag16 з 56,17% gain — одна-єдина ознака, що сумарно перевищує внесок усіх інших разом узятих. Це підтверджує ключове методологічне рішення з підрозділу 2.4.3.

Другою характеристикою є особлива роль статичних категоріальних ознак як умовних змінних. Ознаки family, store_nbr, t_idx та week_of_year показують відносно низький gain (0,40–1,70%), але дуже високий split (6,50–19,17%). Модель майже не використовує ці ознаки для масштабних розщеплень, але постійно звертається до них для тонкого налаштування правил по контекстах. Ідентифікатор магазину store_nbr з 19,17% split є найбільш активно використовуваним розщепленням.

Третя характеристика стосується внеску екзогенних факторів. Найбільший gain серед них мають дві святкові ознаки, відібрані за data-driven процедурою з пункту 3.3.2: hol_navidad (0,54%) та hol_primer_dia_del_ano (0,39%) — обидві відповідають дням масового закриття магазинів. Решта шести гранулярних ознак свят разом дають менше ніж 0,01% gain, а прапорець вікна землетрусу — також практично нульовий внесок. Ефекти менш виражених свят та структурної події вже частково кодуються ковзними середніми та календарними ознаками; явні бінарні прапорці додають деталізацію лише там, де вона є необхідною (різдвяні та новорічні закриття).

3.5.2 Розподіл помилок за товарними категоріями

Якість прогнозу нерівномірно розподілена між 33-ма товарними категоріями. Для кожної категорії на валідаційному вікні фолду 3 обчислено три величини: RMSLE, середнє значення продажів та середнє логарифмічне зміщення bias — останнє показує, чи систематично модель занижує (від'ємний bias) або завищує (додатний bias) прогноз. Top-5 найскладніших та top-5 найпростіших категорій наведено у таблиці 3.14.

Розподіл якості за категоріями виявляє чітку закономірність: модель найкраще працює на високообсягових товарах щоденного попиту (DAIRY,

Таблиця 3.14 – Якість прогнозу за товарними категоріями (фолд 3)

Категорія	*Сер. продажі*	*RMSLE*	*Bias*
Найскладніші:			
SCHOOL AND OFFICE SUPPLIES	59,9	0,676	+0,218
GROCERY II	29,0	0,664	−0,194
LINGERIE	7,8	0,633	+0,084
CELEBRATION	12,5	0,549	−0,054
HARDWARE	1,4	0,526	+0,016
Найпростіші:			
DAIRY	834,5	0,146	−0,041
GROCERY I	4 607,0	0,156	−0,020
PRODUCE	2 288,8	0,158	−0,042
BREAD/BAKERY	540,3	0,181	−0,034
DELI	303,1	0,186	−0,052

GROCERY I, PRODUCE, BREAD/BAKERY) з RMSLE 0,15–0,19, і найгірше — на низькообсягових категоріях з нерегулярним або сезонним попитом (SCHOOL AND OFFICE SUPPLIES, GROCERY II, LINGERIE, CELEBRATION) з RMSLE 0,55–0,68. Така залежність відображає природу дозованого сигналу: на щоденних товарах історія є щільною, ковзні середні стабільні; на сезонних та переривчастих категоріях історія часто містить нулі або поодинокі сплески.

Аналіз `bias` додає шар деталізації. Більшість категорій мають невеликий `bias`, близький до нуля. Виділяються дві помітні аномалії: SCHOOL AND OFFICE SUPPLIES демонструє додатний `bias` +0,218 (модель систематично занижує прогноз), а GROCERY II — від'ємний `bias` −0,194 (модель систематично завищує). Ці два систематичних зміщення вказують на існування специфічних режимів помилок, що детально розглядаються у наступному пункті.

3.5.3 Систематичні режими помилок

Для виявлення конкретних механізмів виникнення великих помилок проаналізовано десять спостережень фолду 3 з найбільшим за модулем логарифмі-

чним відхиленням прогнозу від реального значення (таблиця 3.15).

