НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Навчально-науковий інститут прикладного системного аналізу

Кафедра математичних методів системного аналізу

До захисту допущено:
Завідувач кафедри
_____ Оксана ТИМОЩУК
«___»_____2026 р.

Дипломна робота

на здобуття ступеня бакалавра

за освітньо-професійною програмою «Системний аналіз і управління»

спеціальності 124 «Системний аналіз»

на тему: «Аналіз структури федеральних видатків США на основі статистичних даних»

Виконав:
Студент IV курсу, групи КА-22
Ващенко Віктор Костянтинович _____

Керівник:
асистент кафедри штучного інтелекту
Сидорський Володимир Сергійович _____

Консультант з економічного розділу:
доцент кафедри ЕК, к.е.н., доц.
Рощина Надія Василівна _____

Консультант з нормоконтролю:
к.ф.-м.н. Статкевич Віталій Михайлович _____

Рецензент:
доцент кафедри штучного інтелекту, доктор філософії
Гуськова Віра Геннадіївна _____

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

Київ – 2026 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Навчально-науковий інститут прикладного системного аналізу
Кафедра математичних методів системного аналізу

Рівень вищої освіти – перший (бакалаврський)
Спеціальність – 124 «Системний аналіз»
Освітня програма «Системний аналіз і управління»

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Оксана ТИМОЩУК
«___»_____2026 р.

ЗАВДАННЯ
на дипломну роботу студенту
Ващенку Віктору Костянтиновичу

1. Тема роботи “Аналіз структури федеральних видатків США на основі статистичних даних”, керівник роботи Сидорський Володимр Сергійович, затверджені наказом по університету від 27.05.2026 р. №1944-с.

2. Термін подання студентом роботи «10» червня 2026 р.

3. Вихідні дані до роботи: відкритий набір статистичних даних Budget Spending Datasets, наукові джерела з прогнозування часових рядів, машинного навчання та інтерпретації моделей, а також результати власного програмного аналізу в Python.

4. Зміст роботи: теоретичний огляд методів прогнозування, первинний аналіз і підготовка даних, побудова базової та XGBoost моделей, застосування Prophet, SHAP і Permutation Importance, порівняння результатів та формування висновків.

5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо): презентація дипломної роботи, графіки динаміки витрат, порівняння прогнозів і метрик моделей, SHAP-графіки, графіки Prophet, залишків та аномалій, а також таблиці результатів аналізу й моделювання.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Функціонально-вартісний аналіз	Рощина Н.В., доц., к.е.н.		

7. Дата видачі завдання: 10.02.2026 р.

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Примітка
1	Вибір і погодження теми	10.02.2026	виконано
2	Аналіз наукових джерел	28.02.2026	виконано
3	Огляд методів прогнозування	20.03.2026	виконано
4	Вибір і аналіз датасету	05.04.2026	виконано
5	Підготовка даних	20.04.2026	виконано
6	Побудова моделей	05.05.2026	виконано
7	Аналіз результатів	20.05.2026	виконано
8	Оформлення роботи	05.06.2026	виконано

Студент _____ Віктор ВАЩЕНКО

Керівник _____ Володимир СИДОРСЬКИЙ

РЕФЕРАТ

Дипломна робота: 178 с., 54 табл., 21 рис., 2 дод, 63 джерела.

БЮДЖЕТНІ ВИТРАТИ, ФІНАНСОВЕ ПРОГНОЗУВАННЯ, ЧАСОВІ РЯДИ, PROPHET, XGBOOST, BLACK-BOX МОДЕЛЬ, SHAP-АНАЛІЗ, PERMUTATION IMPORTANCE, МЕТРИКИ ЯКОСТІ, МАШИННЕ НАВЧАННЯ.

Об’єктом дослідження є аналіз і прогнозування бюджетних витрат, а предметом – методи прогнозування фінансово-бюджетних часових рядів та інтерпретації моделей машинного навчання. Метою роботи є побудова й порівняння моделей прогнозування та визначення найвпливовіших ознак.

У роботі розглянуто Prophet, XGBoost Regressor, Permutation Importance і SHAP-аналіз. У практичній частині використано відкритий датасет Budget Spending Datasets, який через відсутність даних про країну, рік і належність до конкретної бюджетної системи розглядається як прикладова вибірка.

Проведено первинний аналіз, перевірено пропуски, дублікати й коректність відхилень. Дані агреговано за місяцями, департаментами та кварталами, після чого сформовано ознаки для прогнозування.

Побудовано базову математичну модель і black-box модель на основі XGBoost. Порівняння показало, що XGBoost забезпечив кращу якість: MAPE зменшилося з 10,67% до 7,63%. Найбільший вплив мають лагові значення фактичних витрат, ковзне середнє та планові витрати.

Додатково застосовано Prophet для аналізу квартального ряду, досліджено тренд, залишки та потенційні аномалії. Практичне значення роботи полягає у формуванні підходу до аналізу й прогнозування бюджетних витрат, оцінювання відхилень і підвищення прозорості моделей.

ABSTRACT

Bachelor’s thesis: 178 pages, 54 tables, 21 figures, 2 appendices, 63 references.

BUDGET EXPENDITURES, FINANCIAL FORECASTING, TIME SERIES, PROPHET, XGBOOST, BLACK-BOX MODEL, SHAP ANALYSIS, PERMUTATION IMPORTANCE, QUALITY METRICS, MACHINE LEARNING.

The object of the research is the analysis and forecasting of budget expenditures, while the subject is the set of methods for forecasting financial and budgetary time series and interpreting machine learning models. The aim of the thesis is to build and compare forecasting models and identify the most influential features.

The thesis examines Prophet, XGBoost Regressor, Permutation Importance, and SHAP analysis. The practical part uses the open Budget Spending Datasets, which is treated as an illustrative sample because it does not specify a country, year, or connection to a particular budget system.

An initial data analysis was performed, including checks for missing values, duplicates, and variance correctness. The data were aggregated by months, departments, and quarters, after which features for forecasting were created.

A baseline mathematical model and a black-box model based on XGBoost were developed. The comparison showed that XGBoost provided better performance: MAPE decreased from 10.67% to 7.63%. The most influential features were lagged actual spending values, the moving average, and budgeted expenditures.

Prophet was additionally applied to analyze the quarterly time series, including trend, residuals, and potential anomalies. The practical value of the thesis lies in developing an approach to budget expenditure analysis and forecasting, variance assessment, and improving model transparency.

ЗМІСТ

ВСТУП....8
РОЗДІЛ 1 ОГЛЯД МОДЕЛЕЙ І МЕТОДІВ ПРОГНОЗУВАННЯ ФІНАНСОВИХ ЧАСОВИХ РЯДІВ.... 12
1.1 Загальна постановка задачі прогнозування часових рядів.... 14
1.2 Специфіка задачі фінансового прогнозування....18
1.3 Методи й моделі прогнозування часових рядів....20
1.4 Метрики оцінки якості прогнозування....23
1.5 Сучасні методи прогнозування часових рядів....27
1.6 Методи аналізу важливості ознак.... 32
Висновки до розділу 1....33
РОЗДІЛ 2 ЗАСТОСУВАННЯ СУЧАСНИХ МЕТОДІВ ПРОГНОЗУВАННЯ ДЛЯ АНАЛІЗУ БЮДЖЕТНИХ ВИТРАТ....35
2.1 Prophet....36
2.2 Permutation Importance.... 38
2.3 SHAP-аналіз.... 40
2.4 Огляд популярних вибірок даних.... 43
Висновки до розділу 2....44
РОЗДІЛ 3 МЕТОД АНАЛІЗУ СТРУКТУРИ БЮДЖЕТНИХ ВИТРАТ.... 46
3.1 Первинний аналіз датасету....46
3.2 Підготовка даних до моделювання.... 55
3.3 Побудова базової моделі прогнозування.... 62
3.4 Побудова XGBoost Regressor моделі.... 68
3.5 Аналіз роботи XGBoost моделі та дослідження важливості ознак..... 73
3.6 Математична модель прогнозування бюджетних витрат....77
3.7 Порівняння математичної та XGBoost моделі.... 81
3.8 SHAP-аналіз моделі....87
3.8.1 Глобальна важливість ознак.... 88
3.8.2 Аналіз напрямку впливу ознак....89

3.8.3 Локальне пояснення окремих прогнозів..... 91
3.8.4 Порівняння SHAP із Permutation Importance..... 92
3.8.5 Обмеження SHAP-аналізу..... 93
3.8.6. Висновок до SHAP-аналізу..... 93
3.9 Абляційне дослідження ознак..... 94
3.10 Прогнозування часових рядів методом Prophet.....97
3.11 Порівняння прогнозу з фактичними даними..... 102
3.12 Аналіз тренду та залишків.....108
3.13 Аналіз аномалій та дослідження моделі Prophet..... 113
3.13.1 Дослідження поведінки моделі Prophet.....115
3.13.2 Аномалії та їхній вплив на прогноз.....117
3.13.3 Висновки за результатами аналізу..... 119
3.14 Порівняльний аналіз підходів.....119
Висновки до розділу 3.....125
РОЗДІЛ 4 ЕКОНОМІЧНИЙ АНАЛІЗ ТА ОЦІНКА ВАРТОСТІ ПРОГРАМНОГО ПРОДУКТУ..... 126
4.1 Постановка задачі проектування.....126
4.2 Обґрунтування функцій програмного продукту.....128
4.3 Обґрунтування системи параметрів програмного продукту..... 133
4.4 Аналіз експертного оцінювання параметрів.....136
4.5 Аналіз рівня якості варіантів реалізації функцій..... 139
Висновки до розділу 4.....145
ВИСНОВКИ.....147
ПЕРЕЛІК ПОСИЛАНЬ..... 150
ДОДАТОК А ЛІСТИНГ ПРОГРАМИ.....158

ВСТУП

У сучасних умовах аналіз бюджетних витрат є важливим інструментом фінансового планування, оцінювання ефективності використання ресурсів і підтримки управлінських рішень. Бюджетні показники відображають обсяг витрат, структуру пріоритетів, рівень виконання плану, наявність перевитрат або економії та загальну динаміку фінансової діяльності. Саме тому дослідження бюджетних витрат на основі статистичних даних є актуальним напрямом прикладного системного аналізу.

Особливе значення має аналіз фінансових часових рядів, оскільки бюджетні процеси змінюються в часі під впливом сезонних факторів, управлінських рішень, економічних умов, кризових подій і особливостей окремих департаментів. Через це простого описового аналізу недостатньо: необхідно застосовувати методи прогнозування, які дають змогу оцінити майбутню динаміку показників, порівняти фактичні та очікувані значення й визначити чинники, що найбільше впливають на результат.

Класичні статистичні методи, зокрема авторегресійні моделі, моделі ковзного середнього та ARIMA, тривалий час були основою аналізу фінансових часових рядів. Вони мають чітку математичну структуру та добре підходять для дослідження лінійних залежностей. Водночас сучасні фінансові дані часто містять нелінійні зв’язки, категоріальні ознаки, відхилення від плану, аномальні значення та взаємодії між факторами. Це зумовлює потребу в застосуванні методів машинного навчання та спеціалізованих моделей прогнозування.

Одним із сучасних підходів є використання моделі Prophet, яка дає змогу аналізувати часові ряди через трендову, сезонну та випадкову компоненти. Вона є зручною для візуального аналізу динаміки показника та побудови прогнозу на основі попередніх значень. Іншим напрямом є застосування black-box моделей машинного навчання, зокрема XGBoost

Regressor, який дозволяє працювати з табличними ознаками, лаговими змінними, ковзними середніми, плановими показниками та категоріальними характеристиками. Проблема інтерпретованості моделей є особливо важливою у фінансово-бюджетному аналізі. Якщо модель використовується лише як інструмент отримання прогнозу, її практична цінність є обмеженою. Для аналітичного дослідження необхідно також розуміти, які саме ознаки вплинули на прогноз і чому модель сформувала певний результат. Для цього використовуються методи аналізу важливості ознак, зокрема Permutation Importance та SHAP-аналіз. Вони дозволяють оцінити внесок окремих факторів у результат прогнозування та зробити black-box модель більш прозорою.

Актуальність теми роботи полягає в необхідності поєднання статистичного аналізу, методів прогнозування часових рядів і сучасних інструментів інтерпретації моделей для дослідження структури бюджетних витрат. Такий підхід дозволяє не лише описати наявні дані, а й побудувати прогноз, порівняти різні моделі, дослідити залишки, виявити можливі аномалії та визначити найважливіші ознаки, що впливають на витрати.

Метою роботи є аналіз структури бюджетних витрат на основі статистичних даних і побудова моделей прогнозування, що дозволяють оцінити фактичні витрати, порівняти якість різних підходів та інтерпретувати вплив окремих ознак на результат моделювання.

Для досягнення поставленої мети необхідно виконати такі завдання:

1) розглянути теоретичні основи прогнозування часових рядів та специфіку фінансового прогнозування;

2) проаналізувати класичні та сучасні методи прогнозування, зокрема ARIMA, Prophet, XGBoost і нейромережеві підходи;

3) дослідити метрики оцінювання якості прогнозу, зокрема MAE, RMSE, MAPE та R^2;

4) розглянути методи аналізу важливості ознак, зокрема Permutation Importance та SHAP;

5) провести первинний аналіз датасету бюджетних витрат;

6) підготувати дані до моделювання шляхом агрегації, формування кварталів, лагових ознак і цільової змінної;

7) побудувати базову математичну модель прогнозування;

8) побудувати black-box модель на основі XGBoost Regressor;

9) порівняти якість математичної та black-box моделі;

10) провести аналіз важливості ознак і SHAP-аналіз моделі;

11) застосувати модель Prophet для прогнозування агрегованого часового ряду;

12) виконати аналіз тренду, залишків та потенційних аномалій;

13) сформувати порівняльний аналіз використаних підходів і визначити їхні переваги та обмеження.

