Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Центр післядипломної освіти
(повна назва)
Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти перший (бакалаврський)

Застосування методів машинного навчання для прогнозування попиту товарів
(тема)

Виконав:
здобувач другого року навчання,
групи ІТШІпз-24-1
Микита НАРОЖНИЙ
(власне ім'я, прізвище)
Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-професійна
Освітня програма Штучний інтелект
(повна назва освітньої програми)
Керівник проф. Марина Кудрявцева
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ШІ _____ Лариса ЧАЛА
(підпис) (власне ім'я, прізвище)

2026 р.

Харківський національний університет радіоелектроніки

Факультет Центр післядипломної освіти

Кафедра Штучного інтелекту

Рівень вищої освіти перший (бакалаврський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна

Освітня програма Штучний інтелект
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Нарожному Микиті Сергійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Застосування методів машинного навчання для прогнозування попиту товарів

затверджена наказом університету від 18 травня 20 26 р. № 77 Стз
2. Термін подання здобувачем роботи до екзаменаційної комісії 17 червня 20 26 р.
3. Вихідні дані до роботи Науково-технічна література з аналізу та прогнозування часових рядів; набір даних історичних щоденних продажів товарів сегмента retail FMCG середовище розробки Google Colab; мова програмування Python та її бібліотеки

4. Перелік питань, що потрібно опрацювати в роботі
1) Аналіз предметної галузі та постановка задачі
2) Методи машинного навчання для прогнозування часових рядів
3) Проєктування та розробка системи прогнозування
4) Експериментальні дослідження та аналіз результатів

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / термін виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	18.05.2026	виконано
2	Аналіз предметної галузі	19.05.2026	виконано
3	Постановка задачі	19.05.2026	виконано
4	Аналіз даних	20.05.2026	виконано
5	Експериментальне моделювання та навчання	21.05.2026	виконано
	моделі		
6	Підготовка пояснювальної записки	23.05.2026	виконано
7	Перевірка на академічний плагіат	05.06.2026	виконано
8	Нормоконтроль	09.06.2026	виконано
9	Підготовка презентації та доповіді	09.06.2026	виконано
10	Попередній захист	10.06.2026	виконано
11	Захист перед ЕК	17.06.2026	

Дата видачі завдання 18 травня 20 26 р.

Здобувач _____
(підпис)

Керівник роботи _____ проф. Марина КУДРЯВЦЕВА
(підпис) (посада, власне ім’я, прізвище)

РЕФЕРАТ

Пояснювальна записка: 69 с., 5 рис., 1 табл., 6 дод., 22 джерела.

АНАЛІЗ ДАНИХ, ГРАДІЄНТНИЙ БУСТИНГ, МАШИННЕ НАВЧАННЯ, ПРОГНОЗУВАННЯ ПОПИТУ, РОЗДРІБНА ТОРГІВЛЯ, ЧАСОВІ РЯДИ, FMCG, LIGHTGBM.

Об’єкт дослідження – сфера роздрібної торгівлі.

Предмет дослідження – методи машинного навчання на базі градієнтного бустингу для автоматизації планування попиту.

Мета роботи – підвищення точності прогнозування попиту на товари FMCG шляхом застосування ансамблевих моделей машинного навчання. Методи дослідження: системний аналіз, розвідувальний аналіз даних, алгоритм градієнтного бустингу LightGBM, метод SHAP-значень.

Результати та їх новизна: розроблено систему прогнозування попиту за допомогою LightGBM з динамічним відбором ознак.

Новизною є використання кастомної метрики safe_MAPE для переривчастого попиту та математична інтерпретація впливу нелінійних факторів (ціни, промо-акцій, залишків) за допомогою методу SHAP.

Сфера застосування: інтеграція в операційні ERP-системи ритейлерів та виробників товарів повсякденного попиту.

Висновки та пропозиції щодо напрямку розвитку: розроблена модель успішно виявляє складні макро-тренди. У майбутньому доцільно інтегрувати до масиву даних макроекономічні індикатори та алгоритми глибокого навчання.

ABSTRACT

Bachelor’s thesis contains: 69 pp., 5 fig., 1 tabl., 6 ann., 22 references.

DATA ANALYSIS, DEMAND FORECASTING, FMCG, GRADIENT BOOSTING, LIGHTGBM, MACHINE LEARNING, RETAIL, TIME SERIES.

The object of the study is the retail sector.

The subject of the study is machine learning methods based on gradient boosting for automating demand planning.

The aim of the work is to improve the accuracy of FMCG demand forecasting by applying ensemble machine learning models.

Research methods: systems analysis, exploratory data analysis,

LightGBM gradient boosting algorithm, and SHAP values method.

Results and novelty: a demand forecasting system using LightGBM with dynamic feature selection was developed.

The novelty lies in applying a custom safe_MAPE metric for intermittent demand and mathematically interpreting the impact of nonlinear factors (prices, promotions, stock levels) using the SHAP method.

Scope of application: integration into operational ERP systems of retailers and FMCG manufacturers.

Conclusions and development proposals: the developed model successfully captures complex macro-trends.

Future work should focus on integrating macroeconomic indicators and deep learning algorithms into the dataset.

ЗМІСТ

Перелік умовних позначень, символів, одиниць, скорочень і термінів 8
Вступ 9
1 Аналіз предметної галузі та постановка задачі 11
1.1 Специфіка ринку FMCG та роль аналітики у плануванні попиту 11
1.1.1 Загальна характеристика та економічна специфіка ринку FMCG 11
1.1.2 Проблематика ланцюгів постачання та «вартість тиші» 11
1.1.3 Еволюція аналітики: від реактивної до предиктивної 12
1.1.4 Вплив аналітики на операційну ефективність та стратегічне планування 13
1.2 Аналіз факторів, що впливають на обсяги продажів 14
1.2.1 Вплив цінової політики 15
1.2.2 Роль маркетингових заходів 15
1.2.3 Рівень товарних запасів та доступність на полиці 16
1.2.4 Зовнішні фактори впливу окрім ціни, промо та залишків 17
1.3 Огляд існуючих підходів до прогнозування попиту в ритейлі 18
1.3.1 Класичні статистичні методи та їх обмеження 18
1.3.2 Методи машинного навчання та ансамблеві моделі 20
1.3.3 Моделі глибокого навчання та новітні підходи 21
1.3.4 Порівняння підходів та вибір оптимального методу, гіпотеза «Size Matters» 22
1.4 Постановка задачі прогнозування як задачі регресії 23
2 Методи машинного навчання для прогнозування часових рядів 26
2.1 Класичні статистичні методи та їх обмеження 26
2.1.1 Методи експоненційного згладжування 26
2.1.2 Моделі сімейства ARIMA 27
2.1.3 Фундаментальні обмеження класичних моделей у ритейлі 27
2.2 Ансамблеві методи навчання: випадковий ліс 29

2.2.1 Архітектура та принцип роботи 29
2.2.2 Переваги використання Random Forest у ритейлі 30
2.2.3 Обмеження алгоритму для часових рядів 30
2.3 Градієнтний бустинг: особливості архітектури LightGBM 31
2.3.1 Концепція градієнтного бустингу 31
2.3.3 Переваги LightGBM у прогнозуванні попиту FMCG 33
3 Проєктування та розробка системи прогнозування 35
3.1 Опис та огляд структури об’єкта дослідження 35
3.1.2 Результати розвідувального аналізу та оцінка якості даних 36
3.1.3 Аналіз часової динаміки наборів даних 37
3.2 Попередня обробка даних: очищення від аномалій та обробка пропусків 40
3.2.1 Очищення від логічних аномалій 41
3.2.2 Часове відсікання для стабілізації ряду 41
3.2.3 Кодування категоріальних змінних 42
3.3 Генерація ознак: часові лаги та ковзні середні 42
3.4 Ковзні статистичні показники та тренди 44
3.5 Лаговані екзогенні фактори та групові метрики 44
3.6 Тимчасові, календарні та сезонні сигнали 45
3.7 Сигнали дефіциту та інтерпретація бізнес-взаємодій 46
3.8 Кодування категоріальних змінних та стратегія валідації 46
4 Експериментальні дослідження та аналіз результатів 48
4.1 Аналіз точності моделей на тестовій вибірці 48
4.2 Візуалізація результатів та інтерпретація важливості ознак 51
Висновки 54
Перелік джерел посилання 55
Додаток А Вихідний код програмної реалізації 57
Додаток Б Відомість кваліфікаційної роботи 69

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ARIMA – AutoRegressive Integrated Moving Average – авторегресія інтегрованого ковзного середнього;

EDA – Exploratory Data Analysis – розвідувальний аналіз даних;

EFB – Exclusive Feature Bundling – взаємовиключне об'єднання ознак;

ERP – Enterprise Resource Planning – система планування ресурсів підприємства;

ETS – Exponential Smoothing – експоненційне згладжування;

FMCG – Fast-Moving Consumer Goods – товари повсякденного попиту;

LightGBM – Light Gradient Boosting Machine – легка машина градієнтного бустингу;

LSTM – Long Short-Term Memory – довга короткострокова пам'ять;

MAE – Mean Absolute Error – середня абсолютна похибка;

MAPE – Mean Absolute Percentage Error – середня абсолютна помилка у відсотках;

ML – Machine Learning – машинне навчання;

MSE – Mean Squared Error – середньоквадратична похибка;

OOS – Out-of-Stock – дефіцит товару (відсутність товару на полиці або складі);

RMSE – Root Mean Square Error – корінь із середньоквадратичної похибки;

SHAP – SHapley Additive exPlanations – адитивні пояснення Шеплі;

SKU – Stock Keeping Unit – ідентифікатор (одиниця) товарної позиції;

SMAPE – Symmetric Mean Absolute Percentage Error – симетрична середня абсолютна помилка у відсотках;

TFT – Temporal Fusion Transformer – часовий ф'южн-трансформер.

ВСТУП

Актуальність теми. Стрімкий розвиток цифрових технологій та методів штучного інтелекту став визначальним фактором трансформації сучасних бізнес-процесів. Одним із найбільш динамічних секторів світової економіки є ринок товарів повсякденного попиту, де ефективність діяльності компаній безпосередньо залежить від точності прогнозування споживчої поведінки. В умовах жорсткої конкуренції та високої волатильності ринку традиційні підходи до планування запасів, що базуються на інтуїтивних рішеннях або простих статистичних методах, виявляються недостатніми. Помилки у прогнозуванні призводять до значних фінансових втрат: виникнення дефіциту товарів спричиняє втрату лояльності клієнтів, тоді як надлишкові запаси заморожують оборотний капітал та збільшують витрати на логістику. Актуальність виконання даної кваліфікаційної роботи визначається необхідністю створення інтелектуальних систем, здатних обробляти великі масиви багатофакторних даних для побудови високоточних прогнозів попиту.

Використання сучасних алгоритмів машинного навчання, зокрема градієнтного бустингу LightGBM, дозволяє трансформувати аналітику з описової у предиктивну, перетворюючи її на повноцінний «рушій прийняття рішень» для бізнесу.

Мета роботи полягає у підвищенні точності прогнозування попиту на товари сегменту FMCG шляхом дослідження, розробки та застосування моделей машинного навчання на базі ансамблевих алгоритмів.