Таблиця 3.15 – Десять найгірших прогнозів фінальної моделі (фолд 3)

Дата	*Магазин*	*Категорія*	*y, факт*	*ŷ, прогноз*	*log err*
2017-08-12	38	SCHOOL AND OFFICE SUPPLIES	48,0	0,0	+3,87
2017-08-06	19	GROCERY II	0,0	33,3	−3,54
2017-08-13	19	GROCERY II	1,0	54,2	−3,32
2017-08-15	38	SCHOOL AND OFFICE SUPPLIES	27,0	0,0	+3,29
2017-08-13	26	GROCERY II	0,0	19,7	−3,03
2017-08-14	19	GROCERY II	2,0	52,0	−2,87
2017-08-08	14	GROCERY II	0,0	16,3	−2,85
2017-08-08	19	GROCERY II	1,0	32,1	−2,81
2017-08-09	26	GROCERY II	0,0	12,6	−2,61
2017-08-08	36	HOME AND KITCHEN I	182,0	14,1	+2,49

Топ найгірших прогнозів демонструє два чітких систематичних режими помилок.

Перший режим — припинення реалізації категорії у конкретних магазинах. Сім з десяти найгірших прогнозів є спостереженнями GROCERY II у магазинах 14, 19 та 26 у серпні 2017 року. Реальні продажі цієї категорії різко впали до нуля або одиниць, тоді як модель продовжувала прогнозувати десятки одиниць. Природа помилки прозора: ковзні середні історії продажів відображають середній рівень продажів за попередні 7–56 днів, і якщо категорія перестає продаватися у магазині, модель не має жодного механізму ідентифікації цієї події — вона продовжує прогнозувати від попереднього рівня, поки ковзне середнє не догодиться до нового стану. Систематичний від'ємний `bias` категорії GROCERY II у пункті 3.5.2 є агрегованим проявом цього самого режиму.

Другий режим — слабка реакція на старт сезонного попиту. Два з найгірших прогнозів стосуються категорії SCHOOL AND OFFICE SUPPLIES у магазині 38 у середині серпня. Реальні продажі склали 27 та 48 одиниць (характерний сплеск початку навчального року), а прогноз був близьким до нуля. Цей режим є дзеркальним до першого: реальні продажі різко *зросли*, а ковзні середні ще не встигли це відобразити. Систематичний додатний `bias` категорії SCHOOL AND OFFICE SUPPLIES також пояснюється цим самим механізмом.

Спільна риса обох режимів полягає у структурній властивості моделі —

сильній залежності прогнозу від нещодавньої історії. Ця властивість є джерелом 94% `gain` моделі і причиною її загальної високої якості, водночас і обмежувальною характеристикою: модель за визначенням не здатна випереджати реальні структурні зрушення, що виходять за межі діапазону історичних коливань. Виявлені обмеження є не дефектом конкретної реалізації, а характеристикою класу моделей, що базуються на лагованих представленнях власної історії ряду.

Сукупний внесок десяти найгірших спостережень у CV-RMSLE є помітним, але обмеженим: навіть гіпотетична ідеальна корекція всіх десяти прогнозів змінила б агреговану метрику приблизно на 0,003. Описані систематичні режими є рідкісними подіями у масштабі всього валідаційного вікна (10 з 28 512), і саме тому модель досягає високої агрегованої якості, навіть демонструючи яскраві окремі помилки.

3.6 Порівняльний аналіз результатів

У попередніх підрозділах побудовано і оцінено за крос-валідацією сім різних моделей. Повний обсяг порівняння можливий лише за наявності оцінок на незалежному цільовому прогнозному вікні (16–31 серпня 2017 року) — даних, що не входили у жоден етап навчання чи валідації.

Оцінка моделі на цільовому прогнозному вікні отримана у такий спосіб. Модель повторно навчено на повному навчальному наборі без виділення валідаційного фолду; кількість раундів градієнтного бустингу зафіксовано на округленому середньому значенні `best_iter` по трьох фолдах крос-валідації, що для основних моделей становить 2 600. Сформовані прогнози оцінено за RMSLE на основі реальних значень продажів цього періоду, які не входять до відкритого навчального набору. Для ансамблю всі чотири моделі-компоненти переучуються на повному наборі окремо за тими ж 2 600 раундами, після чого їхні прогнози комбінуються з округленими вагами $(0{,}40,\ 0{,}40,\ 0{,}15,\ 0{,}05)$.

Зведене порівняння всіх моделей наведено у таблиці 3.16.