Об’єктом дослідження є процес аналізу та прогнозування бюджетних витрат на основі статистичних даних.

Предметом дослідження є методи, моделі та алгоритми прогнозування фінансово-бюджетних часових рядів, а також методи інтерпретації результатів моделей машинного навчання.

У практичній частині роботи використано відкритий набір даних Budget Spending Datasets. Оскільки в ньому відсутня інформація про країну, рік та належність до конкретної державної бюджетної системи, цей датасет розглядається не як офіційна статистика федеральних видатків США, а як прикладова статистична вибірка для апробації методів аналізу та прогнозування бюджетних витрат. Такий підхід дозволяє перевірити роботу моделей на структурованих бюджетних даних, що містять планові та фактичні витрати, департаменти й відхилення.

Інформаційною базою дослідження є відкритий набір даних Budget Spending Datasets, що містить інформацію про місяць, департамент, заплановані витрати, фактичні витрати та відхилення між ними. На основі цих даних проведено первинний статистичний аналіз, поквартальну агрегацію, формування ознак і побудову моделей прогнозування. Оскільки

датасет не містить повної календарної шкали з роками, у роботі він розглядається як прикладова статистична вибірка для апробації методів аналізу та прогнозування бюджетних витрат. Методологічну основу роботи становлять методи статистичного аналізу, методи прогнозування часових рядів, регресійні моделі машинного навчання, методи оцінювання якості прогнозу та методи інтерпретації моделей. Для практичної реалізації використано мову програмування Python і бібліотеки Pandas, NumPy, Matplotlib, Scikit-learn, XGBoost, Prophet та SHAP. Науково-практичне значення роботи полягає в розробленні послідовного підходу до аналізу бюджетних витрат, який поєднує первинний аналіз даних, підготовку ознак, побудову базової та black-box моделей, використання Prophet, аналіз залишків, дослідження аномалій і пояснення результатів за допомогою методів інтерпретації. Такий підхід може бути використаний для аналізу бюджетних показників, оцінювання відхилень між плановими та фактичними витратами, виявлення важливих факторів прогнозування та підтримки прийняття фінансових рішень.

Робота складається зі вступу, чотирьох розділів, висновків, переліку посилань і двох додатків. У першому розділі розглянуто теоретичні основи прогнозування фінансових часових рядів, класичні та сучасні методи прогнозування, метрики якості та методи аналізу важливості ознак. У другому розділі описано застосування сучасних методів прогнозування та інтерпретації моделей для аналізу бюджетних витрат. У третьому розділі проведено практичний аналіз датасету, підготовку даних, побудову моделей, порівняння результатів, SHAP-аналіз, дослідження Prophet, аналіз залишків, аномалій та порівняння використаних підходів. Четвертий розділ присвячений функціонально-вартісному аналізу програмного продукту.

РОЗДІЛ 1 ОГЛЯД МОДЕЛЕЙ І МЕТОДІВ ПРОГНОЗУВАННЯ ФІНАНСОВИХ ЧАСОВИХ РЯДІВ

Проблема прогнозування часових рядів є важливою задачею статистичного аналізу, економетрики та інтелектуального аналізу даних. Часовий ряд розглядається як послідовність спостережень, упорядкованих у часі, а метою прогнозування є побудова моделі, здатної виявити закономірності та оцінити майбутні значення. Hyndman та Athanasopoulos [1] розглядають прогнозування як інструмент підтримки прийняття рішень, що потребує аналізу тренду, сезонності, залишків і якості прогнозу. Це особливо важливо для фінансово-бюджетних даних, які можуть мати довгострокові тенденції, нестабільність і структурні зміни.

Класичну основу прогнозування становлять моделі AR, MA, ARMA та ARIMA, описані Box, Jenkins, Reinsel та Ljung [2]. AR-модель пояснює поточне значення через попередні значення ряду, MA враховує попередні похибки, а ARIMA дає змогу працювати з нестаціонарними рядами. Тому ARIMA часто використовується як базова модель для порівняння з сучаснішими методами.

Важливим джерелом є також праця Hamilton [3], де розглянуто економетричне моделювання динамічних процесів, стаціонарність, авторегресійні процеси та прогнозування економічних показників. Це є релевантним для бюджетних даних, які залежать від попередніх значень, економічних умов і змін у фінансовому плануванні.

Специфіку фінансових часових рядів досліджено в роботі Tsay [4]. Автор зазначає, що такі ряди можуть мати змінну дисперсію, викиди, нестабільність і залежність від зовнішніх факторів. Для бюджетних витрат це

означає необхідність врахування не лише середньої тенденції, а й аномальних періодів та обмежень прогнозування за історичними даними.

Окреме місце займають декомпозиційні підходи, у яких часовий ряд подається як поєднання тренду, сезонності та випадкової компоненти [1]. Такий підхід дає змогу відокремити довгострокові зміни від короткострокових коливань і виявити нетипові періоди у структурі бюджетних витрат.

Серед сучасних декомпозиційних моделей важливою є Prophet, запропонована Taylor та Letham [5]. Вона поєднує тренд, сезонність, подієві ефекти та випадкову похибку. Для фінансово-бюджетних даних Prophet може використовуватися як для прогнозування, так і для аналізу тренду та виявлення відхилень від очікуваної динаміки.

Методи машинного навчання дозволяють розглядати прогнозування часових рядів як задачу регресії. Для цього формуються лагові значення, ковзні середні, прирости та інші ознаки. Загальні принципи таких підходів описано James, Witten, Hastie та Tibshirani [6]. Вони є корисними, коли зв’язки в даних складні й не описуються простими статистичними моделями.

Одним із ефективних класів моделей є ансамблеві методи, зокрема градієнтний бустинг. Friedman [7] описав його як послідовну побудову моделей, кожна з яких зменшує помилки попередніх. XGBoost, запропонований Chen та Guestrin [8], поєднує високу точність, регуляризацію та ефективність роботи з великими наборами даних. У цій роботі такі моделі можуть застосовуватися для прогнозування бюджетних витрат на основі часових і категоріальних ознак.

Для послідовних даних також використовуються нейронні мережі, зокрема LSTM. Hochreiter та Schmidhuber [9] запропонували архітектуру, здатну зберігати інформацію на довгих часових інтервалах. LSTM може бути корисною для складних нелінійних залежностей, однак потребує значного обсягу даних, налаштування параметрів і обережної інтерпретації.

Оцінювання якості прогнозу є важливим етапом дослідження. Hyndman та Koehler [10] розглядають метрики MAE, RMSE, MAPE та MASE. MAE показує середню абсолютну помилку, RMSE сильніше враховує великі помилки, MAPE подає похибку у відсотках, але може бути нестабільною за малих значень, а MASE дає змогу коректніше порівнювати моделі.

Оскільки моделі машинного навчання часто є «чорними скриньками», важливою є інтерпретація результатів. Breiman [11] розглянув важливість змінних у випадкових лісах, а Fisher, Rudin та Dominici [12] описали permutation importance як спосіб оцінити залежність моделі від окремих ознак.

Поширеним методом пояснення прогнозів є SHAP, запропонований Lundberg та Lee [13]. Він базується на значеннях Шеплі та дозволяє визначати внесок кожної ознаки в прогноз. У межах цієї роботи SHAP може використовуватися для пояснення того, які характеристики часового ряду найбільше впливають на прогноз бюджетних витрат.

Отже, огляд літератури показує доцільність комплексного підходу до прогнозування фінансово-бюджетних часових рядів. Класичні моделі формують статистичну основу дослідження, Prophet дає змогу аналізувати тренд і відхилення, а методи машинного навчання та інтерпретації результатів розширюють можливості прогнозування й пояснення отриманих результатів.

1.1 Загальна постановка задачі прогнозування часових рядів

Прогнозування часових рядів є задачею побудови моделі, яка на основі попередніх значень певного показника дозволяє оцінити його майбутню динаміку. Часовий ряд можна подати як послідовність спостережень, упорядкованих за часом:

$$y_1, y_2, \ldots, y_t,$$

де y_t – значення досліджуваного показника в момент часу t. Основна мета прогнозування полягає в тому, щоб, маючи історичні значення ряду, побудувати функцію або модель, яка дозволяє отримати прогноз майбутніх значень:

$$\hat{y}_{t+1}, \hat{y}_{t+2}, \ldots, \hat{y}_{t+h},$$

де h – горизонт прогнозування, а $\hat{y}_{t+h}$ – прогнозоване значення показника через h періодів.

У загальному вигляді задача прогнозування часових рядів передбачає виявлення закономірностей у минулих спостереженнях та використання цих закономірностей для оцінювання майбутніх значень. Як зазначають Hyndman та Athanasopoulos [1], прогнозування не зводиться лише до механічного продовження ряду вперед. Воно потребує аналізу структури даних, вибору відповідної моделі, оцінювання похибок і перевірки того, наскільки отриманий прогноз є корисним для практичного застосування. Особливістю часових рядів є те, що спостереження в них зазвичай не є незалежними. На відміну від звичайних табличних даних, де порядок рядків може не мати суттєвого значення, у часовому ряді порядок спостережень є принциповим. Поточне значення показника може залежати від попередніх значень, від довгострокового тренду, циклічних коливань, сезонних закономірностей або випадкових зовнішніх факторів. Саме тому для прогнозування часових рядів використовуються спеціальні методи, здатні враховувати часову залежність.

У класичній постановці задача прогнозування може розглядатися як побудова моделі виду:

$$y_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-p}, x_t) + \varepsilon_t$$

де y_t – фактичне значення показника в момент часу t, $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$ – попередні значення часового ряду, x_t – додаткові пояснювальні змінні, якщо вони використовуються, f – функція, яка описує залежність між вхідними

даними та прогнозованим значенням, а ε_t – випадкова похибка моделі. У класичних статистичних моделях функція f часто має лінійний вигляд, тоді як у сучасних методах машинного навчання вона може бути складною нелінійною функцією.

Важливим етапом є визначення горизонту прогнозування. Він може бути короткостроковим, середньостроковим або довгостроковим. У фінансово-економічних задачах короткострокові прогнози використовуються для оперативного планування, а довгострокові – для стратегічного аналізу та прийняття управлінських рішень. У межах цієї роботи прогнозування розглядається як інструмент оцінювання майбутньої динаміки фінансово-бюджетних показників.

Задача прогнозування часових рядів складається з кількох основних підзадач.

1. Збирання та підготовка даних: перевірка повноти, коректності часової шкали, наявності пропусків, повторів або помилкових значень. Якість початкових даних безпосередньо впливає на точність прогнозу.
2. Попередній аналіз ряду включає візуалізацію динаміки, виявлення тренду, сезонності, циклічності та аномальних значень. Це дозволяє сформувати початкове уявлення про структуру даних і вибрати відповідні методи прогнозування.
3. Перевірка стаціонарності для класичних моделей, зокрема ARMA та ARIMA, причому важливо, щоб ряд мав стабільні статистичні характеристики або міг бути приведений до стаціонарного вигляду. Для цього можуть використовуватися диференціювання, логарифмування або інші перетворення [2].
4. Формування ознак – у класичних моделях використовуються значення ряду та їхні лаги, а в моделях машинного навчання можуть додаватися ковзні середні, темпи приросту, календарні змінні, індикатори періодів та інші похідні характеристики.

5. Вибір і побудова моделей – для прогнозування можуть застосовуватися як класичні статистичні методи AR, MA, ARMA, ARIMA, так і сучасні підходи: градієнтний бустинг, нейронні мережі, LSTM, Prophet та інші моделі, здатні враховувати складні й нелінійні залежності.
6. Навчання та перевірка моделей. У часових рядах дані не можна випадково перемішувати, оскільки це порушує часову структуру. Зазвичай ранні спостереження використовуються для навчання, а пізніші – для перевірки якості прогнозу.
7. Оцінювання якості прогнозу – для цього використовуються метрики MAE, MSE, RMSE, MAPE та інші. MAE показує середню абсолютну помилку, RMSE сильніше враховує великі відхилення, а MAPE подає похибку у відсотках, проте може бути нестабільною для малих значень [10].
8. Інтерпретація результатів. Для практичного застосування важливо не лише отримати прогноз, а й пояснити, які фактори вплинули на результат. Для цього використовуються permutation importance та SHAP, які дають змогу оцінити важливість окремих ознак.
9. Аналіз аномалій і відхилень від очікуваної динаміки. У фінансових даних такі відхилення можуть бути пов’язані з кризами, змінами політики, надзвичайними подіями або змінами у структурі витрат. Їх виявлення допомагає уникнути помилкової інтерпретації разових подій як сталої закономірності.
10. Формування висновків і практична інтерпретація прогнозу. На цьому етапі порівнюються результати моделей, оцінюється їх точність, стабільність, інтерпретованість і придатність для аналізу фінансово-бюджетних показників.

Таким чином, постановка задачі прогнозування часових рядів передбачає підготовку даних, попередній аналіз, перевірку властивостей ряду, формування ознак, вибір і навчання моделей, оцінювання якості прогнозу, інтерпретацію результатів та аналіз аномалій. У межах цієї роботи такий

підхід дозволяє системно дослідити фінансово-бюджетні показники, порівняти класичні й сучасні методи та визначити чинники, що впливають на прогноз.

1.2 Специфіка задачі фінансового прогнозування

Фінансове прогнозування є окремим напрямом аналізу часових рядів, оскільки фінансові показники формуються під впливом економічних, соціальних, політичних та управлінських чинників. На відміну від технічних або природничих процесів, фінансові ряди не завжди мають стабільну структуру. Їхня динаміка може змінюватися через кризи, реформи, бюджетні рішення, інфляцію, зміни законодавства або надзвичайні події. Тому прогнозування фінансових показників потребує не лише застосування моделей, а й змістовного аналізу даних. У межах цієї роботи фінансове прогнозування розглядається на прикладі бюджетних витрат. Такі дані відображають не лише економічну динаміку, а й пріоритети державної політики. Зміна обсягів або часток видатків може бути пов’язана з військовими витратами, соціальними програмами, охороною здоров’я, освітою, інфраструктурою чи обслуговуванням боргу. Тому аналіз такого ряду має як статистичне, так і аналітичне значення.