Задачі дослідження. Для досягнення поставленої мети необхідно вирішити такі задачі:

– проаналізувати предметну галузь, специфіку ринку FMCG-ритейлу та фактори, що найбільше впливають на обсяги продажів;

– дослідити існуючі підходи до прогнозування часових рядів, визначити обмеження класичних статистичних моделей та обґрунтувати вибір методів машинного навчання;

– здійснити розвідувальний аналіз EDA, попередню обробку історичного масиву даних продажів та генерацію нових ознак, включаючи часові лаги й ковзні вікна;

– виконати програмну реалізацію системи навчання регресійних моделей на базі алгоритму LightGBM для різних товарних категорій;

– провести експериментальну оцінку точності розроблених моделей на тестовій вибірці за допомогою відповідних метрик RMSE, MAPE, SMAPE, R2 та виконати інтерпретацію впливу ознак.

Об’єкт дослідження – сфера роздрібної торгівлі.

Предмет дослідження – методи, моделі та алгоритми машинного навчання на базі градієнтного бустингу для автоматизації планування попиту на рівні окремих товарних позицій.

Методи дослідження. Для розв’язання поставлених у роботі задач використано: методи системного аналізу (для дослідження предметної галузі та постановки задачі); методи математичної статистики та розвідувального аналізу даних; методи машинного навчання, зокрема ансамблеві алгоритми дерев рішень та градієнтний бустинг для побудови предиктивних регресійних моделей; метод SHAP-значень для інтерпретації результатів моделювання та визначення важливості ознак.

Практичне значення отриманих результатів. Результати кваліфікаційної роботи мають безпосереднє практичне значення для ритейлерів, дистриб’юторів та виробників FMCG-сектору. Впровадження розробленого конвеєра прогнозування на базі моделі LightGBM дозволяє враховувати складні нелінійні фактори попиту, значно підвищити операційну ефективність, оптимізувати рівні страхових запасів та автоматизувати прийняття рішень щодо поповнення складів у динамічному ринковому середовищі.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Специфіка ринку FMCG та роль аналітики у плануванні попиту

1.1.1 Загальна характеристика та економічна специфіка ринку FMCG

Ринок товарів повсякденного попиту (FMCG – Fast-Moving Consumer Goods) визначається екстремальною швидкістю обертання продукції, мінімальним рівнем маржинальності (так звані «razor-thin profit margins») та споживчою базою, якій притаманна надзвичайно висока чутливість до цінових коливань і низька вартість перемикання між брендами. Товари у цьому секторі надходять від виробничих ліній до полиць роздрібних магазинів дуже швидко, а поріг їхньої оборотності часто вимірюється днями або тижнями, а не місяцями. Глобальний обсяг цього ринку є колосальним: для прикладу, у 2024 році він оцінювався у 14,1 трлн доларів США.

Високий рівень конкуренції у сфері FMCG зумовлений трьома основними факторами: жорстко обмеженим простором на полицях магазинів, низькою диференціацією товарів (оскільки різниця між продуктами часто полягає лише у маркетинговому позиціонуванні) та постійними ціновими війнами. Через це компаніям необхідно інвестувати значні кошти в маркетинг, стимулювання збуту та постійно адаптуватися до мінливих споживчих уподобань. У цьому середовищі межа між прибутковим кварталом і значними фінансовими втратами часто залежить виключно від здатності компанії вчасно синхронізувати обсяги постачання з надзвичайно волатильними сигналами попиту.

1.1.2 Проблематика ланцюгів постачання та «вартість тиші»

Фундаментальною проблемою галузі є так звана «вартість тиші» (cost of silence) у ланцюгах постачання. Історично FMCG-бренди функціонували

в реактивному стані: про відсутність товару на полиці (Out-of-Stock, OOS) або провал промо-акції дізнавалися лише після скарг клієнтів або зі звітів, що надходили вже після завершення звітного періоду.

Ситуація ускладнюється надзвичайно фрагментованими мережами дистрибуції, які охоплюють гіпермаркети, міні-маркети, аптеки та невеликі незалежні магазини. На відміну від бізнес-моделей прямих продажів (direct-to-consumer), виробники FMCG часто не мають точної інформації про кінцеву точку продажу в режимі реального часу, покладаючись на дані багаторівневих посередників. Водночас дефіцит товару є критичним явищем: згідно з дослідженнями, близько 64% покупців у разі відсутності потрібного товару просто купують продукт конкурента. Це призводить до безповоротної втрати продажів та довгострокового зниження лояльності до бренду. У глобальному масштабі втрати роздрібної торгівлі через ситуації OOS оцінюються у сотні мільярдів доларів щорічно (близько 4% загального річного доходу).[8] З іншого боку, надлишкові запаси (overstocking) заморожують оборотний капітал компанії, значно збільшують витрати на зберігання та неминуче призводять до псування товарів з коротким терміном придатності.

1.1.3 Еволюція аналітики: від реактивної до предиктивної

У відповідь на описані вище виклики, роль аналітики даних у секторі FMCG зазнала кардинальних змін. Сучасна аналітика більше не є просто інструментом описового вимірювання минулих результатів (descriptive reporting) – вона перетворилася на повноцінний «рушій прийняття рішень» (decision engine). Перехід від традиційних електронних таблиць, які оновлювалися щомісяця, до інтегрованих систем на базі алгоритмів штучного інтелекту дозволяє компаніям оперативно відповідати на головні комерційні питання: де знаходяться точки росту і які саме тактичні дії необхідні для їх досягнення?.

Провідні організації, які впроваджують предиктивну оптимізацію, здатні досягати зростання доходів приблизно на 17% вище за середні показники по ринку (генеруючи додаткові прибутки за рахунок уникнення втрат та оптимізації процесів). Сучасні моделі машинного навчання (ML) здатні аналізувати понад 200 сигналів попиту одночасно (включаючи POS-дані, погоду, соціальні настрої, макроекономічні індикатори) з щоденним оновленням. Це дозволяє досягати точності прогнозування на рівні окремих артикулів (SKU) до 90–95%, що є колосальним стрибком порівняно з 55–65% точності, яку забезпечували традиційні методи на базі електронних таблиць. Предиктивні моделі здатні виявляти зміни у попиті за 2–4 тижні до того, як вони вплинуть на виробничі лінії.

1.1.4 Вплив аналітики на операційну ефективність та стратегічне планування

Точне планування попиту глибоко інтегроване в усі етапи створення вартості компанії. Оптимізація ланцюга постачання та виробництва: З точки зору операційної діяльності, високоточні прогнози дозволяють скоротити рівень резервних запасів (safety stock) на суму до 35%. Більше того, вони мінімізують залежність від екстрених, незапланованих виробничих циклів, вартість яких може перевищувати планові у 3,2 раза. Прогнозування усуває дисбаланс, забезпечуючи наявність потрібного товару в потрібному місці та в потрібний час.

Управління ціноутворенням: Аналітика відіграє вирішальну роль у розумінні власної та перехресної цінової еластичності (cross-price elasticity). Предиктивне моделювання дозволяє компаніям симулювати різноманітні сценарії для знаходження оптимальних цінових точок. Наприклад, необґрунтоване зниження ціни на 5% може дати приріст обсягу продажів лише на 2%, що у підсумку призведе до падіння загального прибутку компанії на 4%.

Оцінка ефективності промо-акцій (Promo Effectiveness): Маркетингові акції є величезною статтею витрат для FMCG. Аналітика допомагає виявити, які знижки генерують реальний інкрементальний обсяг продажів, а які лише субсидують існуючий попит (коли лояльні споживачі купують зі знижкою те, що вони б придбали й за повною ціною). Системи на базі ML математично враховують складні ефекти канібалізації (коли знижка на один продукт «з'їдає» продажі іншого товару цього ж бренду) та ефекти ореолу (зростання продажів супутніх товарів).

Таким чином, застосування методів машинного навчання у прогнозуванні дозволяє розв'язати проблему так званого «холодного старту» (cold start) під час запуску нових продуктів та ефективно моделювати нелінійні патерни складних взаємозв'язків даних.

Це трансформує підхід FMCG-компаній: від інтуїтивного планування до стратегічної оптимізації, де кожен бізнес-крок спирається на глибинний аналіз даних, мінімізуючи операційні витрати та максимізуючи прибуток у динамічному ринковому середовищі.

1.2 Аналіз факторів, що впливають на обсяги продажів

Обсяги продажів у секторі товарів повсякденного попиту (FMCG) формуються під впливом складної нелінійної взаємодії внутрішніх стратегічних важелів (які контролюються ритейлером або виробником) та зовнішніх ринкових змінних.

Розуміння природи цих факторів та їхнього впливу на поведінку споживачів є фундаментальною передумовою для побудови точних предиктивних моделей прогнозування попиту. До ключових факторів, що безпосередньо зумовлюють динаміку продажів на рівні окремих артикулів (SKU), належать ціноутворення, маркетингові заходи (промо-активності), рівень товарних запасів (залишки), а також низка макроекономічних та сезонних змінних.

1.2.1 Вплив цінової політики

Цінова політика є одним із найбільш безпосередніх та потужних рушіїв споживчої поведінки в галузі FMCG. У сучасному висококонкурентному середовищі роздрібної торгівлі покупці відзначаються винятковою чутливістю до цінових коливань. Це зумовлено обмеженим простором на полицях та низьким рівнем диференціації товарів: часто різниця між продуктами конкуруючих брендів полягає виключно у їхньому маркетинговому позиціонуванні.

Експериментальні дослідження та моделювання важливості ознак (Feature Importance) доводять, що ціна (price_unit) має стійкий негативний вплив на попит: вищі ціни загалом призводять до зменшення обсягів реалізації.

Предиктивна аналітика дозволяє виробникам не просто фіксувати падіння продажів, а математично оцінювати власну та перехресну цінову еластичність (own-price and cross-price elasticity). Це дає можливість симулювати різні сценарії для знаходження оптимальної цінової точки, яка максимізує дохід без критичної ерозії маржинальності. Наприклад, хибне рішення про зниження ціни на 5% у преміальному сегменті може дати приріст загального обсягу продажів лише на 2%, що у підсумковому фінансовому результаті призведе до падіння загального прибутку компанії на 4%. Отже, врахування ціни як вхідної змінної є обов'язковим для уникнення збиткових стратегій.

1.2.2 Роль маркетингових заходів

Маркетингові акції, такі як прямі знижки, пропозиції «купи один, отримай інший у подарунок» (BOGO), тимчасові зниження ціни або подарунки за покупку, генерують значні короткострокові стрибки обсягів реалізації і є однією з найбільших статей витрат для FMCG-

компаній. Наявність активного промо (promotion_flag) є потужним позитивним сигналом, що стимулює продажі.

Проте аналіз впливу промо-активностей є складним завданням через наявність двох взаємопов'язаних ефектів:

– ефект канібалізації: ситуація, коли знижка на один продукт стимулює його продажі, але одночасно «з'їдає» (зменшує) обсяги продажів іншого товару цього ж бренду;

– ефект ореолу: позитивний вплив, за якого промо-акція на один базовий товар стимулює приріст продажів супутніх товарів.

Традиційне ручне (або засноване на електронних таблицях) планування часто переоцінює реальний вплив промо на 18–25%, оскільки аналітики не здатні вручну змоделювати ці складні взаємодії та відрізнити реальний інкрементальний приріст продажів від субсидованого попиту (коли покупець придбав би товар і без знижки). Крім того, існує проблема виконання умов акції на місцях (compliance). Часто прогнозовані продажі не досягаються не через погану стратегію, а через те, що в конкретному магазині просто забули розмістити акційний цінник (shelf talker). Інтеграція даних про промо в моделі машинного навчання дозволяє автоматично будувати криві приросту (uplift curves) і прогнозувати реальний ефект від кампаній.

1.2.3 Рівень товарних запасів та доступність на полиці

Наявність товару на полиці (In-Stock) є фундаментальним фактором, без якого продажі неможливі. Вплив товарних залишків на обсяги продажів можна розглядати через призму двох крайнощів: дефіциту та надлишку.