Таблиця 3.16 – Зведене порівняння моделей розділу 3

Модель	*К-сть ознак*	*CV-RMSLE*	*Тестове вікно*	*Розрив CV → тест*
Базова Ridge (3.1)	18	1,030	1,042	+0,012
Повна LightGBM (3.2)	27	0,388	—	—
Фінальна LightGBM (3.3.3)	29	0,387	0,427	+0,040
Гібрид Ridge + LightGBM (3.4.1)	29	0,391	0,449	+0,058
Короткий-лаговий варіант (3.4.2)	27	0,391	—	—
Довгий-лаговий варіант (3.4.2)	28	0,387	—	—
Зважений ансамбль (3.4.4)	—	∼0,385	**0,425**	+0,040

Узгодженість крос-валідаційної та незалежної оцінок

Розрив між CV-RMSLE та оцінкою на цільовому вікні для базової Ridge (+0,012) та основних моделей LightGBM (+0,040) є стабільним: фінальна однорівнева LightGBM та ансамбль показують ідентичний розрив +0,040. Така узгодженість є одним з найсильніших підтверджень адекватності обраної валідаційної процедури з підрозділу 2.5. Різниця в розривах між базовою моделлю та LightGBM пояснюється точністю прогнозу: Ridge має значно більше шумоподібний прогноз, тоді як точніші моделі LightGBM є більш чутливими до невеликих сезонних відмінностей між валідаційними та тестовим вікнами.

Виняток становить гібридна модель з розривом +0,058 — на 0,018 більшим за основні моделі LightGBM. Це свідчить про дещо нижчу узагальнювальну здатність гібридного підходу: Ridge-компонента, що становить основу гібриду, є більш чутливою до сезонних змін між червнем-липнем (валідаційні вікна) та серпнем (тестове вікно).

Підсумкове ранжування моделей

За оцінкою на цільовому прогнозному вікні моделі розташовуються у такому порядку (за зростанням помилки): зважений ансамбль (0,425) < фінальна однорівнева LightGBM (0,427) < гібрид (0,449) < базова Ridge (1,042). Розрив між ансамблем та фінальною однорівневою моделлю становить лише 0,002, проте напрям виграшу узгоджений з результатом фолду 3, де ансамбль покращував окрему модель на 0,003. Виграш від ансамблювання, попри високу кореляцію помилок між моделями-компонентами, є відтворюваним і на CV, і на незалежному тестовому вікні.

Найбільший якісний приріст відбувається при переході від базової лінійної моделі до глобальної моделі LightGBM: CV-RMSLE падає з 1,030 до 0,387

(зменшення на 62%), оцінка на цільовому вікні — з 1,042 до 0,427 (зменшення на 59%). Це відображає спільний внесок двох ключових архітектурних змін: введення ознак історії продажів та переходу від лінійної функціональної форми до нелінійних дерев градієнтного бустингу. Подальші вдосконалення дають додатковий, проте на порядок менший виграш: близько 0,002 RMSLE на тестовому вікні.

Інтерпретація величин

Абсолютна величина досягнутого RMSLE 0,425 у логарифмічному просторі відповідає середній відносній помилці прогнозу близько 50% — значущій помилці у термінах окремих спостережень, але типовій для прогнозування у роздрібній торгівлі з 1 782 рядами різного масштабу та переривчастого характеру. На високообсягових щоденних товарах (DAIRY, GROCERY I, PRODUCE) якість прогнозу значно вища — RMSLE 0,15–0,16, що відповідає середній відносній помилці близько 17% — рівень, що становить практичну цінність для управління запасами таких товарів.

Висновки до розділу 3

У третьому розділі побудовано та емпірично оцінено серію прогнозних моделей: від базової Ridge-регресії на екзогенних факторах до зваженого ансамблю чотирьох моделей градієнтного бустингу. Оцінювання проведене за двома незалежними процедурами — трифолдовою крос-валідацією та оцінкою на цільовому прогнозному вікні 16–31 серпня 2017 року.

Базова Ridge-регресія досягла CV-RMSLE $= 1{,}030$, що задало нижню межу якості. Введення ознак історії продажів та перехід до нелінійної функціональної форми LightGBM знизили помилку до 0,388 — скорочення на 62%. Подальший data-driven редизайн ознак свят, що замінив три грубі прапорці вісьмома гранулярними, дав фінальну однорівневу модель з CV-RMSLE $= 0{,}387$ та оцінкою 0,427 на цільовому вікні. Зважений ансамбль чотирьох моделей покращив тестову оцінку до 0,425.