Однією з головних особливостей фінансових часових рядів є наявність тренду. У бюджетних даних він може відображати поступове зростання державних витрат або зміну структури фінансування. Водночас важливо відрізняти довгострокову тенденцію від короткочасного відхилення, оскільки неправильна інтерпретація разового стрибка може викривити прогноз.

Іншою важливою особливістю є наявність структурних зламів. Вони виникають тоді, коли закономірності одного періоду перестають діяти в іншому. Для фінансових даних це може бути пов’язано з кризами, війнами,

політичними рішеннями або масштабними державними програмами. У таких випадках модель, навчена на старих даних, може гірше описувати нову динаміку. Фінансові часові ряди також можуть містити аномальні значення. У федеральних видатках вони можуть бути спричинені кризовими періодами, надзвичайними витратами, змінами в обліку або нетиповими бюджетними рішеннями. Такі значення не завжди є помилками, тому їх потрібно не просто видаляти, а аналізувати їхній вплив на модель.

Ще однією проблемою є нестаціонарність. Фінансові дані часто мають змінне середнє значення, дисперсію або довгострокові тенденції. Тому перед моделюванням можуть застосовуватися диференціювання, логарифмування, нормалізація та інші перетворення. Водночас надмірна трансформація може ускладнити інтерпретацію результатів. У фінансовому прогнозуванні точність моделі не є єдиним критерієм якості. Важливо також розуміти, чому модель сформувала певний прогноз. Це особливо актуально для моделей машинного навчання, які можуть бути точними, але складними для пояснення. Для цього в роботі можуть використовуватися permutation pmportance та SHAP, які дозволяють оцінити вплив окремих ознак на результат. Фінансове прогнозування завжди пов’язане з невизначеністю. Навіть якісна модель не гарантує точного передбачення майбутніх значень, оскільки подальша динаміка може залежати від нових політичних, економічних або зовнішніх факторів. Тому прогноз слід розглядати як аналітичну оцінку можливої динаміки, а не як точне передбачення.

У межах дослідження доцільно поєднувати кілька типів моделей. ARIMA може використовуватися як базовий статистичний підхід, Prophet – для аналізу тренду та аномалій, а методи машинного навчання, зокрема градієнтний бустинг, – для роботи з ширшим набором ознак і нелінійними залежностями. Нейронні мережі також можуть бути корисними, але потребують достатнього обсягу даних і обережного застосування.

Таким чином, фінансове прогнозування є комплексною задачею, оскільки дані можуть містити тренди, структурні злами, аномалії,

нестаціонарність і залежність від зовнішніх факторів. Тому необхідно виконати попередній аналіз, сформувати ознаки, порівняти кілька моделей, оцінити якість прогнозу та пояснити отримані результати.

1.3 Методи й моделі прогнозування часових рядів

Методи прогнозування часових рядів можна умовно поділити на кілька груп: прості базові методи, класичні статистичні моделі, декомпозиційні моделі, методи машинного навчання та нейромережеві підходи. Кожна з цих груп має власну сферу застосування, переваги й обмеження. У межах цієї роботи доцільно розглядати не одну окрему модель, а сукупність підходів, оскільки фінансово-бюджетні часові ряди можуть містити тренд, аномальні значення, нелінійні залежності та структурні зміни. Базові методи прогнозування часто використовуються як початкова точка для порівняння зі складнішими моделями [1]. До них належать наївний прогноз, середнє значення, ковзне середнє та експоненційне згладжування. Наївний прогноз передбачає, що майбутнє значення дорівнюватиме останньому відомому значенню ряду, тому він є простим орієнтиром для оцінювання ефективності інших моделей [1]. Метод ковзного середнього застосовується для згладжування випадкових коливань і виявлення загальної тенденції. Він корисний для попереднього аналізу фінансових рядів, однак має обмежену прогностичну здатність, оскільки не враховує складні залежності між значеннями [1].

Експоненційне згладжування надає більшу вагу останнім спостереженням. Його різновиди дозволяють враховувати відсутність або наявність тренду й сезонності. У фінансово-бюджетних даних такі методи можуть використовуватися для базового прогнозу та аналізу загальної динаміки [1].

Однією з основних класичних моделей часових рядів є авторегресійна модель AR. Вона ґрунтується на припущенні, що поточне значення ряду залежить від його попередніх значень. У загальному вигляді авторегресійну модель порядку p можна подати так:

$$y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + \varepsilon_t,$$

де y_t – значення ряду в момент часу t, c – стала, $\varphi_1,\varphi_2,\ldots,\varphi_p$, – параметри моделі, p – кількість лагів, а ε_t – випадкова похибка. Така модель є корисною тоді, коли у часовому ряді існує залежність між поточними та попередніми значеннями. Наприклад, у фінансових даних поточний рівень витрат може частково залежати від рівня витрат у попередні роки.

Іншим класичним підходом є модель ковзного середнього MA. На відміну від AR-моделі, вона описує поточне значення ряду через попередні випадкові похибки. У загальному вигляді MA-модель порядку q записується так:

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q},$$

де μ – середнє значення ряду, $\theta_1,\theta_2,\ldots,\theta_q$ - параметри моделі, а ε_t - випадкова похибка. Модель MA дозволяє враховувати вплив випадкових шоків або відхилень, які могли впливати на динаміку показника в минулі періоди.

Поєднання авторегресійної моделі та моделі ковзного середнього формує модель ARMA. Вона одночасно враховує як попередні значення самого ряду, так і попередні похибки. ARMA-модель є ефективною для стаціонарних часових рядів, тобто таких, у яких середнє значення, дисперсія та структура залежностей залишаються відносно стабільними в часі. Проте багато реальних фінансових рядів не є стаціонарними, оскільки містять тренд або змінну дисперсію. Саме тому на практиці частіше використовується розширення ARMA - модель ARIMA.

Модель ARIMA є однією з найпоширеніших класичних моделей прогнозування часових рядів. Її назва розшифровується як AutoRegressive Integrated Moving Average. Вона має три основні параметри p, d, q. Параметр

p визначає порядок авторегресійної частини, d – кількість диференціювань, необхідних для приведення ряду до стаціонарного вигляду, а q – порядок частини ковзного середнього. У загальному вигляді ARIMA позначають як ARIMA (p,d,q).

Головною перевагою ARIMA є можливість працювати з нестаціонарними рядами. Якщо часовий ряд має тренд, його можна продиференціювати, тобто перейти від абсолютних значень до різниць між сусідніми спостереженнями. Наприклад, замість значень y_t аналізується різниця:

$$y'_t = y_t - y_{t-1}.$$

Таке перетворення дозволяє зменшити вплив тренду й зробити ряд більш придатним для статистичного моделювання. Для фінансово-бюджетних даних це має важливе значення, оскільки показники видатків часто мають довгострокове зростання або спад.

Якщо часовий ряд містить сезонність, може використовуватися сезонна модель SARIMA. Вона розширює ARIMA за рахунок додаткових сезонних параметрів. SARIMA застосовується тоді, коли значення ряду повторюють певні закономірності через фіксовані проміжки часу. Для щомісячних або квартальних фінансових даних це може бути особливо актуальним. Однак якщо дані мають річну частоту, як у випадку багатьох бюджетних наборів даних, сезонність може бути менш вираженою або взагалі відсутньою.

Окремо варто розглянути модель Prophet. Вона належить до декомпозиційних моделей і подає часовий ряд як суму кількох компонентів: тренду, сезонності, подієвих ефектів і випадкової похибки. У спрощеному вигляді модель можна подати так:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t,$$

де $g(t)$ – трендова компонента, $s(t)$ – сезонність, $h(t)$ – вплив свят або подій, а ε_t – випадкова похибка. Перевагою Prophet є інтерпретованість і можливість окремо аналізувати складові часового ряду. У межах фінансового

прогнозування ця модель може бути корисною для виявлення довгострокового тренду та періодів, які суттєво відхиляються від очікуваної динаміки.

1.4 Метрики оцінки якості прогнозування

Оцінювання якості прогнозу є одним із ключових етапів побудови моделей часових рядів. Сам факт отримання прогнозних значень ще не означає, що модель є придатною для практичного використання. Необхідно визначити, наскільки прогнозовані значення відрізняються від фактичних, чи є похибка прийнятною для поставленої задачі, а також чи справді складніша модель має перевагу над простішими базовими підходами. Саме для цього використовуються метрики оцінки якості прогнозування.

У загальному вигляді якість прогнозу оцінюється шляхом порівняння фактичних значень часового ряду y_t та прогнозованих значень $\hat{y}_t$. Різниця між ними називається похибкою прогнозу:

$$e_t = y_t - \hat{y}_t,$$

де e_t – похибка прогнозу в момент часу t, y_t – фактичне значення показника, а $\hat{y}_t$ – значення, отримане за допомогою моделі. Якщо похибка є малою, модель краще відтворює реальну динаміку показника. Якщо похибка велика, прогноз є менш точним і потребує перегляду моделі, ознак або способу підготовки даних.

Однією з найпоширеніших метрик є середня абсолютна похибка – MAE. Вона обчислюється як середнє значення модулів різниць між фактичними та прогнозованими значеннями:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - \widehat{y}_t\right|.$$

Перевага MAE полягає в простоті інтерпретації. Ця метрика показує, на скільки одиниць у середньому модель помиляється під час прогнозування. Наприклад, якщо аналізуються федеральні видатки у відсотках або в частках, MAE показує середній абсолютний розмір відхилення прогнозу від реального значення. MAE є зручною для практичного аналізу, оскільки вона вимірюється в тих самих одиницях, що й досліджуваний показник.

Іншою поширеною метрикою є середньоквадратична похибка – MSE. Вона визначається як середнє значення квадратів похибок:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \widehat{y}_t)^2.$$

Особливість MSE полягає в тому, що великі помилки отримують значно більшу вагу через піднесення до квадрата. Це означає, що модель з окремими великими відхиленнями буде мати значно гірше значення MSE. Така властивість є корисною тоді, коли великі помилки прогнозування є особливо небажаними. Водночас MSE складніше інтерпретувати, оскільки вона вимірюється у квадраті одиниць початкового показника.

Для зручнішої інтерпретації часто використовується корінь із середньоквадратичної похибки – RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \widehat{y}_t)^2}.$$

RMSE, як і MAE, вимірюється в тих самих одиницях, що й початковий показник, але при цьому зберігає властивість сильніше штрафувати великі помилки. Тому RMSE є корисною метрикою для фінансового прогнозування, де значні відхилення можуть суттєво викривити аналітичні висновки. Якщо RMSE значно перевищує MAE, це може свідчити про наявність окремих великих помилок або аномальних періодів, які модель передбачає недостатньо добре.

Ще однією популярною метрикою є середня абсолютна відсоткова похибка – MAPE. Вона показує середній розмір похибки у відсотках від фактичного значення:

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{y_t - \widehat{y}_t}{y_t}\right|.$$

Перевага MAPE полягає в тому, що вона дозволяє інтерпретувати похибку у відсотках, незалежно від масштабу початкових даних. Наприклад, значення MAPE на рівні 5% означає, що прогноз у середньому відхиляється від фактичних значень приблизно на 5%. Це робить метрику зручною для порівняння моделей і пояснення результатів. Однак MAPE має суттєве обмеження: якщо фактичні значення близькі до нуля або дорівнюють нулю, метрика може ставати нестабільною або взагалі непридатною. Тому її необхідно застосовувати обережно.

Для усунення частини недоліків MAPE іноді використовується симетрична середня абсолютна відсоткова похибка – sMAPE. Вона враховує не лише фактичне, а й прогнозоване значення в знаменнику:

$$sMAPE = \frac{100\%}{n}\sum_{t=1}^{n}\frac{\left|y_t - \widehat{y}_t\right|}{(\left|y_t\right| + \left|\widehat{y}_t\right|)/2}.$$

sMAPE дозволяє частково зменшити проблему асиметричності відсоткових похибок. Проте ця метрика також має обмеження і може бути складнішою для інтерпретації. У фінансових задачах її доцільно використовувати як додатковий показник, а не як єдиний критерій вибору моделі.

Окрему увагу в літературі приділяють масштабованим метрикам. Hyndman та Koehler [10] зазначають, що традиційні метрики точності прогнозування не завжди коректно працюють для порівняння різних часових рядів або наборів даних. Саме тому вони пропонують використовувати MASE – середню абсолютну масштабовану похибку. Ця метрика порівнює похибку моделі з похибкою наївного прогнозу. У спрощеному вигляді її можна подати так:

$$MASE = \frac{MAE_{model}}{MAE_{naive}}.$$

Якщо значення MASE менше 1, це означає, що модель прогнозує краще, ніж наївний підхід. Якщо MASE більше 1, то модель поступається простому базовому прогнозу. Така інтерпретація є зручною, оскільки дозволяє оцінити практичну доцільність використання складнішої моделі.

У задачах прогнозування часових рядів доцільно оцінювати моделі за кількома метриками, оскільки кожна з них по-різному відображає похибку прогнозу. MAE показує середній абсолютний розмір помилки, RMSE сильніше реагує на великі відхилення, MAPE дає відсоткову інтерпретацію, а MASE дозволяє порівняти модель із базовим прогнозом [10]. Тому використання кількох показників дає повніше уявлення про якість моделі.

Для фінансово-бюджетних даних вибір метрик має особливе значення. MAE та RMSE є корисними для аналізу абсолютних обсягів видатків, оскільки показують похибку в одиницях самого показника. Для часток або відсоткових значень доцільно додатково використовувати MAPE або sMAPE. Водночас потрібно враховувати масштаб бюджетних категорій, адже однакова абсолютна похибка може мати різне значення для великих і малих категорій.