Відсутність товару завдає найболючішого удару по фінансових показниках компанії [8]. Згідно з дослідженнями, якщо покупець не знаходить потрібного товару на полиці, у 64% випадків він не відкладає покупку, а просто купує продукт конкурента. Це призводить не лише до

миттєвої втрати доходу, але й до розмиття лояльності до бренду. У глобальному масштабі втрати роздрібної торгівлі через ситуації OOS оцінюються приблизно у 4% від загального річного доходу (близько 984 мільярдів доларів США).

З точки зору прогнозування, ситуації OOS створюють серйозну аналітичну проблему: вони спотворюють історичні дані про продажі. Коли товар відсутній на полиці, продажі дорівнюють нулю, проте це не означає, що попит на нього був нульовим. У таких випадках реальний попит стає «прихованим» або «цензурованим» (censored demand). Для побудови точних моделей прогнозування необхідно розраховувати нецензурований попит – тобто математично відновлювати ті обсяги, які компанія могла б продати, якби товар був у наявності.

Надлишкові запаси (Overstocking), навпаки, заморожують оборотний капітал компанії, збільшують складські витрати та призводять до потенційного списання товарів, що є особливо критичним для продуктів, які швидко псуються (perishables). Наявність змінної доступних залишків (stock_available) є критично важливою для моделей: низький рівень залишків або їхня відсутність виступає потужним фактором, що пригнічує прогнозований попит.

1.2.4 Зовнішні фактори впливу Окрім ціни, промо та залишків

На попит значно впливають зовнішні сигнали, які компанія не може контролювати безпосередньо.

Календарні та сезонні фактори: аналіз часових рядів у FMCG підтверджує наявність сильної сезонності та циклічності. До них належать щотижневі патерни (зростання попиту у вихідні дні), вплив днів виплати заробітної плати, а також різкі щорічні сплески під час великих свят.

Погодні умови: різкі зміни погоди (наприклад, неочікувана спека) можуть миттєво згенерувати стрибок продажів для певних категорій, як-от прохолоджувальні напої чи морозиво.

Макроекономічні індикатори: рівень інфляції, споживча впевненість та ціни на пальне здатні модулювати загальні витрати споживачів. Наприклад, у періоди економічної нестабільності поведінкові патерни змінюються: певні сегменти населення починають купувати товари меншими партіями, але частіше.

Урахування всієї цієї множини факторів ціни, промо, наявності на полиці та макро-сигналів робить класичні лінійні методи прогнозування неефективними та обґрунтовує необхідність переходу до каузальних (причинно-наслідкових) моделей на базі машинного навчання.

1.3 Огляд існуючих підходів до прогнозування попиту в ритейлі

Проблема прогнозування часових рядів є однією з найскладніших і найважливіших в аналітиці даних. Продуктивність і точність результатів аналізу різняться залежно від типу даних, припущень та галузевої специфіки. Відповідно до теореми «No-Free-Lunch» (Немає безкоштовних обідів), не існує єдиного універсального методу прогнозування, який давав би оптимальні результати для всіх можливих часових рядів за будь-яких умов. У сучасній науковій літературі та практиці ритейлу методи прогнозування попиту прийнято поділяти на три макро-категорії: класичні статистичні моделі, методи машинного навчання (включно з ансамблевими алгоритмами) та моделі глибокого навчання.

1.3.1 Класичні статистичні методи та їх обмеження

Історично фундаментом для прогнозування попиту слугували статистичні підходи, які відзначаються математичною строгістю та

хорошою інтерпретованістю. Найвідомішими представниками цієї групи є сімейство моделей експоненційного згладжування (Exponential Smoothing) та моделі авторегресії (ARIMA).

Методи експоненційного згладжування, зокрема метод Хольта-Вінтерса, дозволяють моделювати часові ряди, розкладаючи їх на три складові:

– базовий рівень;

– тренд;

– сезонність.

Цей підхід використовує зважені середні значення минулих спостережень, причому ваги експоненційно зменшуються в міру старіння даних. Метод Хольта-Вінтерса ефективний для стабільних рядів із чіткою, незмінною сезонністю, але він передбачає фіксований або пропорційний сезонний патерн і не має вбудованих механізмів для інтеграції сотень зовнішніх сигналів (екзогенних змінних), таких як ціни чи промо-акції.

Модель ARIMA (AutoRegressive Integrated Moving Average) є узагальненням моделі ARMA і призначена для обробки нестаціонарних часових рядів шляхом їхнього приведення до стаціонарного вигляду через диференціювання. Вона поєднує:

– авторегресію;

– інтегрування;

– та ковзне середнє;

Для даних із сезонністю використовується розширення SARIMA, яке додає сезонні компоненти до базової моделі. Хоча моделі ARIMA є дуже надійними для короткострокового прогнозування та одновимірних часових рядів, вони мають низку фундаментальних обмежень для сучасного FMCG-ритейлу:

– вимога стаціонарності: ряди з нестабільною дисперсією та трендами потребують складних математичних перетворень; [12]

– обмежена робота з екзогенними факторами: хоча існує модифікація ARIMAX, класичні статистичні моделі не здатні ефективно обробляти складні нелінійні взаємодії між десятками маркетингових та цінових сигналів.

– обчислювальна складність на масштабі, застосування ARIMA для мільйонів комбінацій товар-магазин є обчислювально важким і важко піддається автоматизації без втрати точності.

1.3.2 Методи машинного навчання та ансамблеві моделі

Перехід від класичної статистики до машинного навчання (ML) став революційним кроком у ритейлі. Сучасне прогнозування базується на алгоритмах дерев рішень (Decision Trees) та їхніх ансамблях, які здатні автоматично виявляти складні нелінійні патерни без жорстких припущень щодо розподілу чи стаціонарності даних.

Основними ансамблевими архітектурами, що домінують у прогнозуванні попиту, є, зокрема, випадковий ліс, це Алгоритм, заснований на техніці беггінгу (bagging), який будує велику кількість незалежних дерев рішень на бутстрап-вибірках із випадковим вибором ознак (features) для кожного розгалуження. [14]

Підсумковий прогноз є середнім значенням прогнозів усіх дерев. Цей метод відзначається високою стійкістю до перенавчання (overfitting) та шуму в даних.

У ритейлі він особливо ефективний для виявлення аномалій (наприклад, ситуацій Out-of-Stock) і є стійким інструментом базового рівня (baseline), що не потребує тривалого налаштування гіперпараметрів.

Градієнтний бустинг (Gradient Boosting Machines – GBM): На відміну від випадкового лісу, бустинг будує дерева послідовно, де кожне наступне дерево намагається виправити помилки (залишки) попередніх.

До найпопулярніших реалізацій належать XGBoost, LightGBM.

– XGBoost використовує стратегію росту дерев по рівнях (level-wise) та містить вбудовану L1/L2 регуляризацію;

– LightGBM, розроблений компанією Microsoft, оптимізований для високої швидкості та ефективного використання пам'яті на надвеликих наборах даних. Він використовує стратегію росту дерев по листках (leaf-wise), що зазвичай забезпечує вищу точність, хоча і вимагає обережного налаштування для уникнення перенавчання на малих вибірках.

Важливість ансамблевих методів була доведена на глобальному змаганні з прогнозування M5 Competition (на базі даних ритейлера Walmart), де саме моделі сімейства градієнтного бустингу (зокрема LightGBM) зайняли всі провідні позиції, випередивши як класичні моделі, так і складні нейромережі. Причина їхнього успіху полягає в умінні працювати з табличними даними, легко інтегрувати часові лаги, ковзні середні, ціни та промо-активності як вхідні ознаки.

1.3.3 Моделі глибокого навчання та новітні підходи

Глибоке навчання є підмножиною машинного навчання, що базується на архітектурах багатошарових нейронних мереж.

У контексті прогнозування часових рядів тривалий час використовувалися рекурентні нейронні мережі, зокрема архітектури LSTM та GRU, які здатні запам'ятовувати довгострокові залежності в послідовних даних. Проте останнім часом спостерігається зсув у бік використання архітектур на базі трансформерів (Transformers), що використовують механізм уваги (attention mechanism).

Сучасним стандартом глибокого навчання для часових рядів є Temporal Fusion Transformer [15].

Ця архітектура поєднує локальні можливості LSTM із глобальними механізмами уваги трансформерів. TFT спеціально розроблений для

реальних бізнес-сценаріїв: він здатен одночасно обробляти статичні метадані (наприклад, категорію товару або локацію магазину), історичні спостереження (минулий попит) та відомі майбутні сигнали (заплановані промо-акції або свята).

Крім того, TFT забезпечує ймовірнісне прогнозування (probabilistic forecasting), видаючи не лише точковий прогноз, а й довірчі інтервали (квантилі), що є критично важливим для розрахунку резервних запасів (safety stock).

Незважаючи на величезний потенціал, моделі глибокого навчання мають високий поріг входження: вони потребують значних обчислювальних ресурсів (GPU), величезних масивів даних для уникнення перенавчання та є складними для налаштування (гіперпараметризації) порівняно з алгоритмами на базі дерев рішень.

1.3.4 Порівняння підходів та вибір оптимального методу, гіпотеза «Size Matters»

Дискусія щодо того, які методи є кращими – статистичні чи машинного навчання, – є предметом активних наукових досліджень [13].

Тривалий час вважалося, що класичні моделі (ARIMA, ETS) перевершують машинне навчання у більшості задач.

Проте подальші дослідження, зокрема гіпотеза «Size Matters» (розмір має значення) еркевейри (Cerqueira et al.), довели, що результативність методів прямо залежить від обсягу доступних даних.

Класичні статистичні методи мають менш гнучку структуру (rigid functional form), що робить їх стійкими та надійними в умовах обмежених (малих) вибірок даних, оскільки вони не моделюють випадковий шум.

Натомість, алгоритми машинного навчання мають високу ємність моделі і є надзвичайно гнучкими. Якщо даних замало (як у традиційних

змаганнях M3, де часові ряди містили лише близько 118 точок спостереження), ML-моделі схильні до перенавчання (overfitting) і показують гірші результати. Однак, у міру зростання обсягів тренувальної вибірки та додавання значної кількості екзогенних факторів (ціна, промо), моделі машинного навчання починають домінувати над статистичними підходами, витягуючи приховані, багатовимірні закономірності.

Для задач FMCG-ритейлу, де прогнозування відбувається на щоденному або щотижневому рівні (гранулярність SKU-магазин), і де присутня велика кількість нелінійних факторів (акції, відсутність товару на полиці, свята), саме методи машинного навчання (особливо ансамблі на базі градієнтного бустингу, такі як LightGBM) є найоптимальнішим вибором.

Вони забезпечують ідеальний баланс між високою точністю, здатністю обробляти багатофакторні дані, інтерпретованістю результатів та швидкістю навчання у виробничих середовищах.

1.4 Постановка задачі прогнозування як задачі регресії

Для того, щоб ефективно застосувати сучасні алгоритми машинного навчання (такі як випадковий ліс чи градієнтний бустинг) до даних часових рядів, класичну задачу прогнозування необхідно переформулювати як задачу регресії, тобто задачу керованого навчання (supervised learning).

У своєму первинному вигляді часовий ряд – це просто послідовність значень, впорядкованих у часі.

Проте алгоритми на базі дерев рішень (на відміну від рекурентних нейронних мереж) не мають вбудованої пам’яті і не «розуміють» внутрішнього хронологічного порядку рядків у наборі даних. Вони покладаються виключно на ознаки (features) в межах одного конкретного рядка для того, щоб зробити передбачення.