Абляційний аналіз виявив суттєву нерівноцінність екзогенних сигналів. Промоакції є єдиним екзогенним фактором з вимірюваним позитивним внеском ($-0{,}034$ RMSLE при видаленні). Грубі прапорці свят та ознаки транзакцій залишились у межах шумового порогу, ціна нафти показала навіть невелику негативну дельту. Аналіз важливості ознак підтвердив центральну роль історії продажів: п'ять відповідних ознак сумарно дають 93,7% `gain`, причому семиденне ковзне середнє самостійно становить 56%. Статичні категоріальні ознаки виконують функцію умовних змінних — з низьким `gain`, але високим `split`.

Узгодженість розриву CV $\rightarrow$ тестове вікно ($+0{,}012$ для Ridge, $+0{,}040$ для основних LightGBM-моделей) підтверджує адекватність обраної валідаційної процедури з підрозділу 2.5. Аналіз помилок виявив два систематичні режими великих помилок — припинення реалізації категорії у конкретних магазинах та слабка реакція на сезонний старт. Обидва пов'язані з фундаментальною властивістю історико-залежних моделей: модель за визначенням не може випередити структурні зрушення, що виходять за межі діапазону історичних коливань.

Сукупно у розділі побудовано систему прогнозування з якістю 0,425 RMSLE на цільовому вікні — скорочення помилки відносно базової моделі на 59%. Методологічна цінність дослідження виявляється у data-driven процедурі редизайну ознак свят, у детальному декомпозиційному аналізі внеску екзогенних факторів та у виявленні структурних режимів помилок, що характеризують клас історико-залежних моделей прогнозування.

ВИСНОВКИ

У дипломній роботі досліджено та реалізовано інтелектуальну систему прогнозування часових рядів продажів у мережі роздрібної торгівлі з урахуванням екзогенних факторів різної природи. Дослідження виконано на реальних даних мережі продовольчих магазинів Еквадору (1 782 часових ряди, період з 1 січня 2013 року по 15 серпня 2017 року) та охоплює повний цикл — від аналізу даних до побудови, навчання та порівняння прогнозних моделей.

У першому розділі досліджено предметну область та виконано огляд методів прогнозування часових рядів продажів — від класичних статистичних підходів до методів машинного навчання та глибокого навчання. Запропоновано класифікацію екзогенних факторів за природою впливу, доступністю на момент прогнозування та характером взаємодії з цільовою змінною. Обґрунтовано вибір градієнтного бустингу на деревах рішень як основного класу моделей з огляду на високу розмірність задачі та необхідність глобального моделювання тисяч взаємопов'язаних рядів.

У другому розділі виконано розвідувальний аналіз даних та сформовано робочий ознаковий простір. Виявлено множинну сезонність, виражений тренд, диференційований вплив промоакцій ($+43\%$ до $+208\%$ залежно від категорії), спурійний характер кореляції з ціною нафти та структурний злом, спричинений землетрусом 2016 року. Обґрунтовано логарифмічне перетворення цільової змінної, що забезпечує точну еквівалентність оптимізації стандартної функції втрат MSE та цільової метрики RMSLE. Сконструйовано 27-ознаковий простір з семи груп та розроблено валідаційну схему з трьома непересічними фолдами розширювального вікна.

У третьому розділі побудовано та емпірично оцінено серію прогнозних моделей: базову Ridge-регресію, глобальну LightGBM, її фінальну версію з гранулярними ознаками свят, гібридну модель Ridge + LightGBM, два варіанти з модифікованим лаговим горизонтом та зважений ансамбль чотирьох моделей. Базова модель досягла оцінки 1,042 RMSLE на цільовому прогнозному вікні,

фінальна однорівнева LightGBM — 0,427, зважений ансамбль — 0,425. Скорочення помилки відносно базової моделі складає 59% на тестовій вибірці.