Також важливо враховувати аномалії. Різкі стрибки або падіння можуть суттєво збільшувати RMSE, тому погіршення метрик потрібно доповнювати змістовним аналізом причин відхилень. У фінансових рядах такі аномалії можуть бути пов’язані з кризами, політичними рішеннями або змінами бюджетної структури.

Оцінювання прогнозу в часових рядах має зберігати часовий порядок даних. На відміну від звичайних задач машинного навчання, дані не можна випадково перемішувати: модель навчається на попередніх періодах і перевіряється на наступних [1]. Це дозволяє відтворити реальну ситуацію прогнозування, коли майбутні значення ще невідомі.

У межах цієї роботи метрики якості використовуються для порівняння класичних статистичних моделей, декомпозиційних підходів, моделей

машинного навчання та нейронних мереж. Метою є не лише вибір моделі з найменшою похибкою, а й визначення того, який підхід краще підходить для фінансово-бюджетних часових рядів. Якщо складна модель лише незначно покращує результат порівняно з простою, доцільність її використання потребує додаткового обґрунтування.

Таким чином, метрики якості є важливим інструментом перевірки ефективності моделей прогнозування. Вони дозволяють кількісно оцінити похибку, порівняти моделі та визначити надійність прогнозу. Для фінансового прогнозування важливо поєднувати кількісну оцінку з аналізом аномалій і практичною інтерпретацією результатів.

1.5 Сучасні методи прогнозування часових рядів

Сучасні методи прогнозування часових рядів значною мірою пов’язані з розвитком машинного навчання та нейронних мереж. На відміну від класичних статистичних моделей, таких як AR, ARMA або ARIMA, ці підходи не завжди потребують жорсткого припущення про лінійність, стаціонарність або просту структуру залежностей у даних. Їхня головна перевага полягає в здатності виявляти складні нелінійні закономірності, взаємодії між ознаками та приховані патерни, які можуть бути непомітними для традиційних моделей.

У задачах прогнозування часових рядів сучасні методи зазвичай застосовуються у двох основних підходах. Перший підхід полягає в перетворенні часового ряду на табличний набір даних, де цільовою змінною є майбутнє значення показника, а вхідними ознаками є лагові значення, ковзні середні, темпи приросту, трендові характеристики та інші похідні показники. У такому випадку задача прогнозування розглядається як задача регресії. Другий підхід передбачає використання моделей, які безпосередньо

працюють із послідовностями, зокрема рекурентних нейронних мереж, LSTM, GRU або інших архітектур, орієнтованих на часові залежності.

Одним із найважливіших класів сучасних моделей машинного навчання є градієнтний бустинг. Його теоретичну основу сформулював Friedman [7], який розглядав бустинг як послідовне наближення невідомої функції за допомогою слабких моделей. У задачах регресії такими слабкими моделями найчастіше є дерева рішень. Ідея градієнтного бустингу полягає в тому, що моделі будуються не незалежно одна від одної, а послідовно: кожна наступна модель намагається виправити помилки попередніх. Завдяки цьому ансамбль поступово покращує якість прогнозування.

У загальному вигляді прогноз градієнтного бустингу можна подати як суму прогнозів окремих базових моделей:

$$\widehat{y} = F_M(x) = \sum_{m=1}^{M} \gamma_m h_m(x),$$

де $h_m(x)$ – окрема базова модель, найчастіше дерево рішень, γ_m – вага цієї моделі, а M – кількість моделей в ансамблі. Кожне наступне дерево будується таким чином, щоб зменшити похибку попередньої сукупності моделей. Саме ця послідовність корекції помилок робить градієнтний бустинг потужним інструментом для прогнозування.

На практиці одним із найпоширеніших алгоритмів градієнтного бустингу є XGBoost, запропонований Chen та Guestrin [8]. XGBoost поєднує градієнтний бустинг дерев рішень із регуляризацією, що дозволяє зменшити ризик перенавчання. Крім того, ця модель ефективно працює з табличними даними, підтримує обробку пропущених значень, дозволяє враховувати нелінійні залежності та взаємодії між ознаками. У задачах прогнозування часових рядів XGBoost часто використовується після попереднього формування ознак на основі часового ряду.

Для фінансово-бюджетних даних градієнтний бустинг є особливо корисним, оскільки такі дані можуть містити складні й нелінійні залежності. Наприклад, майбутнє значення певної категорії федеральних видатків може

залежати не лише від значення попереднього року, а й від середньої динаміки за кілька років, темпу приросту, відхилення від тренду або попередніх аномальних періодів. Класична ARIMA-модель описує залежність переважно через лінійну структуру, тоді як XGBoost може виявляти складніші зв'язки між сформованими ознаками.

Водночас застосування градієнтного бустингу до часових рядів має певні обмеження. Ця модель не "розуміє" час безпосередньо, тому дослідник має самостійно сформувати ознаки, які передають часову структуру. До таких ознак можуть належати лаги:

$$y_{t-1}, y_{t-2}, \ldots, y_{t-p},$$

ковзні середні:

$$MA_k = \frac{1}{k} \sum_{i=0}^{k-1} y_{t-i},$$

темпи приросту, різниці між періодами, індикатори тренду та інші характеристики. Якщо ці ознаки сформовано некоректно, модель може не виявити реальні часові закономірності або, навпаки, перенавчитися на випадкові особливості навчальної вибірки. Ще одним важливим сучасним напрямом є використання нейронних мереж. Нейронні мережі є універсальними апроксиматорами функцій і можуть моделювати складні нелінійні залежності між вхідними та вихідними змінними. У найпростішому випадку для прогнозування можна використовувати багатошарову нейронну мережу прямого поширення. У такій моделі на вхід подаються сформовані ознаки часового ряду, наприклад лагові значення або ковзні середні, а на виході отримується прогноз майбутнього значення. Однак звичайні нейронні мережі не мають спеціального механізму для роботи з послідовностями. Тому для часових рядів частіше використовують рекурентні нейронні мережі. Рекурентні мережі відрізняються тим, що враховують не лише поточні вхідні дані, а й інформацію з попередніх кроків. Це дозволяє їм працювати з послідовностями та зберігати певний "контекст" попередніх значень. У загальному вигляді рекурентну модель можна подати так:

$$h_t = f(x_t, h_{t-1}),$$

де h_t – прихований стан моделі в момент часу t, x_t – вхідне значення або вектор ознак, а h_{t-1} – прихований стан попереднього моменту часу. Таким чином, модель може використовувати інформацію з минулих періодів для формування поточного прогнозу.

Проблемою звичайних рекурентних нейронних мереж є складність збереження довгострокових залежностей. Якщо важлива інформація міститься у віддалених попередніх значеннях, класична RNN може втрачати її під час навчання. Для розв’язання цієї проблеми Hochreiter та Schmidhuber [9] запропонували архітектуру LSTM, яка використовує спеціальні механізми пам’яті для збереження, оновлення або відкидання інформації.

LSTM є популярною в задачах прогнозування часових рядів, оскільки здатна враховувати довші часові залежності. У фінансовому прогнозуванні це може бути корисним, коли поточна динаміка показника залежить не лише від останніх значень, а й від триваліших попередніх тенденцій.

Поряд із LSTM застосовуються GRU [63]. Ця архітектура має подібну логіку, але є простішою та містить менше параметрів. У деяких задачах GRU може забезпечувати подібну якість прогнозування за меншої складності моделі, що важливо для невеликих наборів даних.

Останніми роками також використовуються моделі з механізмом уваги та трансформери. Вони можуть визначати, які попередні періоди є найбільш важливими для прогнозу. Проте такі моделі зазвичай потребують великого обсягу даних і складного налаштування, тому їх застосування в роботах з обмеженим часовим рядом має бути додатково обґрунтоване.

У цій роботі важливо враховувати, що дані про федеральні видатки мають річну структуру та відносно невелику кількість спостережень. Тому надто складні нейронні мережі можуть не мати достатньо даних для якісного навчання, а простіші моделі можуть давати стабільніші та краще інтерпретовані результати.

Саме тому доцільно порівнювати сучасні методи з класичними моделями. ARIMA може використовуватися як базова статистична модель, Prophet – для аналізу тренду, XGBoost – для роботи зі сформованими ознаками, а LSTM – як приклад нейромережевого підходу. Таке порівняння дозволяє оцінити, чи справді складні методи дають перевагу для конкретного набору даних.

Важливою проблемою сучасних методів є ризик перенавчання. Воно виникає тоді, коли модель добре запам’ятовує навчальні дані, але погано працює на нових спостереженнях. Для часових рядів це особливо актуально, оскільки кількість даних часто обмежена, а часовий порядок не дозволяє випадково перемішувати спостереження.

Ще однією важливою характеристикою є інтерпретованість. Градієнтний бустинг і нейронні мережі можуть мати високу точність, але їх складніше пояснити. Тому їх доцільно поєднувати з методами аналізу важливості ознак, зокрема permutation importance та SHAP [13].

У межах цієї роботи сучасні методи прогнозування використовуються для перевірки можливого покращення якості прогнозу порівняно з класичними моделями, а також для визначення ознак, які найбільше впливають на результат.

Отже, сучасні методи, зокрема XGBoost, LSTM, GRU та трансформерні підходи, розширюють можливості аналізу часових рядів. Водночас їх потрібно застосовувати обережно, особливо за обмеженого обсягу даних. Для фінансово-бюджетного прогнозування доцільним є порівняльний підхід, за якого сучасні та класичні моделі оцінюються за однаковими метриками й доповнюються інструментами інтерпретації результатів.

1.6 Методи аналізу важливості ознак

Після побудови моделі прогнозування важливо не лише оцінити її точність, а й визначити, які ознаки найбільше впливають на результат. Це особливо актуально для моделей машинного навчання, зокрема випадкових лісів, градієнтного бустингу та нейронних мереж, які часто розглядаються як моделі типу «чорної скриньки». У фінансовому прогнозуванні результат має бути не лише точним, а й зрозумілим для подальшого аналізу.

У задачах прогнозування часових рядів ознаками можуть бути лагові значення, ковзні середні, темпи приросту, трендові характеристики, різниці між періодами та інші похідні змінні. Для прогнозування федеральних видатків це можуть бути значення витрат за попередні роки, середні значення за кілька періодів або показники довгострокового тренду. Аналіз важливості ознак дозволяє визначити, які з них справді впливають на прогноз.

Одним із простих підходів є внутрішня важливість ознак, яка використовується в деревах рішень, випадкових лісах і градієнтному бустингу. Вона показує, як часто ознака використовується для поділу даних і наскільки вона зменшує похибку моделі [11]. Проте такий підхід може залежати від структури моделі та не завжди прямо пояснює вплив ознаки на прогноз.

Більш універсальним методом є permutation importance. Його суть полягає в тому, що значення певної ознаки випадково перемішуються, після чого повторно оцінюється якість моделі. Якщо похибка суттєво зростає, ознака вважається важливою [12]. Цей метод можна застосовувати до різних моделей, зокрема регресії, випадкових лісів, градієнтного бустингу та нейронних мереж.

Для фінансового прогнозування permutation importance корисний тим, що дозволяє оцінити роль сформованих часових ознак. Наприклад, можна визначити, чи важливішим є лагове значення за попередній рік, ковзне середнє або темп приросту. Водночас результати потрібно інтерпретувати

обережно, оскільки сильно пов’язані між собою ознаки можуть частково замінювати одна одну.

Більш розвиненим методом пояснення моделей є SHAP, який базується на значеннях Шеплі з теорії ігор [13] (див. також [44, 61]). Він дозволяє оцінити внесок кожної ознаки в конкретний прогноз. Перевагою SHAP є можливість аналізу як загальної важливості ознак для всієї моделі, так і пояснення окремих прогнозних значень.

SHAP є особливо корисним для моделей градієнтного бустингу, зокрема XGBoost, оскільки такі моделі можуть мати складну внутрішню структуру [8, 13]. За допомогою SHAP можна визначити, які ознаки збільшують або зменшують прогнозоване значення, що робить результати моделі більш прозорими.

Водночас аналіз важливості ознак не слід ототожнювати з причинно-наслідковим аналізом. Якщо певна ознака є важливою для моделі, це означає, що вона допомагає покращити прогноз, але не обов’язково є реальною причиною зміни показника. Для федеральних видатків це особливо важливо, оскільки бюджетна динаміка залежить від багатьох зовнішніх економічних і політичних факторів.

Таким чином, аналіз важливості ознак є важливою частиною сучасного прогнозування часових рядів. Внутрішня важливість, permutation importance та SHAP дозволяють не лише оцінити якість моделі, а й пояснити, які характеристики історичних даних найбільше впливають на прогноз федеральних видатків.

Висновки до розділу 1

У результаті розгляду теоретичних основ прогнозування часових рядів встановлено, що ця задача включає не лише побудову моделі, а й попередній

аналіз даних, виявлення трендів, формування ознак, оцінювання якості прогнозу та інтерпретацію результатів. Оскільки часові ряди мають впорядковану структуру, для їх аналізу необхідно використовувати методи, які враховують часову залежність.

Огляд літератури показав, що класичні моделі AR, MA, ARMA та ARIMA залишаються важливою основою прогнозування. Вони мають зрозумілу структуру та можуть використовуватися як базові моделі для порівняння, однак переважно орієнтовані на лінійні залежності й потребують урахування стаціонарності ряду.

Фінансово-бюджетні часові ряди мають низку особливостей: тренди, структурні злами, аномальні значення та періоди нестабільності. У випадку федеральних видатків США дані відображають не лише економічну динаміку, а й зміну пріоритетів державної політики, тому потребують змістовної інтерпретації.

Сучасні методи, зокрема XGBoost, градієнтний бустинг, нейронні мережі та LSTM, розширюють можливості прогнозування, оскільки дозволяють враховувати нелінійні залежності й взаємодії між ознаками.