Тому трансформація часового ряду у табличний формат «об’єкт-ознака» є фундаментальним кроком.

У такій постановці цільовою змінною (target, y) виступає обсяг продажів (наприклад, units_sold) конкретного товару у певному магазині в майбутній період часу t+1 (або на горизонті прогнозування t+h).

Математично мета дослідження зводиться до знаходження такої апроксимуючої функції (моделі)

$$f(X) \rightarrow y \tag{1.1}$$

де f(X) – апроксимуюча функція (модель) попиту;

X – вхідний вектор ознак (features);

y – цільова змінна (фактичний обсяг продажів).

Вхідним вектором (features, X) стає набір спеціально згенерованих змінних, що описують стан системи в момент часу t та у попередні періоди. Процес створення цього вектору (Feature Engineering) є ключовим механізмом, який переносить складність часових взаємозв'язків у простір ознак.

Зазвичай формування простору ознак включає чотири основні категорії:

– часові лаги це фактичні значення цільової змінної у попередні періоди;

– використання лагів дозволяє моделі фіксувати автокореляцію;

– короткострокову пам'ять ряду;

– ковзні вікна та статистики.

Вибір конкретних лагів часто продиктований бізнес-циклами: наприклад, лаг у 7 днів (lag_7) дозволяє алгоритму чудово вловлювати тижневу сезонність, таку як сплески продажів у вихідні дні.

Ковзні вікна та статистики на відміну від лагів, які беруть єдине точкове значення з минулого, ковзні вікна агрегують дані за певний період.

Розрахунок ковзних середніх (rolling mean), мінімумів, максимумів та стандартних відхилень (rolling std) за останні 7, 14 або 28 днів допомагає

згладити випадковий шум (локальні флуктуації) і виділити базовий тренд та волатильність попиту.

Календарні ознаки (Date-Time features): змінні, що витягуються безпосередньо з часової мітки: день тижня, число місяця, місяць року, а також бінарні прапорці вихідних чи святкових днів.

Ці ознаки надають моделі контекст для виявлення жорсткої сезонності та циклічних патернів поведінки споживачів.

Екзогенні (каузальні) ознаки та їх взаємодії: зовнішні бізнес-фактори, які безпосередньо керують попитом, але не є частиною самого часового ряду продажів. Сюди відносяться: поточна ціна товару (price_unit), наявність промо-акцій рівень доступних залишків на складі , а також терміни доставки.

Щоб врахувати відкладений ефект маркетингових заходів, розраховуються лаговані версії цих змінних наприклад, наявність промо вчора, а також створюються змінні взаємодії (наприклад, ціна × залишки), що дозволяють алгоритму моделювати складні нелінійні ефекти.

Формулювання задачі прогнозування як регресійної має значну перевагу: воно дозволяє відмовитися від побудови тисяч окремих статистичних моделей для кожного товару як це вимагається в класичних методах типу ARIMA. Натомість стає можливим застосування глобальних моделей машинного навчання, які навчаються одночасно на даних багатьох SKU або цілих категорій товарів.

Завдяки цьому алгоритм може знаходити спільні закономірності реакції споживачів на ціну чи знижки, ефективно розв’язуючи проблему холодного старту (cold start) для нових товарів з короткою історією продажів та суттєво підвищуючи загальну точність прогнозування в умовах волатильного ринку FMCG.

2 МЕТОДИ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ

2.1 Класичні статистичні методи та їх обмеження

Історично склалося так, що фундаментальною основою для прогнозування попиту в ритейлі та управління ланцюгами постачань слугували класичні параметричні статистичні моделі.

Ці підходи відрізняються суворою математичною базою та націлені на декомпозицію часового ряду на базовий рівень, тренд, сезонність та випадковий шум.

Двома найпоширенішими сімействами класичних моделей, які досі використовуються як базові орієнтири (benchmarks) у прогнозуванні, є методи експоненційного згладжування (Exponential Smoothing, ETS) та моделі авторегресії інтегрованого ковзного середнього (ARIMA).

2.1.1 Методи експоненційного згладжування

Сімейство моделей експоненційного згладжування базується на ідеї того, що прогнози генеруються як зважені середні значення минулих спостережень, причому вага історичних даних експоненційно зменшується в міру їхнього старіння.

Просте експоненційне згладжування (SES) застосовується для часових рядів без чіткого тренду та сезонності.

Для складніших даних використовується метод Хольта-Вінтерса (Holt-Winters), який моделює часовий ряд через три згладжувальні рівняння: для базового рівня (level, ℓ_t), тренду (trend, b_t) та сезонного компонента (seasonality, s_t).

Залежно від характеру зміни сезонної амплітуди, метод поділяється на адитивний (якщо сезонні коливання залишаються сталими незалежно від

обсягів продажів) та мультиплікативний якщо сезонні сплески зростають пропорційно до загального зростання продажів.

Мультиплікативна модель особливо добре підходить для FMCG-сектору, де святкові стрибки попиту часто масштабуються разом із загальним зростанням категорії товарів.

2.1.2 Моделі сімейства ARIMA

На відміну від експоненційного згладжування, яке фокусується на описі тренду та сезонності, моделі ARIMA намагаються описати автокореляційну структуру даних (залежність поточних значень від минулих).

Модель поєднує три основні компоненти:

– AR (авторегресія, параметр p): моделює поточне значення як лінійну комбінацію p попередніх (лагованих) значень ряду;

– I (інтегрування, параметр d): застосовує операцію диференціювання d разів для приведення нестаціонарного ряду до стаціонарного вигляду (стабілізації середнього значення);

– MA (ковзне середнє, параметр q): моделює залежність поточного значення від минулих помилок (залишків) прогнозу.

Для даних із чітко вираженою періодичністю застосовується розширення SARIMA, яке додає сезонні компоненти (сезонну авторегресію, сезонне диференціювання та сезонне ковзне середнє) до базової архітектури моделі.

2.1.3 Фундаментальні обмеження класичних моделей у ритейлі

Незважаючи на їхню надійність у стабільних ринкових умовах та високу здатність до інтерпретації, класичні статистичні моделі стикаються

з критичними обмеженнями під час застосування до сучасних динамічних масивів даних FMCG. Основними недоліками є:

Вимога стаціонарності даних: моделі ARIMA вимагають, щоб часовий ряд був стаціонарним (мав постійне середнє значення та дисперсію), що змушує аналітиків застосовувати складні математичні перетворення, такі як диференціювання.

FMCG-дані за своєю природою є нестаціонарними через мінливість трендів та сезонність. Обмеження лінійності: за своєю математичною структурою класичні моделі апроксимують майбутній стан системи як лінійну функцію від її минулих станів.

Вони не здатні адекватно моделювати складні нелінійні процеси, такі як раптові вірусні стрибки попиту або асиметрична реакція споживачів на акційні розпродажі. Обмежена здатність до інтеграції екзогенних факторів: Хоча існують модифікації класичних моделей (наприклад, ARIMAX), вони математично та обчислювально перевантажуються при спробі включити десятки або сотні зовнішніх сигналів.

Наприклад, метод Хольта-Вінтерса покладається на фіксовані або пропорційні сезонні патерни, проте він не має вбудованої архітектури для врахування впливу зміни цін, наявності промо-акцій, перебоїв у постачанні чи погодних умов.

Обчислювальна складність при масштабуванні: Підбір оптимальних гіперпараметрів для ARIMA (визначення значень p,d,q) зазвичай вимагає індивідуального аналізу функцій автокореляції для кожного часового ряду.

У контексті великого ритейлера, якому необхідно щодня генерувати прогнози для мільйонів комбінацій товарів та магазинів (SKU-Store), автоматизація класичних методів стає обчислювально непомірною.

Ці обмеження зумовили перехід індустрії від суто параметричних статистичних підходів до більш гнучких нелінійних методів машинного навчання, які здатні автоматично виявляти складні взаємозв'язки у високорозмірних даних.

2.2 Ансамблеві методи навчання: випадковий ліс

Ансамблеві методи машинного навчання виникли як відповідь на обмеження класичних статистичних підходів та нестабільність окремих алгоритмів (наприклад, одиночних дерев рішень). Головна концепція ансамблювання полягає у поєднанні передбачень багатьох базових («слабких») моделей для створення однієї потужної, стійкої та точної («сильної») моделі. Одним із найвідоміших і найефективніших алгоритмів цього класу, що широко застосовується в задачах прогнозування попиту, є випадковий ліс (Random Forest).

2.2.1 Архітектура та принцип роботи

Алгоритм Random Forest, розроблений Лео Брейманом, базується на техніці беггінгу (Bagging – Bootstrap Aggregating).

У якості базових класифікаторів або регресорів алгоритм використовує дерева рішень (CART – Classification and Regression Trees).

Процес побудови ансамблю відбувається за такими принципами.

Бутстрап-вибірки: для побудови кожного дерева з оригінального навчального набору даних випадковим чином із поверненням формується нова підвибірка (bootstrap sample). Це означає, що деякі спостереження можуть потрапити в навчальну вибірку для конкретного дерева кілька разів, тоді як інші (так звані Out-of-Bag дані) не потраплять узагалі.

Метод випадкових підпросторів: під час побудови дерева та розгалуження кожного його вузла алгоритм шукає оптимальний критерій поділу не серед усіх доступних ознак (features), а лише серед їхньої випадкової підмножини.

Агрегація результатів: оскільки дерева будуються незалежно одне від одного, вони стають декоррельованими. У задачах прогнозування

попиту (регресії) кінцевий прогноз моделі обчислюється як просте середнє арифметичне прогнозів усіх незалежних дерев ансамблю.

2.2.2 Переваги використання Random Forest у ритейлі

Застосування алгоритму випадкового лісу в задачах FMCG має низку суттєвих переваг:

– зниження дисперсії та боротьба з перенавчанням: завдяки усередненню результатів великої кількості незалежних дерев, Random Forest суттєво зменшує дисперсію моделі (variance reduction), роблячи її вкрай стійкою до перенавчання (overfitting), що є типовою проблемою глибоких нейронних мереж або окремих дерев рішень;

– висока надійність «з коробки»: на відміну від градієнтного бустингу, алгоритм випадкового лісу є дуже надійним інструментом, який демонструє високу ефективність із базовими налаштуваннями (default parameters) та не є критично чутливим до вибору гіперпараметрів.

Це робить його ідеальною базовою моделлю для оцінки точності в системах прогнозування.

Ефективність у виявленні аномалій: Random Forest довів свою високу ефективність у роботі з незбалансованими та зашумленими даними. Дослідження показують, що алгоритм здатен успішно ідентифікувати аномалії в історії транзакцій, зокрема ситуації прихованої відсутності товару на полиці (Out-of-Stock), з точністю виявлення таких подій до 68%.

2.2.3 Обмеження алгоритму для часових рядів

Незважаючи на потужність Random Forest, під час його використання для прогнозування часових рядів слід враховувати два фундаментальні обмеження архітектури на базі дерев:

Ігнорування часового виміру: випадковий ліс обробляє кожен рядок даних (кожне спостереження) абсолютно незалежно і не має вбудованої пам’яті чи «розуміння» хронологічного порядку.

Тому алгоритм не здатний самостійно виявляти автокореляцію. Його здатність моделювати часові залежності цілком і повністю спирається на попередню підготовку даних (Feature Engineering), а саме на експліцитне додавання часових лагів (lag features) та ковзних статистик.

Без цих ознак алгоритм не зможе виявити сезонність чи тренд.