Розроблена інтелектуальна система прогнозування являє собою інтегрований конвеєр з шести основних компонент: модулю попередньої обробки даних, що узгоджує п’ять взаємопов’язаних джерел у єдиний навчальний кадр; модулю конструювання ознак, що формує 29-вимірний ознаковий простір на основі семи типів сигналів; data-driven процедури відбору ознак свят, що автоматично вимірює ефект кожного свята та обирає оптимальний розмір набору через крос-валідацію; ансамблю чотирьох моделей LightGBM з різними архітектурними та ознаковими конфігураціями; механізму автоматичного підбору ваг ансамблю шляхом SLSQP-оптимізації; та валідаційної підсистеми з трифолдовою крос-валідацією та незалежним оцінюванням на цільовому прогнозному вікні. Кожна компонента незалежно обґрунтована та емпірично оцінена, що забезпечує модульність системи та можливість заміни окремих компонент без перебудови всього конвеєра. Окремим оригінальним внеском роботи є процедура data-driven редизайну ознак свят, систематичний абляційний аналіз внеску груп ознак та виявлення двох систематичних режимів помилок, що характеризують фундаментальне обмеження класу історико-залежних моделей.

Усі задачі, поставлені у підрозділі 1.10, виконані повністю: проведено огляд методів прогнозування та обґрунтовано вибір моделювання; виконано розвідувальний аналіз даних і сформовано стратегію передобробки; розроблено ознаковий простір; побудовано та навчено прогнозні моделі різних класів із застосуванням коректної стратегії валідації; виконано порівняльний аналіз моделей та дослідження впливу окремих груп екзогенних факторів на якість прогнозу. Отримані результати мають практичну цінність для управління запасами у мережах роздрібної торгівлі: RMSLE на високообсягових категоріях щоденного попиту на рівні 0,15–0,16 відповідає середній відносній помилці прогнозу близько 17%, що становить рівень практичної застосовності у виробничих системах. Виявлені обмеження визначають напрями подальшого розвитку — доповнення системи механізмами виявлення режимних змін та інтеграція майбутніх відомих коваріатів про планові структурні зміни в асортименті магазинів.

ПЕРЕЛІК ПОСИЛАНЬ

1. National Retail Federation, “NRF forecasts 4.4% annual retail sales growth with new economic model,” 2026. [Online]. Available: https://nrf.com/media-center/press-releases/nrf-forecasts-4-4-annual-retail-sales-growth-with-new-economic-model.

2. R. Fildes, S. Ma, and S. Kolassa, “Retail forecasting: Research and practice,” *International Journal of Forecasting*, vol. 38, no. 4, pp. 1283–1318, 2022.

3. R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. OTexts, 2021.

4. G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Wiley, 2015.

5. McKinsey & Company, “AI-driven operations forecasting in data-light environments,” 2022. [Online]. Available: https://www.mckinsey.com/capabilities/operations/our-insights/ai-driven-operations-forecasting-in-data-light-environments.

6. F. Lazzeri, *Machine Learning for Time Series Forecasting with Python*. Wiley, 2020.

7. M. Joseph, *Modern Time Series Forecasting with Python*. Packt Publishing, 2022.

8. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

9. P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, 3rd ed. Springer, 2016.

10. S. J. Taylor and B. Letham, “Forecasting at scale,” *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

11. M. Peixeiro, *Time Series Forecasting in Python*. Manning Publications, 2022.

12. J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

13. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “LightGBM: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 3146–3154.

14. B. Auffarth, *Machine Learning for Time-Series with Python*. Packt Publishing, 2021.

15. S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “The M5 competition: Background, organization, and implementation,” *International Journal of Forecasting*, vol. 38, no. 4, pp. 1325–1336, 2022.

16. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

17. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

18. B. Lim, S. O. Arik, N. Loeff, and T. Pfister, “Temporal Fusion Transformers for interpretable multi-horizon time series forecasting,” *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

19. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

20. A. Popov, “Efficiency comparison of missing data imputation methods in predictive model creation,” *System Research and Information Technologies*, no. 1, pp. 32–43, 2025. DOI: https://doi.org/10.20535/SRIT.2308-8893.2025.1.03.

21. A. Cook, DanB, Inversion, and R. Holbrook, “Store Sales — Time Series Forecasting,” Kaggle, 2021. [Online]. Available: https://kaggle.com/competitions/store-sales-time-series-forecasting.

22. “Economy of Ecuador,” *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Economy_of_Ecuador.

23. U.S. Geological Survey, “Magnitude 7.8 Earthquake in Ecuador,” 2016. [Online]. Available: https://www.usgs.gov/news/featured-story/magnitude-78-earthquake-ecuador.

ДОДАТОК А ЛІСТИНГ ПРОГРАМИ

Лістинг програми розміщено за посиланням:
https://github.com/kuzvera/ecuador-store-sales/blob/main/thesis.ipynb