Оцінювання якості прогнозу є необхідним для порівняння моделей. Для цього використовуються метрики MAE, MSE, RMSE, MAPE, sMAPE та MASE, які відображають різні аспекти похибки. Тому доцільно застосовувати кілька метрик одночасно.

Окреме значення має аналіз важливості ознак. Методи permutation importance та SHAP дозволяють визначити, які лагові значення, ковзні середні, темпи приросту й трендові характеристики найбільше впливають на прогноз. Це підвищує прозорість моделей і аналітичну цінність результатів.

Таким чином, теоретичний аналіз підтверджує доцільність комплексного підходу до прогнозування фінансово-бюджетних часових рядів. Поєднання класичних моделей дозволяє оцінити надійність моделей та визначити ключові чинники динаміки показника.

РОЗДІЛ 2 ЗАСТОСУВАННЯ СУЧАСНИХ МЕТОДІВ ПРОГНОЗУВАННЯ ДЛЯ АНАЛІЗУ БЮДЖЕТНИХ ВИТРАТ

У цьому розділі розглядається практичне застосування методів прогнозування та інтерпретації моделей для аналізу бюджетних витрат. На відміну від попереднього розділу, основна увага приділяється прикладному використанню моделей для дослідження динаміки бюджетних витрат.

Мета цього етапу полягає не лише в побудові прогнозу, а й у поясненні чинників, що впливають на результат моделювання. Для цього використовуються модель Prophet, метод permutation importance та SHAP-аналіз.

Prophet застосовується для прогнозування часових рядів із трендом і можливими змінами динаміки. Це важливо для бюджетних даних, оскільки федеральні видатки можуть змінюватися під впливом економічних процесів, політичних рішень, криз або зміни державних пріоритетів.

Permutation importance використовується для оцінювання важливості ознак. Метод показує, наскільки погіршується якість прогнозу після випадкового перемішування значень певної ознаки. Якщо похибка суттєво зростає, така ознака має значний вплив на результат.

SHAP-аналіз дає змогу детальніше пояснити роботу моделі, оцінюючи внесок кожної ознаки як у загальну поведінку моделі, так і в окремі прогнозні значення. Це дозволяє зрозуміти, які характеристики даних сприяють зростанню або зменшенню прогнозованого показника.

Таким чином, розділ поєднує прогнозування та інтерпретацію результатів. Такий підхід дозволяє не лише отримати прогноз бюджетних витрат, а й змістовно пояснити динаміку досліджуваного показника.

2.1 Prophet

Prophet є сучасною моделлю прогнозування часових рядів, розробленою для роботи з даними, які мають трендову структуру, сезонність, а також можливі різкі зміни динаміки. У межах цієї роботи Prophet використовується як інструмент для прогнозування бюджетних витрат і аналізу загальної тенденції їх зміни в часі.

На відміну від класичних моделей ARIMA, які значною мірою спираються на стаціонарність ряду та авторегресійну структуру, Prophet використовує декомпозиційний підхід. Це означає, що часовий ряд подається як сукупність кількох компонентів: тренду, сезонності, впливу подій та випадкової похибки. У загальному вигляді модель можна подати так:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t,$$

де $g(t)$ – трендова компонента, $s(t)$ – сезонна компонента, $h(t)$ – вплив свят або спеціальних подій, а ε_t – випадкова похибка.

У задачі аналізу бюджетних витрат основне значення має трендова складова g(t), оскільки вона описує загальний напрям зміни витрат у часі. У спрощеному вигляді тренд можна подати як функцію, що змінюється залежно від часу:

$$g(t) = k(t)t + m(t),$$

де $k(t)$ – швидкість зміни тренду, $m(t)$ – зміщення трендової лінії. Якщо в часовому ряді відбувається зміна динаміки, Prophet може враховувати точки зміни тренду, у яких параметри $k(t)$ та $m(t)$ коригуються.

Для прогнозування модель оцінює майбутнє значення показника як:

$$\widehat{y}(t + h) = g(t + h) + s(t + h) + h(t + h),$$

де h – горизонт прогнозування, а $\widehat{y}(t + h)$ – прогнозоване значення бюджетних витрат через h періодів.

У межах цієї роботи сезонна та подієва компоненти мають допоміжне значення, оскільки датасет містить обмежену кількість квартальних спостережень. Тому Prophet використовується насамперед для оцінювання загальної трендової динаміки агрегованого квартального ряду.

Для аналізу бюджетних витрат найважливішою є трендова компонента Prophet, оскільки вона відображає довгострокову зміну показника. Федеральні видатки можуть змінюватися під впливом економічного розвитку, державної політики, соціальних програм, оборонних витрат або кризових періодів. Prophet дозволяє виділити цю тенденцію та використати її для прогнозування [5].

Особливістю Prophet є здатність враховувати зміни тренду. У фінансових даних темп зростання або зниження може змінюватися в окремі періоди, наприклад під час криз, реформ або масштабних державних програм. Такі моменти розглядаються як точки зміни тренду, які модель може автоматично визначати [5]. У межах дослідження Prophet використовується для прогнозування бюджетних витрат на наступні роки. Для цього дані приводяться до потрібного формату: стовпець із датою позначається як ds, а стовпець зі значенням показника – як y. Після навчання на історичних даних модель формує прогноз на заданий горизонт [5].

Перевагою Prophet є зручне графічне подання результатів. Модель дозволяє показати фактичні значення, прогнозну траєкторію та інтервал невизначеності. Це важливо для фінансового прогнозування, оскільки майбутні бюджетні витрати не можуть бути визначені абсолютно точно [1, 5]. Окрім основного прогнозу, Prophet дає змогу аналізувати окремі компоненти моделі: тренд, сезонність і можливі подієві ефекти. Для річних бюджетних даних сезонність може бути менш суттєвою, тому основну увагу доцільно приділяти трендовій компоненті та змінам динаміки [5]. Застосування Prophet до федеральних видатків дозволяє оцінити загальний напрям розвитку показника. Якщо прогнозна лінія зростає, це свідчить про збереження тенденції до збільшення витрат; якщо стабілізується або уповільнюється –

про можливу зміну динаміки. Водночас модель не враховує майбутні політичні чи економічні події, яких не було в історичних даних. Разом із тим Prophet має певні обмеження. Його точність може знижуватися за невеликої кількості спостережень або значної залежності показника від зовнішніх факторів. Тому результати Prophet доцільно порівнювати з іншими моделями та оцінювати за метриками якості прогнозу [10].

Отже, Prophet є зручним інструментом для прогнозування та візуального аналізу бюджетних витрат. Він дозволяє виділити довгостроковий тренд, врахувати зміни динаміки та сформувати основу для подальшого порівняння з іншими методами прогнозування [5].

2.2 Permutation Importance

Permutation Importance є методом оцінювання важливості ознак, який використовується для пояснення роботи моделей машинного навчання. У межах цієї роботи він застосовується після побудови прогнозної моделі, щоб визначити, які характеристики бюджетного часового ряду найбільше впливають на результат прогнозування [12].

Основна ідея методу полягає в тому, що значення окремої ознаки випадково перемішуються, після чого повторно оцінюється якість моделі. Якщо після цього похибка суттєво зростає, ознака вважається важливою. Якщо якість майже не змінюється, її вплив на прогноз є незначним [12].

Для прогнозування бюджетних витрат часовий ряд попередньо перетворюється на набір ознак. До них можуть належати лагові значення, ковзні середні, темпи приросту, різниці між роками та трендові індикатори. Наприклад, для прогнозування федеральних видатків можуть використовуватися значення за попередній рік, середнє за кілька років або темп зміни витрат.

Практичне застосування Permutation Importance передбачає кілька етапів: формування ознак, поділ даних на навчальну й тестову частини з урахуванням часової структури, навчання моделі, оцінювання базової якості прогнозу та почергове перемішування кожної ознаки. Зміна похибки після перемішування показує важливість відповідної ознаки.

Формально важливість ознаки x_j за методом Permutation Importance можна визначити як зміну якості моделі після випадкового перемішування значень цієї ознаки:

$$PI_j = L(y, \hat{y}_{perm(j)}) - L(y, \hat{y}),$$

де PI_j – важливість j-ї ознаки, L – функція втрат або метрика похибки, y – фактичні значення, $\hat{y}$ – прогноз моделі на початкових даних, $\hat{y}_{perm(j)}$ – прогноз моделі після перемішування значень ознаки x_j.

Якщо після перемішування ознаки значення похибки суттєво зростає, тобто

$$PI_j > 0,$$

то ознака є важливою для моделі. Якщо ж

$$PI_j \approx 0$$

то перемішування цієї ознаки майже не впливає на якість прогнозу, а її роль у моделі є незначною.

Результати методу зазвичай подаються у вигляді таблиці або графіка, де ознаки розташовуються за рівнем важливості. Це дозволяє визначити, чи модель більше спирається на попередні значення, довгострокову тенденцію або темпи зміни показника.

Водночас результати Permutation Importance потрібно інтерпретувати обережно. Висока важливість ознаки не означає наявності прямого причинно-наслідкового зв’язку, а лише свідчить про її корисність для прогнозування. Також метод може давати занижені оцінки для ознак, які сильно корелюють між собою.

Крім того, важливість ознак залежить від обраної метрики якості. Якщо використовується MAE, оцінюється зміна середньої абсолютної похибки, а якщо RMSE – метод сильніше реагує на великі помилки [10].

У межах цієї роботи Permutation Importance використовується як проміжний інструмент між прогнозуванням і глибшою інтерпретацією результатів. Його результати можуть бути доповнені SHAP-аналізом, який дає змогу пояснити внесок ознак не лише загалом, а й для окремих прогнозних значень [13].

Таким чином, Permutation Importance є простим і універсальним методом аналізу важливості ознак. Він дозволяє визначити, які сформовані часові характеристики найбільше впливають на прогноз федеральних видатків, і робить результати моделювання більш зрозумілими з аналітичної точки зору [12].

2.3 SHAP-аналіз

SHAP-аналіз є одним із сучасних методів пояснення результатів моделей машинного навчання. Його основна мета полягає в тому, щоб визначити, який внесок зробила кожна ознака у формування прогнозу. У межах цієї роботи SHAP використовується для інтерпретації моделі прогнозування бюджетних витрат, зокрема для пояснення того, які характеристики історичних даних найбільше впливали на прогноз федеральних видатків [61].

На відміну від звичайних метрик якості, які лише показують, наскільки точно модель передбачає значення, SHAP дозволяє глибше проаналізувати внутрішню логіку моделі. Це особливо важливо для моделей машинного навчання, таких як градієнтний бустинг або XGBoost, оскільки вони можуть демонструвати високу точність, але залишатися складними для прямої

інтерпретації. SHAP частково розв’язує цю проблему, перетворюючи результат роботи моделі на набір зрозумілих внесків окремих ознак.

Метод SHAP базується на значеннях Шеплі з теорії кооперативних ігор. У такій інтерпретації кожна ознака розглядається як окремий “учасник”, який робить внесок у підсумковий прогноз моделі. Мета методу полягає в тому, щоб справедливо розподілити прогнозоване значення між усіма ознаками, які брали участь у його формуванні. Тобто SHAP показує, наскільки кожна ознака збільшила або зменшила прогноз порівняно з базовим середнім значенням моделі.

У загальному вигляді пояснення прогнозу за допомогою SHAP можна подати так:

$$f(x) = E[f(x)] + \sum_{i=1}^{n} \phi_i,$$

де $f(x)$ – прогноз моделі для конкретного спостереження, $E[f(x)]$ – базове очікуване значення прогнозу, i – SHAP-значення для i-ї ознаки, а n – кількість ознак. Якщо SHAP-значення ознаки є додатним, ця ознака збільшує прогнозоване значення. Якщо SHAP-значення є від’ємним, ознака зменшує прогноз.

Значення Шеплі для j-ї ознаки визначається як середній граничний внесок цієї ознаки в усіх можливих підмножинах ознак:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(m-|S|-1)!}{m!} [f(S \cup \{j\}) - f(S)],$$

де F – множина всіх ознак, S – підмножина ознак без j-ї ознаки, m – загальна кількість ознак, $f(S)$ – прогноз моделі з використанням підмножини ознак S, $f(S \cup \{j\})$ – прогноз після додавання j-ї ознаки.

У задачі аналізу бюджетних витрат це має важливе практичне значення. Наприклад, якщо модель прогнозує зростання певної категорії федеральних видатків, SHAP дозволяє визначити, які саме ознаки сприяли такому прогнозу. Це можуть бути високі значення витрат у попередні роки, позитивний темп приросту, високе ковзне середнє або виражена

довгострокова тенденція. Якщо ж прогнозоване значення знижується або стабілізується, SHAP дозволяє визначити, які ознаки вплинули на це зниження.

SHAP-аналіз використовується для пояснення роботи моделей машинного навчання та може застосовуватися на глобальному й локальному рівнях. Глобальний аналіз показує, які ознаки є найважливішими для моделі загалом, зазвичай через середнє абсолютне SHAP-значення. Локальний аналіз пояснює окремий прогноз і дозволяє визначити, які ознаки вплинули на результат для конкретного періоду [13]. У практичній частині SHAP доцільно застосовувати після навчання моделі машинного навчання, наприклад XGBoost або іншої моделі градієнтного бустингу [8, 13]. Для цього спочатку формуються ознаки часового ряду: лагові значення, ковзні середні, темпи приросту, різниці між роками та трендові характеристики. Після навчання моделі SHAP дозволяє оцінити внесок кожної ознаки у сформований прогноз. Результати SHAP-аналізу зазвичай подаються у вигляді графіків. Summary plot показує загальну важливість ознак і характер їхнього впливу, bar plot відображає середню абсолютну важливість, а force plot або waterfall plot пояснюють окремі прогнозні значення. Це дає змогу визначити, які ознаки підвищують або знижують прогнозований рівень бюджетних витрат [13]. У контексті федеральних видатків SHAP робить результати прогнозування більш змістовними. Наприклад, якщо модель прогнозує зростання витрат, SHAP може показати, що на це вплинули високі значення попередніх років або позитивний тренд. Якщо прогноз нижчий за очікуваний, це може бути пов’язано зі зниженням темпу приросту чи відхиленням від попередньої тенденції.