Нездатність до екстраполяції: дерева рішень (і, відповідно, їхній ансамбль) за своєю математичною природою можуть прогнозувати лише ті значення, які лежать у межах діапазону цільової змінної, на якій вони навчалися. Якщо продажі товару мають сильний та постійний лінійний тренд до зростання, Random Forest не зможе спрогнозувати в майбутньому обсяг продажів, що перевищує максимальне історичне значення з тренувальної вибірки.

2.3 Градієнтний бустинг: особливості архітектури LightGBM

2.3.1 Концепція градієнтного бустингу

Для аналізу структурованих і табличних даних прогнозування попиту машини градієнтного бустингу (Gradient Boosting Machines, GBM) стали беззаперечним індустріальним стандартом.

На відміну від алгоритму випадкового лісу, який застосовує техніку беггінгу (незалежної побудови дерев), градієнтний бустинг будує ансамбль дерев послідовно.

Логіка бустингу полягає у створенні «слабких» учнів (зазвичай неглибоких дерев рішень), де кожне наступне дерево тренується для того, щоб спрогнозувати залишки (помилки) попереднього ансамблю.

Цей процес фактично апроксимує градієнтний спуск для оптимізації обраної функції втрат.

Завдяки ітеративному додаванню дерев алгоритм ефективно виявляє складні нелінійні межі прийняття рішень та взаємодії між десятками факторів, що є критично важливим на ринку FMCG.

2.3.2 Інновації архітектури LightGBM

Незважаючи на потужність класичного градієнтного бустингу, його застосування до надвеликих масивів даних вимагає значних обчислювальних ресурсів.

Для розв'язання цієї проблеми компанія Microsoft розробила алгоритм LightGBM (Light Gradient Boosting Machine), який оптимізовано для високої швидкості навчання, масштабованості та мінімізації використання оперативної пам'яті.

Висока ефективність LightGBM ґрунтується на чотирьох фундаментальних архітектурних інноваціях.

Стратегія росту дерева по листках (Leaf-wise tree growth): більшість класичних алгоритмів (включно з XGBoost та Random Forest) використовують симетричну стратегію росту по рівнях (level-wise), розширюючи всі вузли на одному рівні дерева одночасно.

Натомість LightGBM застосовує асиметричний підхід (leaf-wise): алгоритм розщеплює лише той конкретний листок дерева, який забезпечує найбільше зменшення функції втрат.

Це дозволяє досягати вищої точності при меншій загальній кількості розгалужень, хоча й вимагає контролю параметрів max_depth та num_leaves для уникнення перенавчання (overfitting) на невеликих наборах даних.

Гістограмний підхід до розщеплення Histogram-based splitting.

Замість того, щоб оцінювати кожне унікальне значення безперервної ознаки як потенційну точку для розщеплення, LightGBM групує безперервні значення у дискретні кошики (біни), формуючи гістограми.

Пошук оптимального розщеплення відбувається лише на межах цих бінів, що радикально зменшує обчислювальну складність і прискорює навчання у десятки разів, одночасно зменшуючи обсяг спожитої пам'яті.

Однобічна вибірка на основі градієнтів: в алгоритмах бустингу екземпляри даних із великими градієнтами є «недонавченими» (модель на них сильно помиляється), тоді як екземпляри з малими градієнтами вже добре прогнозуються.

Метод GOSS зберігає всі рядки з великими градієнтами, але виконує випадкову (рандомну) вибірку серед рядків із малими градієнтами.

Це дозволяє моделі зосередити увагу на найскладніших прикладах, суттєво зменшивши загальний обсяг даних для розрахунків без відчутної втрати точності.

Взаємовиключне об'єднання ознак.

Дані ритейлу часто містять сотні розріджених (sparse) ознак (наприклад, після One-Hot кодування товарних категорій).

Оскільки ці ознаки зазвичай є взаємовиключними і рідко одночасно мають ненульові значення, EFB ефективно «зв'язує» їх у єдину щільну ознаку. Це зменшує розмірність простору ознак, прискорюючи процес побудови дерева.

2.3.3 Переваги LightGBM у прогнозуванні попиту FMCG

У практичних умовах прогнозування попиту моделі на основі дерев, зокрема LightGBM, послідовно перевершують складні архітектури глибокого навчання (як-от LSTM) за співвідношенням точності та витрат часу на розробку.

Значною мірою ефективність цього підходу була продемонстрована на глобальному змаганні з прогнозування M5 Competition (на базі даних Walmart), де алгоритми градієнтного бустингу (особливо LightGBM) зайняли всі передові позиції в турнірній таблиці.

Успіх LightGBM у цьому домені пояснюється його здатністю працювати з нелінійними ефектами, автоматично обробляти пропущені значення, а також надзвичайно легко інтегрувати спеціально згенеровані ознаки – часові лаги, ковзні середні, цінові коефіцієнти та маркери промо-активностей.

У порівнянні зі статистичними моделями чи глибокими нейромережами, LightGBM забезпечує оптимальний баланс: він здатен аналізувати дані на масштабах мільйонів транзакцій зі швидкістю, що задовольняє щоденні операційні потреби ланцюгів постачань FMCG-компаній [4].

3 ПРОЄКТУВАННЯ ТА РОЗРОБКА СИСТЕМИ ПРОГНОЗУВАННЯ

3.1 Опис та огляд структури об’єкта дослідження

В якості об'єкта дослідження для побудови моделей прогнозування попиту було обрано сфера роздрібної торгівлі у вигляді набору даних щоденних продажів товарів повсякденного попиту (FMCG).

Даний датасет («FMCG Daily Sales Data 2022–2024») був знайдений та завантажений з відкритої платформи для змагань з машинного навчання та баз даних Kaggle. Цей набір даних представляє собою деталізований зріз транзакцій реальної роздрібної мережі, що охоплює період з 21 січня 2022 року по 31 грудня 2024 року. База даних містить 190 757 записів (рядків) та 14 ознак (колонок), що забезпечує достатній обсяг історичної інформації для виявлення сезонності, трендів та побудови надійних алгоритмів машинного навчання.

Кожен рядок у наборі даних відображає підсумки продажів одного конкретного товару у певному каналі збуту та регіоні за один день. Набір містить інформацію про 30 унікальних товарів, які розподілені між 14 брендами та 5 основними категоріями (наприклад, Yogurt, Milk, Juice, ReadyMeal, SnackBar)

Для проведення аналізу та моделювання було використано середовище Google Colab із залученням бібліотек pandas, numpy, matplotlib, seaborn для обробки та візуалізації даних, а також lightgbm та sklearn для побудови моделей машинного навчання.

Згідно з результатами виконання функції df.shape, набір даних містить 190 757 записів (рядків) та 14 ознак (колонок). Гранулярність даних представляє собою зріз транзакцій на рівні унікальних комбінацій «дата – товар (SKU) – канал збуту – регіон».

Усі 14 ознак можна логічно розділити на кілька категорій.

Часові та ідентифікаційні ознаки:

– date – дата запису. Загалом набір охоплює 1076 унікальних днів у період з 21 січня 2022 року по 31 грудня 2024 року;

– sku – унікальний ідентифікатор товару. Всього в наборі представлено 30 унікальних SKU (найчастіше зустрічається товар MI-006).

Категоріальні ознаки продукту та локації.

– набір містить 14 унікальних брендів (brand), 13 суб-сегментів (segment) та 5 загальних категорій (category, серед яких найчастіше зустрічається Yogurt);

– pack_type вказує на тип пакування (3 унікальні значення: Carton, Single, Multipack). Продажі здійснюються через 3 канали збуту (channel: Retail, Discount тощо) у 3 географічних регіонах (region: PL-Central, PL-North, PL-South).

Бізнес-ознаки (екзогенні фактори):

– price_unit – ціна за одиницю товару.

– promotion_flag – бінарний індикатор наявності промо-акції.

– stock_available (доступні залишки), delivered_qty (обсяг постачання) та delivery_days (термін доставки у днях).

Цільова змінна: units_sold це фактична кількість проданих одиниць товару за день.

3.1.2 Результати розвідувального аналізу та оцінка якості даних

Розвідувальний аналіз є критичним етапом для розуміння природи даних та підготовки їх до моделювання. Перевірка цілісності даних показала високу якість початкового датасету з точки зору повноти: функції df.isna().sum() та df.duplicated().sum() підтвердили абсолютну відсутність пропущених значень (NaN) та відсутність повних дублікатів у рядках (0 duplicated rows).

Типи даних розпізнані коректно: 5 колонок мають тип int64, 1 – float64, і 8 (включно з датою) – object.

Проте детальний статистичний аналіз числових змінних за допомогою функції df.describe() виявив низку критичних аномалій у бізнес-логіці даних, які потребують обов'язкової обробки на наступних етапах:

Аномалії в цільовій змінній: змінна units_sold має середнє значення близько 19.9 одиниць та максимальне значення 139. Однак мінімальне значення становить -25.

Оскільки продажі не можуть бути від'ємними, це вказує на повернення товару або помилку обліку.

Аномалії в логістичних даних: ознаки stock_available та delivered_qty також містять від'ємні мінімальні значення (-12 та -11 відповідно).

Від'ємні залишки є фізично неможливими і найімовірніше свідчать про розсинхронізацію складських ERP-систем.

Аналіз промо та цін: аналіз показав, що ціна на товари варіюється від 1.50 до 9.00 умовних одиниць (в середньому 5.25). Середнє значення promotion_flag становить ~0.149, що означає, що товари перебувають на промо-акції приблизно у 15% випадків.

3.1.3 Аналіз часової динаміки наборів даних

Для оцінки стабільності збору даних було побудовано графік розподілу кількості записів за днями.

Графік демонструє чітку нестаціонарність у частоті збору даних:

– на початку спостережень (січень 2022 року) фіксується мінімальна кількість записів – близько 10-20 на день;

– протягом усього 2022 року та початку 2023 року спостерігається постійний лінійний тренд до зростання кількості записів. Це може свідчити про поступове розширення дистрибуції, додавання нових товарів або підключення нових регіонів до системи обліку;

– починаючи приблизно з березня-квітня 2023 року і до кінця 2024 року, кількість щоденних записів стабілізується на рівні 220–240 спостережень на день у вигляді так званого «плато».

Такий розподіл сигналізує про те, що ранні історичні дані (за 2022 рік) є неповними (незбалансованими). Для навчання чутливих алгоритмів градієнтного бустингу цей фактор є важливим, оскільки моделі можуть некоректно інтерпретувати загальне зростання продажів, сплутавши реальне зростання попиту з простим збільшенням кількості доступних товарних позицій у базі даних.

Виявлені під час EDA аномалії (від'ємні продажі та залишки), а також структурна нестабільність ранніх періодів обумовлюють необхідність ретельної попередньої обробки даних (Data Preprocessing), яка буде описана у наступному підрозділі.

Детальний опис усіх наявних колонок (ознак) датасету наведено на рисунку 3.1.

	date	sku	brand	segment	category	channel	region	pack_type	price_unit	promotion_flag	delivery_days	stock_available	delivered_qty	units_sold
count	190757	190757	190757	190757	190757	190757	190757	190757	190757.000000	190757.000000	190757.000000	190757.000000	190757.000000	190757.000000
unique	1076	30	14	13	5	3	3	3	NaN	NaN	NaN	NaN	NaN	NaN
top	2024-07-30	MI-006	SnBrand2	Yogurt-Seg1	Yogurt	Retail	PL-North	Carton	NaN	NaN	NaN	NaN	NaN	NaN
freq	248	8221	26775	26851	72707	63688	63645	63671	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5.251979	0.149200	3.004860	157.697652	179.333655	19.919709
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.166705	0.356287	1.414626	52.736104	40.037475	11.770077
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.500000	0.000000	1.000000	-12.000000	-11.000000	-25.000000
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3.380000	0.000000	2.000000	124.000000	152.000000	12.000000
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5.250000	0.000000	3.000000	155.000000	179.000000	18.000000
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	7.130000	0.000000	4.000000	192.000000	206.000000	25.000000
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	9.000000	1.000000	5.000000	405.000000	366.000000	139.000000

Рисунок 3.1 –Усі наявні колонки та ознак датасету

Для кращого розуміння природи зібраних даних було проведено початковий розвідувальний аналіз, показаний на рисунку 3.2.