Порівняно з Permutation Importance, SHAP дає детальніше пояснення. Permutation Importance показує загальну важливість ознак через зміну якості моделі після їх перемішування [12], а SHAP дозволяє оцінити внесок кожної ознаки в окремі прогнози [13]. Тому ці методи доцільно використовувати разом. Водночас SHAP має певні обмеження. Його результати залежать від

якості самої моделі та не доводять причинно-наслідкових зв’язків. Крім того, у часових рядах лагові значення, ковзні середні й трендові ознаки можуть бути пов’язані між собою, тому внесок може розподілятися між кількома подібними ознаками.

Отже, SHAP-аналіз є важливим інструментом інтерпретації моделей прогнозування бюджетних витрат. Він дозволяє оцінити загальну важливість ознак і пояснити окремі прогнозні значення, що підвищує прозорість моделі та обґрунтованість результатів дослідження [13].

2.4 Огляд популярних вибірок даних

Для дослідження аналізу та прогнозування бюджетних витрат важливим етапом є вибір відповідного набору даних. Якість вибірки визначає можливість коректного аналізу, побудови прогнозних моделей, виявлення трендів, аномалій та оцінювання впливу окремих факторів. У фінансовому та бюджетному аналізі часто використовуються відкриті датасети з Kaggle, World Bank, FRED, OECD, а також спеціалізовані набори часових рядів [14–22].

Одним із найбільш придатних для прикладного аналізу є Budget Spending Datasets [14]. Він містить дані про бюджетні витрати за різними категоріями, що дозволяє досліджувати структуру витрат, порівнювати показники, аналізувати відхилення та формувати ознаки для прогнозування. Окрему групу становлять набори державних фінансів. Наприклад, Federal Budget Shares 1941–2020 містить історичні дані про фінансово-бюджетні показники за тривалий період, що дає змогу аналізувати довгострокові зміни бюджетних пріоритетів [16]. Для порівняння планових і фактичних витрат можуть використовуватися вибірки типу Financial Dataset [Expenses]: Budget

vs Actual, які підходять для аналізу відхилень і якості фінансового планування [17].

Важливими джерелами макроекономічних даних є World Bank – Expense (% of GDP), FRED – Federal Government: Current Expenditures та OECD – General Government Spending [18–20]. Вони містять офіційні показники державних витрат, придатні для аналізу бюджетної динаміки, міжкраїнного порівняння та побудови моделей часових рядів.

Окремо можна виділити узагальнені набори часових рядів, зокрема M4 Forecasting Competition Dataset та M5 Forecasting – Accuracy [21, 22]. Вони не є безпосередньо бюджетними, проте корисні для порівняння методів прогнозування та вивчення складних структур часових даних.

Отже, для дослідження бюджетних витрат найбільш релевантними є набори, що містять часову ознаку, категорії витрат, планові й фактичні значення та можливість аналізу відхилень. У межах цієї роботи доцільно використовувати Budget Spending Datasets як основну вибірку [1], а Federal Budget Shares, World Bank, FRED та OECD – як допоміжні джерела для ширшого аналізу фінансово-бюджетних тенденцій [16, 18–20].

Висновки до розділу 2

У другому розділі було розглянуто сучасні підходи до прогнозування та інтерпретації моделей, які можуть бути використані для аналізу бюджетних витрат. Основну увагу приділено моделі Prophet, методу Permutation Importance та SHAP-аналізу. Встановлено, що Prophet є доцільним інструментом для аналізу загальної динаміки часового ряду та виявлення трендової компоненти, тоді як Permutation Importance і SHAP дозволяють пояснити роботу моделей машинного навчання та визначити вплив окремих ознак на прогноз. Також було розглянуто популярні набори даних, які можуть

застосовуватися для аналізу бюджетних і фінансових показників. Визначено, що для практичної частини роботи доцільно використовувати датасет бюджетних витрат, який містить планові та фактичні значення, департаменти, місяці та відхилення. Таким чином, у розділі сформовано методичну основу для подальшого практичного аналізу, побудови моделей прогнозування та інтерпретації отриманих результатів.

РОЗДІЛ 3 МЕТОД АНАЛІЗУ СТРУКТУРИ БЮДЖЕТНИХ ВИТРАТ

На даному етапі дослідження проведено первинний аналіз набору даних Budget Spending Datasets, отриманого з платформи Kaggle. Метою цього етапу є вивчення структури даних, визначення основних характеристик змінних, виявлення можливих проблем у даних, а також формування загального уявлення про поведінку досліджуваного процесу.

3.1 Первинний аналіз датасету

Датасет містить дані про бюджетні витрати у вигляді транзакцій. Основними змінними є дата, категорія витрат і сума витрат, що дозволяє використовувати як статистичні методи, так і методи машинного навчання. На першому етапі було проаналізовано структуру даних, типи змінних і їх розподіл. Ключовою змінною для моделювання є сума витрат, а дата використовується для формування часового ряду. Для подальшого аналізу дані було агреговано за місяцями. Це дозволило зменшити вплив випадкових коливань і сформувати часовий ряд, у якому кожному місяцю відповідає загальна сума витрат (табл. 3.1).

Особливістю цього набору даних є те, що змінна Month містить лише назву місяця без зазначення року. Тому датасет не є повноцінним багаторічним часовим рядом у класичному вигляді. Його можна використовувати для аналізу сезонної або місячної структури витрат, порівняння департаментів, оцінки перевищення чи економії бюджету, а також як основу для формування агрегованого ряду за місяцями. Для подальшого

прогнозування цю особливість необхідно врахувати, оскільки відсутність року обмежує можливість аналізу довгострокового тренду.

Таблиця 3.1 – Структура датасету

Змінна	Тип даних	Опис
Month	категоріальна	місяць, до якого належить запис
Department	категоріальна	департамент, для якого зафіксовано витрати
Budgeted Amount	числова	запланований обсяг витрат
Actual Spending	числова	фактичний обсяг витрат
Variance	числова	відхилення фактичних витрат від запланованих

Перевірка якості даних показала, що у датасеті немає пропущених значень. Також не виявлено повністю дубльованих рядків. Було додатково перевірено правильність розрахунку змінної Variance. Для всіх 5000 записів вона дорівнює різниці між фактичними та запланованими витратами:

$$Variance = Actual\ Spending - Budgeted\ Amount.$$

Це свідчить про внутрішню узгодженість числових змінних у датасеті (табл. 3.2).

Таблиця 3.2 – Перевірка якості даних

Показник	Значення
Кількість записів	5000
Кількість змінних	5
Пропущені значення	0
Дублікати рядків	0
Кількість місяців	12
Кількість департаментів	6
Помилки у розрахунку Variance	0

На наступному етапі було проаналізовано основні числові характеристики планових витрат, фактичних витрат і відхилення. Загальна сума запланованих витрат у датасеті становить 62 212 846, тоді як загальна

сума фактичних витрат становить 62 308 993. Отже, сумарне відхилення дорівнює 96 147, тобто фактичні витрати загалом перевищують заплановані приблизно на 0,15%. Це свідчить про те, що в цілому бюджет у датасеті є досить збалансованим, хоча на рівні окремих записів, місяців і департаментів можуть спостерігатися суттєві відхилення (табл. 3.3).

Таблиця 3.3 – Описова статистика числових змінних

Показник	**Budgeted Amount**	**Actual Spending**	**Variance**
Сума	62 212 846	62 308 993	96 147
Середнє значення	12 442,57	12 461,80	19,23
Медіана	12 400,50	12 504,00	10,00
Стандартне відхилення	4 333,26	4 619,23	1 725,66
Мінімум	5 005	2 316	-2 999
Максимум	19 985	22 854	2 999

Аналіз описової статистики показує, що середнє значення фактичних витрат є трохи вищим за середнє значення запланованих витрат. Різниця між ними невелика, що підтверджує загальну збалансованість бюджету. Водночас стандартне відхилення фактичних витрат є вищим, ніж у запланованих, тобто фактичні витрати мають дещо більшу варіативність. Це означає, що реальні витрати частіше відхиляються від середнього рівня, ніж планові значення. Змінна Variance має майже симетричний розподіл навколо нуля: середнє відхилення становить 19,23, а медіана - 10,00. Це означає, що в середньому фактичні витрати лише незначно перевищують заплановані. Водночас мінімальне та максимальне значення відхилення становлять відповідно -2 999 і 2 999, тобто в окремих записах спостерігаються значні як перевитрати, так і економія бюджету.

Було також проаналізовано співвідношення випадків перевищення та недовикористання бюджету. У 2507 записах фактичні витрати перевищують заплановані, у 2491 записі фактичні витрати є нижчими за планові, і лише у 2 записах фактичні витрати точно дорівнюють запланованим. Отже, частка

перевитрат становить приблизно 50,14%, а частка економії - 49,82%. Це підтверджує, що датасет є збалансованим за напрямом відхилень (табл. 3.4).

Таблиця 3.4 – Співвідношення перевитрат і економії бюджету

Тип відхилення	Кількість записів	Частка, %
Фактичні витрати більші за планові	2507	50,14
Фактичні витрати менші за планові	2491	49,82
Фактичні витрати дорівнюють плановим	2	0,04

Додатково було перевірено наявність викидів за правилом міжквартильного розмаху. Для змінних Budgeted Amount, Actual Spending і Variance статистичних викидів за цим критерієм не виявлено. Це означає, що хоча в датасеті є окремі великі позитивні та негативні відхилення, вони не виходять за межі очікуваного діапазону варіації даних. Тому такі значення доцільно не вилучати, а враховувати як природну частину бюджетного процесу.

Таблиця 3.5 – Кількість записів за департаментами

Департамент	Кількість записів
Finance	787
HR	859
IT	849
Marketing	806
Operations	853
Sales	846

У таблиці 3.5 аналіз категоріальної змінної Department показав, що в датасеті представлено шість департаментів: Finance, HR, IT, Marketing, Operations і Sales. Кількість записів за департаментами є відносно рівномірною. Найбільше записів має департамент HR - 859, а найменше

Finance - 787. Така структура дозволяє порівнювати департаменти між собою без суттєвого перекосу в кількості спостережень.

Найбільший обсяг фактичних витрат припадає на департамент Operations - 10 859 761, що становить приблизно 17,43% від усіх фактичних витрат. Далі йдуть департаменти Sales із фактичними витратами 10 649 297 та IT із фактичними витратами 10 531 198. Найменший обсяг фактичних витрат має департамент Finance - 9 860 113, хоча цей департамент має найменшу кількість записів (табл. 3.6).

Таблиця 3.6 – Агрегований аналіз витрат за департаментами

Департамент	Заплановано	Фактично	Відхилення	Факт до плану, %	Частка фактичних витрат, %
Finance	9 794 075	9 860 113	66 038	100,67	15,82
HR	10 445 694	10 430 477	-15 217	99,85	16,74
IT	10 511 713	10 531 198	19 485	100,19	16,90
Marketing	10 061 169	9 978 147	-83 022	99,17	16,01
Operations	10 835 604	10 859 761	24 157	100,22	17,43
Sales	10 564 591	10 649 297	84 706	100,80	17,09

За результатами аналізу видно, що найбільше позитивне відхилення має департамент Sales: фактичні витрати перевищують заплановані на 84 706. Це означає, що саме Sales дав найбільший внесок у загальне перевищення бюджету. Також значне позитивне відхилення має департамент Finance - 66 038. Найбільша економія бюджету спостерігається в департаменті Marketing, де фактичні витрати є нижчими за заплановані на 83 022. Таким чином, на рівні департаментів структура відхилень є нерівномірною: одні підрозділи систематично формують перевитрати, а інші - економію.

На наступному етапі було проаналізовано структуру даних за місяцями. Кількість записів за місяцями також є відносно рівномірною, хоча певні відмінності присутні. Найбільше записів припадає на червень - 454, а найменше на липень - 377. Це означає, що місячна структура датасету є достатньо збалансованою для порівняння, але під час інтерпретації

агрегованих сум необхідно враховувати різну кількість записів у кожному місяці (табл. 3.7).

Таблиця 3.7 – Кількість записів за місяцями

Місяць	Кількість записів
Jan	391
Feb	417
Mar	429
Apr	427
May	426
Jun	454
Jul	377
Aug	435
Sep	415
Oct	432
Nov	395
Dec	402

Таблиця 3.8 – Агрегований аналіз витрат за місяцями

Місяць	Кількість записів	Заплановано	Фактично	Відхилення	Факт до плану, %
Jan	391	5 023 989	4 983 374	-40 615	99,19
Feb	417	5 206 232	5 242 307	36 075	100,69
Mar	429	5 306 621	5 262 657	-43 964	99,17
Apr	427	5 344 594	5 343 198	-1 396	99,97
May	426	5 201 767	5 231 244	29 477	100,57
Jun	454	5 637 882	5 620 191	-17 691	99,69
Jul	377	4 493 657	4 517 320	23 663	100,53
Aug	435	5 486 535	5 515 506	28 971	100,53
Sep	415	5 379 289	5 429 963	50 674	100,94
Oct	432	5 211 074	5 198 227	-12 847	99,75
Nov	395	4 789 089	4 801 038	11 949	100,25
Dec	402	5 132 117	5 163 968	31 851	100,62

У таблиці 3.8 гregований аналіз за місяцями показав, що найбільший обсяг фактичних витрат зафіксовано у червні - 5 620 191, що становить 9,02%

від загального обсягу фактичних витрат. Високі значення також спостерігаються у серпні - 5 515 506 та вересні - 5 429 963. Найменший обсяг фактичних витрат припадає на липень - 4 517 320, що частково пояснюється найменшою кількістю записів у цьому місяці.