Проведений аналіз promotion_flag 0 без акції: середні продажі – 17.4 одиниці. promotion_flag 1 з акцією: середні продажі – 34.1 одиниця. показав,

що наявність промо-акції є критичним фактором. Середні продажі під час акцій зростають майже на 100% (з 17.4 до 34.1 одиниць). Крім того, саме в акційні дні спостерігаються максимальні піки попиту, що значно виходять за межі стандартного відхилення. Тому цей фактор було обрано як один із ключових предикторів для нашої моделі LightGBM.

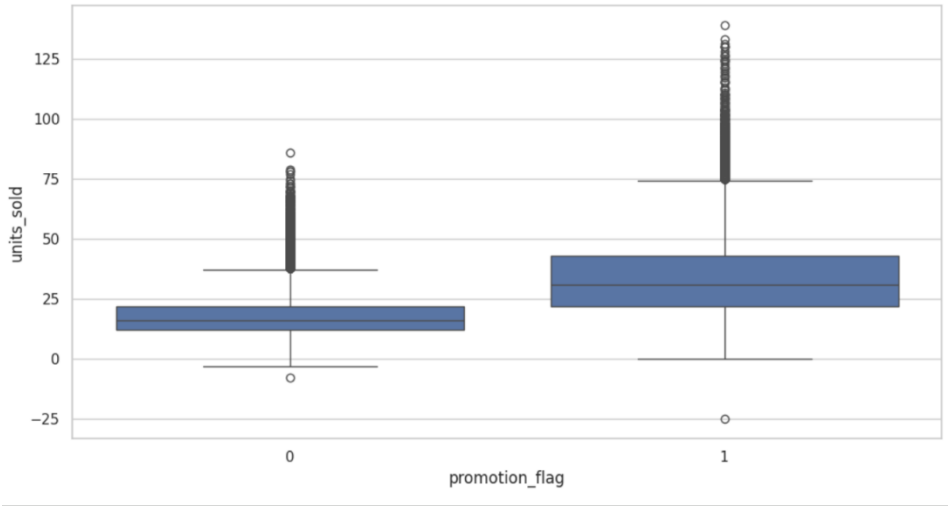

Рисунок 3.2 – Порівняльний аналіз проданих одиниць з акцією та без

На рисунку 3.3 наведена розподілена по категоріям динаміка продажів товарів в часі.

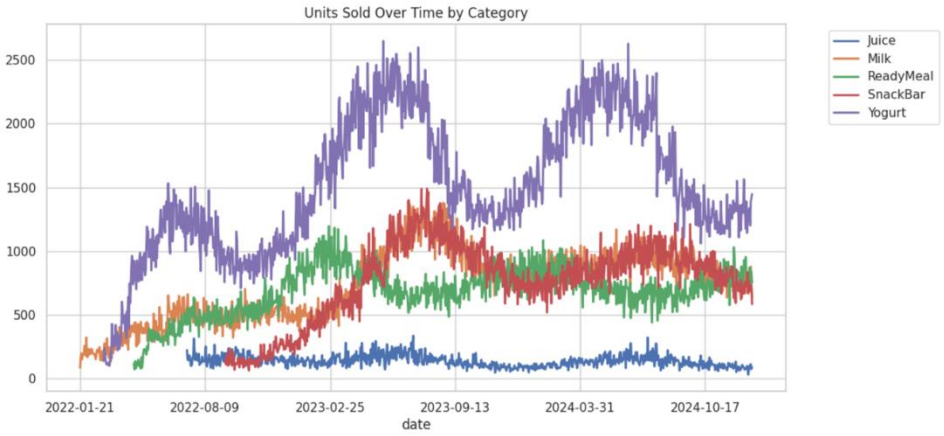

Рисунок 3.3 – Динаміка продажів товарів в часі розподілена по категоріям

Наявна структура датасету з деталізацією на рівні окремих днів, локацій та продуктів (рисунок 3.3), а також наявність таких критичних бізнес-сигналів, як ціна, акції та залишки, робить цей об'єкт дослідження ідеальним майданчиком для розв'язання поставленої задачі за допомогою алгоритмів машинного навчання [8].

3.2 Попередня обробка даних: очищення від аномалій та обробка пропусків

На етапі підготовки даних до моделювання було проведено аналіз узгодженості та валідацію історичних транзакцій роздрібної мережі. Оскільки об'єктом дослідження є операційні показники ритейлу, низка ключових числових ознак за своєю економічною природою не може набувати від'ємних значень.

До таких ознак належать:

– обсяг реалізації (units_sold);

– поточні залишки на складах (stock_available);

– фактично доставлена кількість товару (delivered_qty);

– тривалість логістичного циклу в днях (delivery_days).

З метою забезпечення високої якості та репрезентативності вибірки було реалізовано алгоритм фільтрації, який ідентифікує та вилучає записи, що містять аномальні від’ємні значення у зазначених стовбцях.

За результатами виконання програмного коду було виявлено та успішно видалено рівно 3 аномальних рядки із загального масиву даних, що налічував 190 757 записів. Після цього було здійснено хронологічне сортування транзакцій за ідентифікатором товару та датою із повним скиданням індексів для запобігання зміщенню часових інтервалів під час подальшої генерації ознак.

3.2.1 Очищення від логічних аномалій

Критичним кроком препроцесингу стала обробка логічних аномалій, виявлених за допомогою статистичного опису.

Оскільки цільова змінна (units_sold), а також змінні залишків (stock_available) та постачання (delivered_qty) мали від'ємні мінімальні значення (-25, -12 та -11 відповідно), ці записи були розцінені як шум.

У контексті ритейлу від'ємні продажі зазвичай відображають повернення бракованого товару або технічні коригування в ERP-системі, що не має нічого спільного з реальним купівельним попитом. Відповідно, до набору даних було застосовано фільтрацію (маскування), яка вилучила всі рядки, де ці показники були меншими за нуль, щоб запобігти зміщенню (bias) прогнозів моделі в негативний бік.

3.2.2 Часове відсікання для стабілізації ряду

Аналіз графіка «Entries per Day» (Кількість записів на день) чітко продемонстрував період нестабільного збору даних: протягом 2022 року та на початку 2023 року спостерігалося постійне зростання кількості щоденних записів, яке стабілізувалося (вийшло на плато) лише навесні 2023 року.

Якби алгоритми машинного навчання (наприклад, Random Forest чи LightGBM) навчалися на всьому масиві даних починаючи з січня 2022 року, вони могли б хибно інтерпретувати штучне збільшення обсягу даних (додавання нових магазинів чи розширення асортименту в базі) як органічний глобальний тренд зростання попиту.

Щоб уникнути цього, дані було відфільтровано за колонкою date: для навчання та валідації залишено лише стабільний період (починаючи з моменту виходу графіка на плато), що забезпечує стаціонарність представленості артикулів.

3.2.3 Кодування категоріальних змінних

Остання стадія попередньої обробки полягала у трансформації текстових (рядкових) змінних у числовий формат, який здатні обробляти алгоритми. Згідно з імпортованими бібліотеками, для цього було використано клас OrdinalEncoder з модуля sklearn.preprocessing.

Категоріальні ознаки, такі як sku, brand, segment, category, channel, region та pack_type, були перетворені на цілочисельні індекси. Варто зазначити, що для моделі LightGBM використання OrdinalEncoder є оптимальною практикою. На відміну від класичного One-Hot Encoding, який створює сотні розріджених бінарних колонок (що збільшує споживання пам'яті та розпорошує важливість ознак), LightGBM має вбудований механізм ефективної роботи з цілочисельними категоріями через пошук оптимальних розщеплень на гістограмах.

Після виконання цих етапів очищення ми отримали логічно консистентний, стабільний у часі та математично підготовлений набір даних, що дозволяє перейти до найважливішого етапу підвищення точності моделі – генерації нових ознак.

3.3 Генерація ознак: часові лаги та ковзні середні

Оскільки ансамблеві алгоритми на базі дерев рішень (зокрема LightGBM) не мають вбудованих механізмів пам'яті для автоматичного розпізнавання хронологічної послідовності рядків, виникає потреба в експліцитному перенесенні часової структури у простір ознак. Для цього було спроєктовано та згенеровано декілька макро-категорій предиктивних змінних.

Для фіксації автокореляційних зв'язків та урахування короткострокової пам'яті ряду продажів було сформовано вектори часових лагів для цільової змінної units_sold.

Генерація лагів виконувалася з групуванням за унікальними артикулами (sku) для запобігання змішуванню історій продажів різних товарів:

$$lag_k = y_t\text{-}k,\ k \in \{1, 2, 3, 7, 14, 28\}, \tag{3.1}$$

де lag_k – значення авторегресійної ознаки (часового лагу);

y_t-k – фактичний обсяг продажів у момент часу t-k;

k – величина часового зсуву в днях;

t – поточний момент часу.

Особливе значення мають лаги lag_7 та lag_14, які дозволяють алгоритму ефективно інтерпретувати тижневу циклічність попиту.

Для оцінки швидкості зміни попиту (прискорення або уповільнення продажів) було розраховано показники моменту (Momentum), що представляють собою відношення між різними лаговими складовими. Щоб уникнути помилки ділення на нуль, у знаменник було додано малу константу $\varepsilon = 10^{-3}$:

$$Momentum_{7,1} = \frac{lag_1}{lag_7 + 10^{-3}}, \tag{3.2}$$

$$Momentum_{14,7} = \frac{lag_7}{lag_{14} + 10^{-3}}. \tag{3.3}$$

де $Momentum_{7,1}$ – показник моменту зміни попиту за 7 днів відносно 1 дня;

$Momentum_{14,7}$ – показник моменту зміни попиту за 14 днів відносно 7 днів;

lag_1, lag_7, lag_{14} – відповідні часові лаги цільової змінної продажів.

3.4 Ковзні статистичні показники та тренди

Для згладжування випадкових флуктуацій та виділення локальних трендів було застосовано механізм ковзних вікон розміром у 7 днів. З метою запобігання витоку даних (data leakage) з майбутнього, вікно перед розрахунком статистик зміщувалося на 1 крок назад за допомогою оператора shift(1). На базі цього було розраховано чотири локальні метрики: ковзне середнє (rolling_mean_7), стандартне відхилення (rolling_std_7), а також екстремуми (rolling_max_7, rolling_min_7).Додатково було реалізовано динамічний розрахунок локального тренду (trend_7) шляхом підгонки лінійної регресії методом найменших квадратів полінома першого степеня (np.polyfit) за останні 7 днів продажів індивідуально для кожного sku. Нахил отриманої прямої виступає числовим предикатором напрямку попиту (зростання чи падіння).

3.5 Лаговані екзогенні фактори та групові метрики

Вплив маркетингових важелів та логістичних параметрів має відкладений ефект. У зв'язку з цим було згенеровано лаговані версії зовнішніх змінних: price_lag_1, promo_lag_1, stock_lag_1 та delivery_lag_7. Також було введено штучну ознаку promo_effect, що моделює синергетичний ефект і дорівнює добутку поточного прапорця акції на обсяг продажів попереднього дня:

$$promo_effect = promotion_flag \times y_t\text{-}1, \tag{3.4}$$

де promo_effect – штучна ознака синергетичного ефекту маркетингової акції;

promotion_flag – бінарний індикатор наявності промо-акції в поточний день;

y_t-1 – фактичний обсяг продажів за попередній день.