Найбільше позитивне місячне відхилення спостерігається у вересні, де фактичні витрати перевищують заплановані на 50 674. Також суттєві перевищення бюджету зафіксовано у лютому - 36 075, грудні - 31 851, травні - 29 477 та серпні - 28 971. Найбільша економія бюджету спостерігається у березні, де фактичні витрати є нижчими за заплановані на 43 964, а також у січні - 40 615.

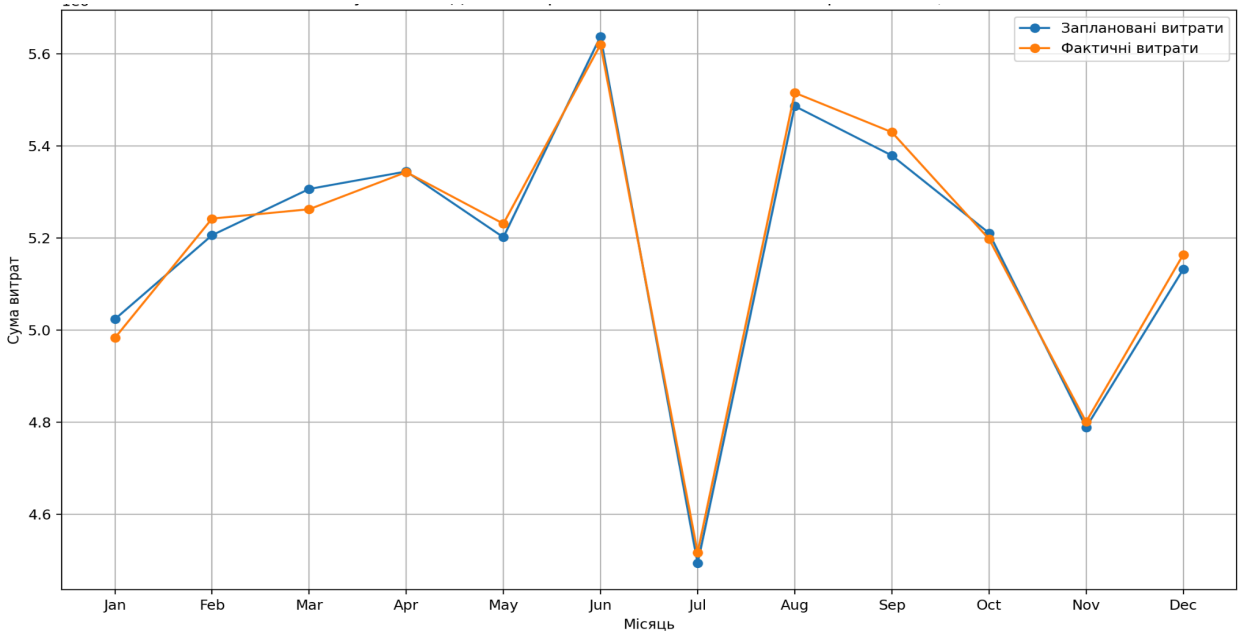

Рисунок 3.1 – Динаміка фактичних і запланованих витрат за місяцями протягом року [52]

На рисунку 3.1 графіку динаміки витрат за місяцями доцільно відобразити дві лінії: суму запланованих витрат і суму фактичних витрат. Такий графік дозволяє візуально порівняти, у які місяці фактичні витрати перевищують план, а в які спостерігається економія. За результатами аналізу можна зробити висновок, що загальна динаміка запланованих і фактичних

витрат є дуже близькою, а відхилення між ними не мають стабільного односпрямованого характеру.

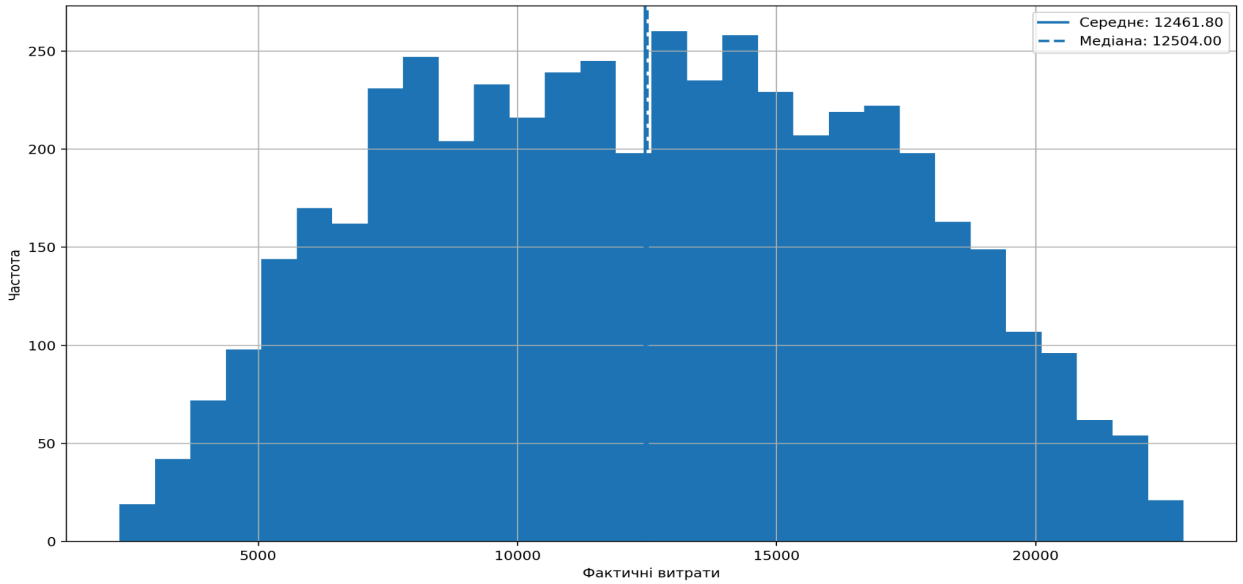

Рисунок 3.2 – Розподіл фактичних витрат [52]

Як показує рисунок 3.2, розподіл фактичних витрат є відносно рівномірним у межах наявного діапазону значень. Мінімальне значення фактичних витрат становить 2 316, максимальне - 22 854, а середнє - 12 461,80. Медіана фактичних витрат дорівнює 12 504, тобто вона дуже близька до середнього значення. Це свідчить про відсутність суттєвої асиметрії в розподілі. Коефіцієнт асиметрії для фактичних витрат становить приблизно 0,03, що також підтверджує майже симетричний характер розподілу.

Структура фактичних витрат за департаментами є відносно збалансованою. Жоден департамент не домінує абсолютно, оскільки частки фактичних витрат коливаються приблизно від 15,82% до 17,43%. Найбільшу частку має Operations, найменшу - Finance. Це означає, що загальний бюджет розподілений між департаментами досить рівномірно, хоча відхилення від плану в різних департаментах суттєво відрізняються (рис. 3.3).

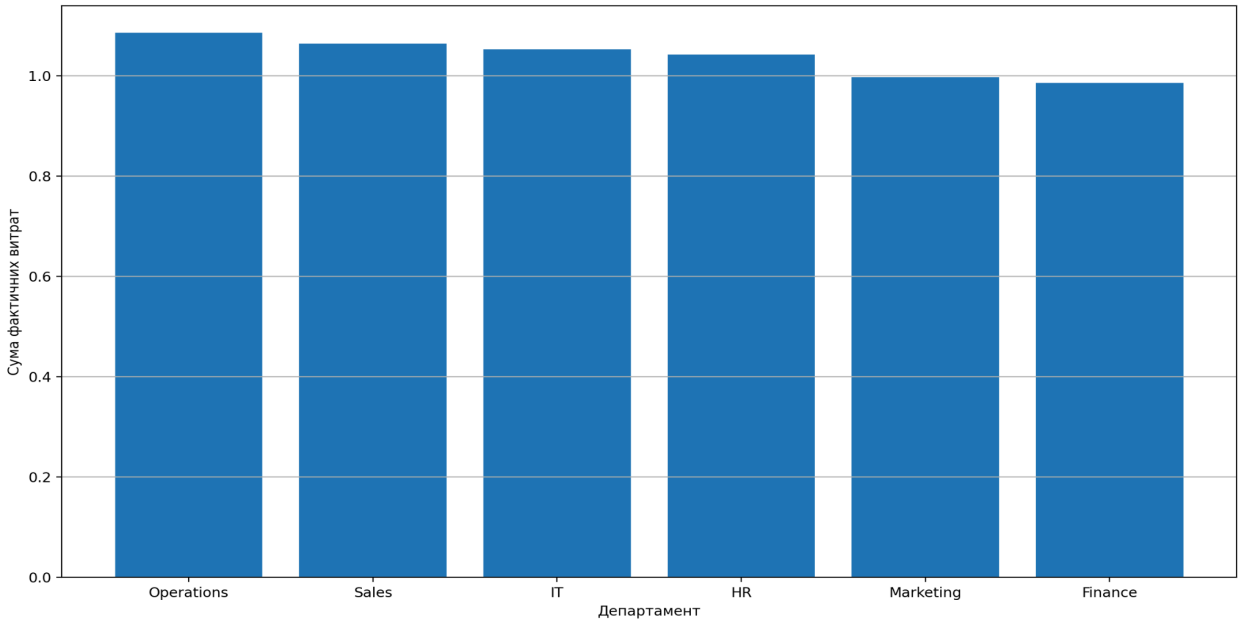

Рисунок 3.3 – Структура фактичних витрат за департаментами [52]

Рисунок 3.3 ілюструє, що окремо було проаналізовано комбінації департаментів і місяців, щоб визначити, де саме виникали найбільші локальні перевитрати та економія. Найбільше позитивне відхилення зафіксовано для департаменту Finance у вересні: фактичні витрати перевищили заплановані на 34 374. Також значні перевищення бюджету спостерігаються для Sales у лютому та травні, а також для IT у липні (табл. 3.9).

Таблиця 3.9 – Кореляція між числовими змінними

Змінні	Кореляція
Budgeted Amount - Actual Spending	0,928
Budgeted Amount - Variance	-0,028
Actual Spending - Variance	0,347

Таким чином, запланований бюджет є основним фактором рівня фактичних витрат, однак перевищення або економія не завжди прямо залежать від його розміру. Тому змінна Budgeted Amount може бути важливою ознакою для прогнозування фактичних витрат, тоді як для аналізу відхилень доцільно враховувати також департамент, місяць та попередню динаміку витрат [1].

Результати первинного аналізу показали, що датасет є придатним для подальшого моделювання: він не містить пропущених значень і дубльованих рядків, а розрахунок відхилень є коректним. Дані охоплюють 12 місяців і 6 департаментів, що дає змогу виконувати часовий та організаційний аналіз. Загалом фактичні витрати перевищують заплановані лише на 0,15 %, що свідчить про майже збалансований бюджет (табл. 2.1).

Аналіз за департаментами та місяцями показав наявність помітних відмінностей. Найбільше перевищення бюджету сформував департамент Sales, тоді як найбільша економія спостерігається в Marketing. У місячному розрізі найбільше перевищення зафіксовано у вересні, а найбільша економія – у березні (рисунки 2.1, 2.2).

Отримані результати підтверджують придатність датасету для прогнозування фактичних витрат, аналізу відхилень від плану та виявлення департаментів і місяців із підвищеним ризиком перевитрат. Водночас датасет не містить року, тому для довгострокового прогнозування часових рядів його потрібно доповнити часовою інформацією або використовувати для аналізу місячної структури бюджетних витрат.

3.2 Підготовка даних до моделювання

Після первинного аналізу датасету було виконано підготовку даних до моделювання. Дані приведено до формату, придатного для прогнозування, побудови black-box моделі, використання Permutation Importance та SHAP-аналізу. Початковий датасет містить місяць, департамент, заплановані витрати, фактичні витрати та відхилення. Оскільки рік у даних відсутній, їх було інтерпретовано як один умовний річний цикл і згруповано за кварталами. Це дозволило перейти від помісячного аналізу до

поквартального, зменшити вплив випадкових коливань і підготувати дані до подальшого прогнозування.

На першому етапі місяці було впорядковано за календарною послідовністю, після чого кожному місяцю присвоєно відповідний квартал: I квартал – січень-березень, II – квітень-червень, III – липень-вересень, IV – жовтень-грудень (табл. 3.10).

Таблиця 3.10 – Розподіл місяців за кварталами

Квартал	**Місяці**	**Позначення**
1 квартал	Jan, Feb, Mar	Q1
2 квартал	Apr, May, Jun	Q2
3 квартал	Jul, Aug, Sep	Q3
4 квартал	Oct, Nov, Dec	Q4

Після формування квартальної ознаки було виконано агрегування даних. Для кожного кварталу розраховано суму запланованих витрат, суму фактичних витрат, загальне відхилення, кількість записів і відсоток виконання бюджету. Такий підхід дозволяє перейти від окремих помісячних транзакцій до узагальненого квартального представлення бюджетних витрат.

У загальному вигляді агреговані показники визначаються так:

$$Budgeted_q = \sum BudgetedAmount_q,$$

$$Actual_q = \sum ActualSpending_q,$$

$$Variance_q = Actual_q - Budgeted_q,$$

де q – номер кварталу .

Додатково розраховується відсоток виконання бюджету:

$$BudgetExecution_q = \frac{Actual_q}{Budgeted_q} \cdot 100\%.$$

У результаті формується агрегований квартальний ряд, який використовується як основа для подальшого прогнозування (табл. 3.11).