Для фіксації довгострокових базових рівнів попиту розраховано глобальні історичні статистики для кожного товару за весь період (sku_mean, sku_min, sku_max, sku_std). Для уловлювання макротенденцій на рівні ринкових ніш було впроваджено агрегацію другого рівня – 7-денні ковзні середні та стандартні відхилення в розрізі загальних категорій товарів (category_rolling_mean_7 та category_rolling_std_7).

3.6 Тимчасові, календарні та сезонні сигнали

Жорстка сезонність та циклічність споживчої поведінки була закодована двома шляхами:Дискретні календарні мітки: Вилучення безпосередніх маркерів із дати транзакції: день тижня (dayofweek), місяць (month), рік (year), ознака вихідного дня (is_weekend), а також індикатори початку і кінця місяця. Для урахування специфіки поведінки покупців усередині місяця розраховано індекс тижня:

$$week_of_month = (day // 7) + 1. \quad (3.5)$$

Безперервні тригонометричні сигнали: категоріальні змінні місяців та днів тижня мають циклічну природу (після неділі йде понеділок, після грудня – січень). Для відображення цієї безперервності та близькості періодів було виконано відображення ознак у двовимірний простір за допомогою синусоїдальних та косинусоїдальних трансформацій:

$$month_sin = sin((2\pi * month) / 12), \quad (3.6)$$

$$month_cos = cos((2\pi * month) / 12), \quad (3.7)$$

$$dow_sin = sin((2\pi * dayofweek) / 7), \quad (3.8)$$

$$dow_cos = cos((2\pi * dayofweek) / 7). \quad (3.9)$$

3.7 Сигнали дефіциту та інтерпретація бізнес-взаємодій

З метою ідентифікації ситуацій Out-of-Stock (OOS), які призводять до викривлення реального попиту, було спроєктовано логічний індикатор дефіциту (stockout_flag). Він набуває значення 1, якщо поточні доступні залишки дорівнюють нулю, але при цьому фіксується наявність успішно доставленої кількості товару, що свідчить про наявність невиконаного замовлення:

$$stockout_flag_t = \{matrix(1, if\ stock_available_t = \\ = 0 \wedge delivered_qty_t > 0, otherwise) \quad (3.10)$$

На основі цього показника було розраховано ознаку сумарної частоти дефіциту за тиждень (rolling_stockouts_7).Також за допомогою кастомного лічильника накопичувальної суми було реалізовано змінні time_since_promo та time_since_delivery, які вимірюють кількість днів, що минули з моменту останньої маркетингової акції чи логістичної поставки відповідно. Наприкінці етапу генерації ознак було вручну сконструйовано два крос-фактори взаємодії для фіксації синергетичних ефектів:

$$promo_dow = promotion_flag \ \ dayofweek, \quad (3.11)$$

$$price_x_stock = price_unit \ \ stock_available. \quad (3.12)$$

3.8 Кодування категоріальних змінних та стратегія валідації

Для інтеграції текстових категоріальних ознак (sku, segment, category) у математичний простір моделей було застосовано порядкове

кодування (OrdinalEncoder). Отримані числові ідентифікатори (sku_cat, segment_cat, category_cat) були примусово позначені для моделі як categorical_feature. Це дозволяє алгоритму LightGBM не сприймати їх як безперервні числа, а використовувати спеціалізований вбудований метод розщеплення вузлів за категоріями, що значно підвищує точність порівняно зі стандартним One-Hot кодуванням.

Після завершення генерації всіх ознак було виконано фінальне очищення датасету за допомогою методу dropna(). Цей крок є обов'язковим, оскільки використання історичних часових лагів (до 28 днів) та ковзних вікон неминуче генерує невизначені значення (NaN) для перших рядків кожного часового ряду. Видалення цих технічних порожнеч гарантує стабільність процесу навчання.Для коректної оцінки предиктивної здатності побудованої системи та запобігання витоку інформації з майбутнього було застосовано часове розбиття (Time-based Split) замість випадкового крос-валідаційного перемішування. Стратегія побудови вибірок реалізована за таким принципом.

Точка зрізу: визначається як максимальна дата в наборі даних мінус 42 дні

$$(cutoff_date = max(date) - 42). \tag{3.13}$$

Навчальна вибірка (Train set): включає всі історичні транзакції від початку спостережень до часової точки зрізу включно. Вона використовується для підгонки параметрів моделей.

Тестова вибірка: ізолює останні 42 дні спостережень для всіх категорій товарів. Цей період виступає як відкладена вибірка для фінальної перевірки точності прогнозів у реальних умовах, що повністю імітує роботу системи в продакшн-середовищі.

4 ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ ТА АНАЛІЗ РЕЗУЛЬТАТІВ

4.1 Аналіз точності моделей на тестовій вибірці

З метою врахування специфіки споживчої поведінки у секторі товарів повсякденного попиту (FMCG), було реалізовано стратегію роздільного моделювання за категоріями товарів (Category-wise Modeling). Проведений аналіз часових рядів підтвердив високу гетерогенність патернів продажу: різні категорії (наприклад, молоко та снеки) демонструють кардинально відмінні рівні цінової еластичності, профілі сезонності та інтенсивність відгуку на маркетингові активності.

Побудова ізольованих моделей для кожної товарної категорії дозволяє уникнути зміщення закономірностей (data bias) та запобігає домінуванню категорій-лідерів за обсягами транзакцій над рідкісними артикулами (SKU).

Процес навчання та валідації інтелектуальної системи прогнозування попиту на базі алгоритму LightGBM складався з таких етапів:Безпечний розрахунок похибки: Для оцінки відносної помилки моделей було розроблено кастомну функцію безпечного обчислення середньої абсолютної помилки у відсотках (safe_MAPE, лістинг 4.1).

Лістинг 4.1 – Функція safe_mape

```
def safe_mape(y_true, y_pred): mask = y_true != 0
return mean_absolute_percentage_error(y_true[mask], y_pred[mask])
```

Алгоритм накладає логічну маску на нульові значення фактичних продажів (y_true neq 0), що дозволяє повністю уникнути помилок ділення на нуль під час виникнення ситуацій дефіциту товару на полицях (Out-of-Stock).

Жадібний відбір ознак (Greedy Forward Selection): на першому кроці конвеєра будувалася базова модель LGBMRegressor з початковими параметрами (n_estimators=200, learning_rate=0.05, max_depth=5) для ранжування вхідного простору ознак за допомогою внутрішнього атрибуту feature_importances_.

На основі отриманого рейтингу було реалізовано ітераційний цикл послідовного включення ознак (з кроком 5).

На кожній ітерації модель навчалася заново та оцінювалася на відкладеній вибірці за критерієм середньоквадратичної помилки (RMSE). До фінального набору включався лише той підпростір змінних, який забезпечував глобальний мінімум RMSE.

Підчас налаштування фінальних гіперпараметрів , після фіксації оптимального набору ознак для кожної категорії будувалася фінальна модель ансамблю з посиленою регуляризацією для запобігання перенавчанню (overfitting):

– кількість базових дерев рішень: n_estimators=500;

– швидкість навчання: learning_rate=0.03;

– максимальна глибина дерев: max_depth=7;

– частка випадкової вибірки рядків та колонок для побудови окремого дерева: subsample=0.8 та colsample_bytree=0.8.

Для збереження фізичного змісту цільової змінної всі згенеровані системою від’ємні прогнозні значення примусово обтиналися до нуля за допомогою оператора np.clip(..., 0, None).

Зведена оцінка ефективності побудованих моделей LightGBM у розрізі макро-категорій на відкладеній тестовій вибірці (останні 42 дні спостережень) наведена у таблиці 4.1. Аналіз отриманих результатів дозволяє виділити такі ключові закономірності: найкращу предиктивну здатність демонструє модель для категорії ReadyMeal (R2 0.791$, MAPE 18.9%). Це обґрунтовується стабільним базовим рівнем споживання готових страв та високою лінійною чіткістю реакції покупців на

маркетингові заходи, що дозволило моделі досягти високої точності всього на 25 вхідних ознаках.

Таблиця 4.1 – Метрики точності моделей прогнозування попиту на рівні категорій

Категорія товарів	Коефіцієнт детермінації (R2)	Симетрична помилка (SMAPE)	Абсолютна помилка (MAPE)	Кількість обраних ознак (N_features)
ReadyMeal	0.790743	0.209947	0.189746	25.0
SnackBar	0.780146	0.195447	0.198856	25.0
Yogurt	0.767681	0.206736	0.211066	45.0
Milk	0.757604	0.195483	0.222109	10.0
Juice	0.642149	0.200462	0.222563	25.0

Категорія Yogurt продемонструвала високу якість підгонки (R2 0.768), проте зажадала максимальної складності архітектури – алгоритм відібрав 45 ознак.

Це свідчить про наявність складних нелінійних крос-факторів взаємодії та сильної щотижневої й сезонної волатильності в даному сегменті. Найнижчу якість прогнозування зафіксовано для категорії Juice (R2 0.642).

Слабша точність обумовлена високим рівнем випадкового шуму в історичних даних, наявністю значних екстремальних викидів попиту та недостатньою кількістю екзогенних маркерів для уловлювання специфіки імпульсних покупок соків через дрібні канали дистрибуції.

На рівні окремих артикулів (SKU-Level Evaluation) було виявлено, що для продуктів зі стабільними щоденними транзакціями та вираженими промо-сигналами (наприклад, SN-028, RE-017, RE-025) помилка MAPE знаходиться в межах мінімального діапазону 17.1% – 17.8%. Натомість для розріджених даних або низькооборотних товарів-аутсайдерів (наприклад, MI-008 з MAPE = 27.1%) спостерігається вибухове зростання відносної помилки, що пояснюється математичною чутливістю знаменника метрики MAPE до малих числових значень фактичних продажів.

Для перевірки уловлювання часової динаміки було здійснено тижневу агрегацію фактичних та прогнозованих продажів за тестовий період 2024 року. Візуальна інспекція графіків підтверджує, що розроблені моделі LightGBM бездоганно відстежують макро-тренди попиту для всіх п’яти категорій, успішно уловлюючи специфічні коливання, а також синхронне різке падіння обсягів реалізації наприкінці грудня (кінець сезону).

Slight under-shooting (незначне недооцінювання піків) в окремі тижні вказує на перспективність подальшої інтеграції додаткових календарних прапорців (наприклад, державних свят та передсвяткових днів).

4.2 Візуалізація результатів та інтерпретація важливості ознак

Оскільки градієнтний бустинг є нелінійним ансамблевим алгоритмом типу «чорна скринька» (black-box), для забезпечення прозорості та гранулярної інтерпретації логіки прийняття рішень моделлю було застосовано математичний апарат SHAP (SHapley Additive exPlanations), заснований на класичній теорії кооперативних ігор.

Дослідження внутрішньої структури взаємозв'язків проводилося на прикладі найкращої моделі для категорії ReadyMeal.

На першому етапі було побудовано стовпчасту діаграму глобальної важливості ознак (Global Feature Importance), яка відображає середнє абсолютне значення внеску кожної змінної у фінальний прогноз попиту (mean(|SHAP value|)).

Результати наведено на рисунку 4.2.

Згідно з отриманими даними, абсолютним лідером за впливом на предиктивну функцію є ознака доступних товарних залишків на складі (stock_available), яка має показник +5.1. Друге місце посідає часовий лічильник моменту з моменту останнього маркетингового заходу (time_since_promo) з вагою +3.24. Третє місце належить календарній ознаці місяця (month, +1.2), що підтверджує критичну роль сезонної

компоненти. Важливий внесок також роблять сконструйований крос-фактор взаємодії ціни та залишків (price_x_stock, +0.53), базовий рівень ціни (price_unit, +0.4) та бінарний прапорець акції (promotion_flag, +0.63).

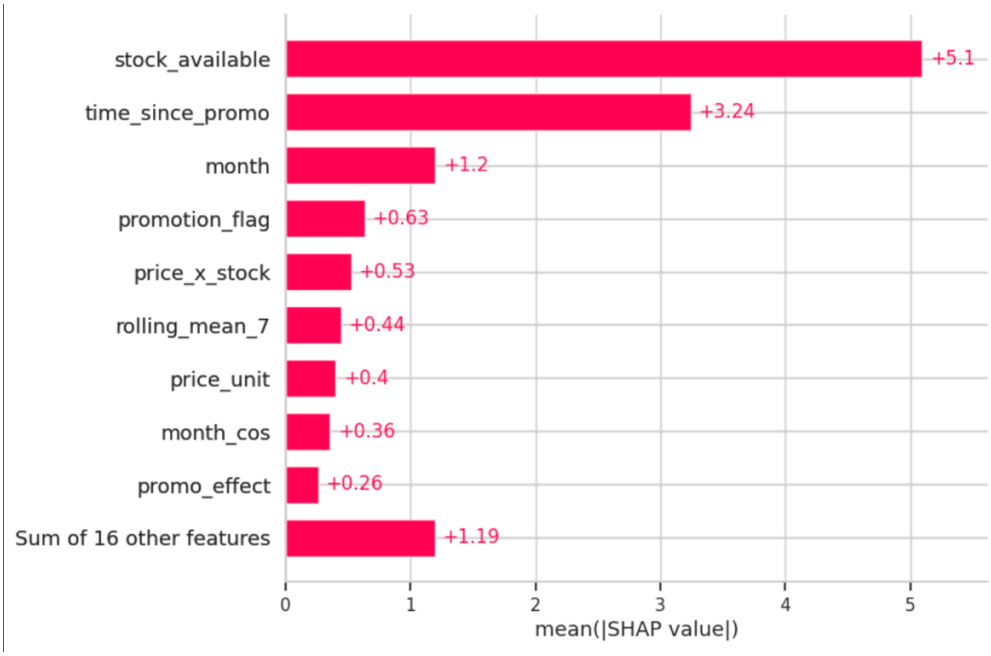

Рисунок 4.2 – Глобальна важливість ознак

Для детального аналізу знаку та характеру цього впливу було побудовано розподільчий графік типу Beeswarm (рисунок 4.3), де кожна точка репрезентує окрему транзакцію, а колір відображає чисельне значення ознаки (червоний – високе, синій – низьке).

Детальний аналіз розподілу SHAP-векторів дозволяє зробити такі науково-практичні висновки щодо логіки роботи моделі:

Фактор товарних залишків (stock_available): низькі значення залишків (сині точки) формують довгий хвіст у негативній зоні графіка, сягаючи позначки -20 за шкалою SHAP. Це математично доводить наявність потужного нелінійного пригнічувального ефекту: модель чітко розпізнає, що відсутність або критичний дефіцит товару на складі фізично унеможливлює здійснення продажу, повністю нівелюючи дію інших позитивних факторів.

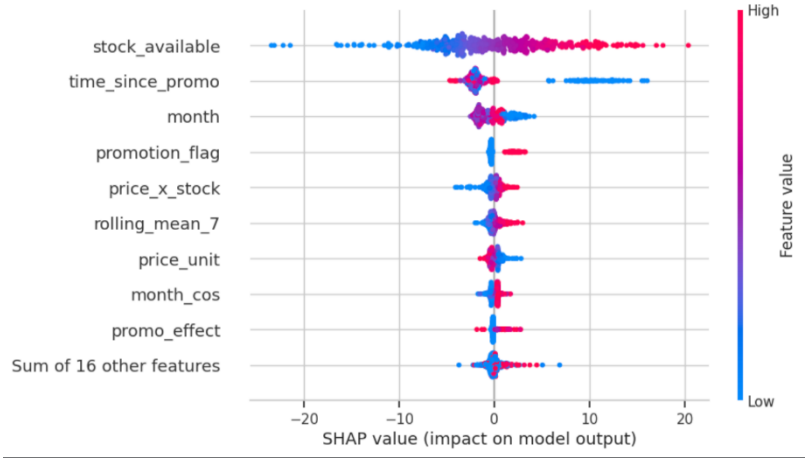

Рисунок 4.3 – Графік Beeswarm

Фактор часу з моменту акції (time_since_promo): високі значення цієї змінної (червоні точки, що означають довгий період без промо-активностей) щільно концентруються у від'ємній зоні впливу на модель.

Натомість мінімальні значення (сині точки, що вказують на нещодавній або поточний запуск акції) зміщують прогноз праворуч, у позитивну зону. Це підтверджує бізнес-гіпотезу про те, що ефект від маркетингового стимулювання збуту має властивість стрімкого згасання в часі.

Еластичність ціни (price_unit та price_x_stock): червоні точки високої вартості товару зміщені ліворуч від нейтральної лінії. Це є прямим математичним підтвердженням дії класичного закону попиту ринку FMCG: підвищення базової ціни одиниці продукції призводить до стійкого та лінійного зниження обсягів її кінцевої реалізації.

Таким чином, використання методів SHAP дозволило повністю демистифікувати «чорну скриньку» градієнтного бустингу LightGBM. Модель не просто вивчила випадкові числові кореляції, а успішно реконструювала реальні причинно-наслідкові економічні закони ритейлу, що підтверджує її високу адекватність, надійність та готовність до інтеграції в реальні операційні процеси планування ланцюгів постачання FMCG-компаній.

ВИСНОВКИ

В рамках виконання кваліфікаційної роботи було розроблено концепцію та реалізовано прототип інтелектуальної системи прогнозування попиту на товари повсякденного вжитку на основі ансамблевих методів машинного навчання, зокрема алгоритму LightGBM для оптимізації управління запасами з урахуванням цінової еластичності, наявності промо-акцій та доступності товару на полиці. Результат роботи зазначеної системи було протестовано на реальних історичних даних транзакцій роздрібної мережі за 2022–2024 роки та отримано високоточні щоденні прогнози обсягів продажів на рівні окремих артикулів. Такі результати роботи можуть бути використані в рамках операційної діяльності ритейлерів, а також у співпраці з дистриб'юторами та виробниками FMCG-сектору, що працюють над проблемами мінімізації втрат від дефіциту та надлишку товарів.

Запропонована система є розвитком і адаптацією сучасних міжнародних рішень у галузі предиктивної аналітики та алгоритмів градієнтного бустингу, враховуючи специфіку високої волатильності споживчого ринку та необхідність обробки багатофакторних нелінійних даних. Порівняно з існуючими класичними статистичними системами (такими як ARIMA), система дозволяє поєднати у собі процес автоматизованого збору інформації, комплексну генерацію часових ознак (Feature Engineering) та побудову індивідуальних прогностичних моделей для різних товарних категорій.

Унікальністю даної системи є запропонований підхід до забезпечення прозорості рішень моделі за допомогою математичного апарату SHAP, що дозволяє не лише отримувати сухий прогноз, а й розуміти причинно-наслідкові зв'язки. В майбутньому планується продовження роботи над висунутою концепцією, інтеграція макроекономічних індикаторів та алгоритмів глибокого навчання, а також перетворення прототипу на повноцінну працюючу систему для ERP-платформ.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Consumer markets – worldwide : statista market forecast. URL: https://www.statista.com/outlook/consumer-markets (date of access: 13.04.2026).

2. Gruen T.W., Corsten D., Bharadwaj S. Retail out-of-stocks: a worldwide examination of causes, rates, and consumer responses. Grocery Manufacturers of America, Washington DC. 2002.

3. Hyndman R. J., Athanasopoulos G. Forecasting: principles and practice. 3rd ed. Melbourne, Australia : OTexts, 2021. 384 p. URL: https://otexts.com/fpp3/ (date of access: 16.04.2026).

4. LightGBM: a highly efficient gradient boosting decision tree / G. Ke et al. *Advances in Neural Information Processing Systems*. 2017. Vol. 30. P. 3146–3154. URL: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9d9bb2154-Abstract.html (date of access: 16.04.2026).

5. Breiman L. Random forests. *Machine learning*. 2001. Vol. 45, no. 1. P. 5–32. URL: https://doi.org/10.1023/a:1010933404324 (date of access: 16.06.2026).

6. Hastie T., Tibshirani R., Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York : Springer, 2009. 764 p.

7. Forecasting: theory and practice / F. Petropoulos et al. *International journal of forecasting*. 2022. URL: https://doi.org/10.1016/j.ijforecast.2021.11.001 (date of access: 16.06.2026).

8. FMCG daily sales data. *Kaggle*. URL: https://www.kaggle.com/datasets/mishashikhov/fmcg-daily-sales-data-to-2022-2024 (date of access: 16.04.2026).

9. McKinney W. Python for data analysis: data wrangling with Pandas, NumPy, and IPython. 3rd ed. Sebastopol : O'Reilly Media, 2022. 578 p.

10. Scikit-learn: machine learning in Python / F. Pedregosa et al. *Journal of Machine Learning Research*. 2011. Vol. 12. P. 2825–2830.

11. Andaur J. M. R., Ruz G. A., Goycoolea M. Predicting out-of-stock using machine learning: an application in a retail packaged foods manufacturing company. *Electronics*. 2021. Vol. 10, no. 22. P. 2787. URL: https://doi.org/10.3390/electronics10222787 (date of access: 16.06.2026).

12. A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks / V. I. Kontopoulou et al. *Future internet*. 2023. Vol. 15, no. 8. P. 255. URL: https://doi.org/10.3390/fi15080255 (date of access: 16.06.2026).

13. Comparing statistical and machine learning methods for time series forecasting in data-driven logistics–a simulation study / L. Schmid et al. *Entropy*. 2024. Vol. 27, no. 1. P. 25. URL: https://doi.org/10.3390/e27010025 (date of access: 16.06.2026).

14. Comparative analysis of LightGBM, Random Forest, and XGBoost with hyperparameter tuning for predicting electricity consumption / M. P. Lukman et al. *Journal of Engineering Science and Technology*. 2026. Vol. 21, no. 1. P. 36–44.

15. A survey of deep learning and foundation models for time series forecasting / J. A. Miller et al. *arXiv*. preprint arXiv:2401.13912. 2024.

20. Lim B., Zohren S. Time-series forecasting with deep learning: a survey. *Philosophical transactions of the royal society A: mathematical, physical and engineering sciences*. 2021. Vol. 379, no. 2194. P. 20200209. URL: https://doi.org/10.1098/rsta.2020.0209 (date of access: 16.06.2026).

21. Філатов В. О., Єрохін А. Л., Золотухін О. В., Кудрявцева М. С. Гібридні імітаційні моделі для задач прийняття рішень в умовах часткової невизначеності. *Відбір і обробка інформації*. 2022. Вип. 50 (126). С. 78–86.

22. A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks / V. I. Kontopoulou et al. *Future internet*. 2023. Vol. 15, no. 8. P. 255. URL: https://doi.org/10.3390/fi15080255 (date of access: 16.06.2026).