Таблиця 3.11 – Агреговані показники за кварталами

Квартал	Кількість записів	Заплановано	Фактично	Відхилення	Факт до плану, %
Q1	1237	15 536 842	15 488 338	-48 504	99,69
Q2	1307	16 184 243	16 194 633	10 390	100,06
Q3	1227	15 359 481	15 462 789	103 308	100,67
Q4	1229	15 132 280	15 163 233	30 953	100,20

Отриманий квартальний ряд не розглядається як замкнений цикл. Тобто після Q4 не формується автоматичний перехід до Q1 з використанням майбутньої інформації. Натомість квартали розглядаються як послідовні спостереження, де кожен наступний елемент прогнозується на основі трьох попередніх.

Основна логіка моделювання має вигляд:

$$Q_{t-3}, Q_{t-2}, Q_{t-1}, \rightarrow Q_t,$$

де Q_t – прогнозований квартал, а Q_{t-3}, Q_{t-2}, Q_{t-1} – три попередні квартали, які використовуються як вхідна інформація для моделі. Таким чином, якщо потрібно спрогнозувати значення четвертого кварталу, модель використовує інформацію про перший, другий і третій квартали. У цій постановці не використовується циклічне повернення до початку ряду, тому всі ознаки відповідають лише попереднім доступним періодам (табл. 3.12).

Таблиця 3.12 – Логіка формування прогнозу на основі трьох попередніх кварталів

Прогнозований квартал	**Використані попередні квартали**	**Логіка прогнозування**
Q4	Q1, Q2, Q3	прогноз Q4 за трьома попередніми кварталами
Наступний умовний квартал	Q2, Q3, Q4	прогноз наступного періоду після Q4 за останніми трьома кварталами

Оскільки в датасеті представлено чотири квартали, перший повноцінний прогноз усередині наявного ряду може бути сформований для Q4, оскільки саме для нього доступні три попередні квартали: Q1, Q2 і Q3. Для подальшого прогнозування наступного періоду після Q4 як вхідні значення використовуються Q2, Q3 і Q4. Такий підхід відповідає логіці ковзного вікна, де на кожному кроці вікно зміщується на один квартал уперед.

Для побудови моделей машинного навчання було сформовано лагові ознаки. Вони відображають значення фактичних витрат у трьох попередніх кварталах:

$$lag1_t = Actual_{t-1},$$

$$lag2_t = Actual_{t-2},$$

$$lag3_t = Actual_{t-3}.$$

У цій постановці lag1_actual, lag2_actual і lag3_actual не є циклічними. Вони завжди відповідають реальним попереднім кварталам у послідовності. Це означає, що для прогнозованого кварталу модель не отримує інформацію з майбутнього періоду. Такий підхід дозволяє уникнути витоку даних і забезпечити коректну інтерпретацію важливості ознак.

Цільова змінна має вигляд:

$$y_t = Actual_t,$$

де $Actual_t$ - фактичні витрати прогнозованого кварталу.

Відповідно, вектор вхідних ознак можна записати так:

$$X_t = \{lag1_t, lag2_t, lag3_t, Budgeted_t, Variance_{t-1}, RollingMean3_t\}.$$

Ознака ковзного середнього за три попередні квартали визначається наступним чином:

$$RollingMean3_t = \frac{lag1_t + lag2_t + lag3_t}{3}.$$

Вказана вище формула дозволяє врахувати не лише окремі значення попередніх кварталів, а й загальний середній рівень витрат за останні три періоди.

Для деталізованого аналізу також було сформовано розширений набір даних на рівні «департамент - квартал». У цьому випадку лагові ознаки формуються окремо для кожного департаменту. Тобто для прогнозування витрат певного департаменту в поточному кварталі використовуються витрати цього ж департаменту за три попередні квартали.

Формально це можна записати так:

$$lag1_t = Actual_{t-1},$$

$$lag2_t = Actual_{t-2},$$

$$lag3_t = Actual_{t-3}.$$

Цільова змінна для департаментального рівня має вигляд:

$$y_{d,t} = Actual_{d,t}.$$

де d – департамент, а t – номер квартального спостереження.

Тобто модель прогнозує фактичні витрати конкретного департаменту в поточному кварталі на основі його витрат у трьох попередніх кварталах та додаткових ознак.

Таблиця 3.13 – Підготовлені ознаки для прогнозування

Показник	**Зміст**	**Роль у моделі**
quarter_num	номер кварталу від 1 до 4	врахування квартальної позиції
Department	назва департаменту	категоріальна ознака
budgeted	сума запланованих витрат	пояснювальна ознака
actual	сума фактичних витрат	основний фінансовий показник
variance	різниця між фактом і планом	ознака перевитрат або економії
records	кількість записів у групі	допоміжна ознака масштабу
lag1_actual	фактичні витрати попереднього кварталу	лагова ознака
lag2_actual	фактичні витрати за два квартали до прогнозу	лагова ознака
lag3_actual	фактичні витрати за три квартали до прогнозу	лагова ознака
rolling_3q_actual	середнє за три попередні квартали	ознака короткострокової динаміки
target_actual	фактичні витрати прогнозованого кварталу	цільова змінна

Таблиця 3.13 показує, що частина ознак є категоріальними, а тому для подальшого використання в моделях машинного навчання їх необхідно перетворити у числовий формат. Для змінної Department доцільно застосувати one-hot encoding, тобто створити окремі бінарні змінні для кожного департаменту. Для кварталу можна використовувати числове кодування від 1 до 4.

Окрему увагу було приділено формуванню навчальних і тестових спостережень. Випадкове перемішування кварталів не використовується, оскільки воно порушує часову логіку задачі. Замість цього застосовується послідовний підхід: модель отримує три попередні квартали як вхідну інформацію та прогнозує наступний квартал (табл. 3.14).

Таблиця 3.14 – Схема прогнозування на основі трьох попередніх кварталів

Крок	Вхідні квартали	Прогнозований квартал	Логіка
1	Q1, Q2, Q3	Q4	прогноз за трьома попередніми кварталами
2	Q2, Q3, Q4	наступний квартал	зсув вікна на один квартал уперед

Така схема відповідає принципу ковзного вікна. На кожному наступному кроці найстаріший квартал вилучається з вікна, а найновіший додається. У результаті модель завжди використовує лише ті дані, які були б доступні на момент прогнозування.

У роботі не використовувалося випадкове розбиття даних на навчальну та тестову вибірки, оскільки для часових рядів важливо зберігати послідовність спостережень. Дані було поділено за часовим принципом: квартали Q1–Q3 становлять історичну частину для формування ознак, тобто 75% агрегованих квартальних спостережень, а Q4 використовується як контрольна тестова частина, тобто 25% спостережень.

Стратифікація не застосовувалася, оскільки поділ виконувався не випадково, а за часовою логікою. Метрики MAE, RMSE та MAPE розраховувалися по всій контрольній вибірці Q4 на рівні департаментів, тобто за шістьма прогнозними спостереженнями.

Для моделі Prophet підготовлені квартальні дані додатково приводяться до формату ds і y, де ds - умовна дата початку кварталу, а y - фактичні витрати за квартал. Оскільки в датасеті немає повної часової шкали з роками, квартали можуть бути представлені як послідовні умовні дати. Це потрібно лише для технічної сумісності з Prophet і не означає, що модель отримує додаткову інформацію про реальні роки.

Для black-box моделі використовується табличне представлення даних. Кожен рядок описує прогнозований квартал або комбінацію «департамент - квартал», а вхідними ознаками є три попередні значення витрат, ковзне

середнє, квартальна ознака, департамент і бюджетні показники. Саме на основі цієї моделі надалі застосовуються Permutation Importance та SHAP-аналіз.

Масштабування числових ознак на цьому етапі не є обов'язковим для моделей на основі дерев рішень або градієнтного бустингу, оскільки вони не є чутливими до масштабу змінних. Водночас для лінійних моделей або нейронних мереж масштабування може бути корисним, тому підготовлений набір даних допускає подальшу нормалізацію або стандартизацію за потреби.

У результаті підготовки даних було сформовано два робочі набори. Перший - агрегований квартальний ряд за всіма департаментами, який використовується для загального аналізу динаміки та моделі Prophet. Другий - розширений набір на рівні «департамент - квартал», який використовується для black-box моделювання, Permutation Importance та SHAP-аналізу (табл. 3.15).

Таблиця 3.15 – Підсумок підготовлених наборів даних

Набір даних	**Рівень агрегації**	**Призначення**
Квартальний ряд	квартал	Prophet, загальний аналіз динаміки витрат
Розширений набір	департамент - квартал	black-box модель, Permutation Importance, SHAP
Цільова змінна	прогнозований квартал	фактичні витрати поточного прогнозованого кварталу

Таким чином, підготовка даних до моделювання включала перехід від помісячного представлення до поквартального, формування послідовної структури кварталів, створення лагових ознак за трьома попередніми кварталами та побудову цільової змінної для прогнозованого кварталу. На відміну від циклічного підходу, оновлена схема не використовує замкнений перехід Q4→Q1 як частину одного циклу, а застосовує принцип ковзного

вікна. Це дозволяє уникнути витоку даних, зробити лагові ознаки рівноправними та підготувати датасет до коректного застосування базової, black-box та інтерпретаційних моделей.

3.3 Побудова базової моделі прогнозування

Після підготовки даних було побудовано базову модель прогнозування, яка використовується як контрольний орієнтир для порівняння зі складнішими підходами. Її мета полягає не в досягненні максимальної точності, а у визначенні мінімального рівня якості прогнозу.

У роботі базова модель ґрунтується на використанні трьох попередніх кварталів для прогнозування наступного. Такий підхід зберігає часову логіку та не допускає використання майбутніх значень як вхідних даних:

$$Q_{t-3}, Q_{t-2}, Q_{t-1} \rightarrow Q_t,$$

де Q_t – прогнозований квартал, а $Q_{t-3}, Q_{t-2}, Q_{t-1}$ – три квартали, що передують йому.

Як базову модель використано прогнозування за середнім значенням трьох попередніх кварталів:

$$\widehat{y}_t = \frac{y_{t-1}+y_{t-2}+y_{t-3}}{3},$$

де $\widehat{y}_t$ – прогнозоване значення фактичних витрат у кварталі t,

y_{t-1} – фактичні витрати попереднього кварталу,

y_{t-2} – фактичні витрати за два квартали до прогнозованого,

y_{t-3} – фактичні витрати за три квартали до прогнозованого.

Таким чином, прогноз для наступного кварталу формується як середній рівень витрат за три попередні квартали. Наприклад, для прогнозування четвертого кварталу використовуються значення першого,

другого та третього кварталів. Такий підхід відповідає логіці короткострокового прогнозування, оскільки майбутнє значення оцінюється на основі найближчої історії (табл. 3.16).

Таблиця 3.16 – Логіка побудови базової моделі

Прогнозований квартал	**Використані значення**	**Формула прогнозу**
Q4	Q1, Q2, Q3	$\widehat{Q}4 = (Q1 + Q2 + Q3)\backslash 3$
Наступний квартал після Q4	Q2, Q3, Q4	$\widehat{Q}_{next} = (Q2 + Q3 + Q4)\backslash 3$

На основі агрегованого квартального ряду було сформовано перший прогноз для четвертого кварталу. Для цього використано фактичні витрати трьох попередніх кварталів: Q1, Q2 та Q3 (табл. 3.17).

Таблиця 3.17 – Результат базового прогнозування для агрегованого квартального ряду

Прогнозований квартал	**Фактичні витрати Q1**	**Фактичні витрати Q2**	**Фактичні витрати Q3**	**Прогноз моделі**	**Фактичні витрати прогнозованого кварталу**	**Помилка**
Q4	15 488 338	16 194 633	15 462 789	15 715 253,33	15 163 233	-552 020,33

За результатами видно, що базова модель спрогнозувала витрати четвертого кварталу на рівні 15 715 253,33, тоді як фактичне значення становить 15 163 233. Помилка прогнозу дорівнює -552 020,33. Це означає, що модель завищила очікуваний рівень витрат, оскільки середнє значення трьох попередніх кварталів виявилося більшим за фактичні витрати четвертого кварталу. Для оцінки якості базової моделі було використано стандартні метрики прогнозування: MAE, RMSE та MAPE. Оскільки на агрегованому рівні за наявними чотирма кварталами можна сформувати один

повноцінний прогноз за трьома попередніми кварталами, значення MAE та RMSE збігаються з абсолютною помилкою прогнозу.

Середня абсолютна помилка визначається за формулою:

$$MAE = \left| y_i - \widehat{y}_i \right|.$$

Середньоквадратична помилка:

$$RMSE = \sqrt{(y_i - \widehat{y}_i)^2}.$$

Середня абсолютна відносна помилка:

$$MAPE = \left| \frac{y_i - \widehat{y}_i}{y_i} \right| \cdot 100\%.$$

Отримане значення MAPE на рівні 3,64% свідчить про те, що базова модель дає відносно невелику відносну похибку на агрегованому рівні. Це пояснюється тим, що загальні квартальні витрати за всіма департаментами мають близький масштаб і не демонструють різких розривів. Водночас така модель є досить обмеженою, оскільки вона враховує лише середній рівень попередніх кварталів і не використовує додаткові ознаки, зокрема департамент, планові витрати, відхилення від бюджету або структуру витрат (табл. 3.18).

Таблиця 3.18 – Метрики якості базової моделі для агрегованого квартального ряду

Метрика	Значення
MAE	552 020,33
RMSE	552 020,33
MAPE	3,64%

Далі базову модель було застосовано до розширеного набору даних на рівні «департамент – квартал». У цьому випадку прогнозування виконується окремо для кожного департаменту. Для прогнозу витрат департаменту в четвертому кварталі використовуються фактичні витрати цього ж департаменту у першому, другому та третьому кварталах:

$$\hat{y}_{d,Q4} = \frac{y_{d,Q1}+y_{d,Q2}+y_{d,Q3}}{3},$$

де d - департамент.

Такий підхід дозволяє оцінити, наскільки стабільною є динаміка витрат у межах окремих департаментів. Якщо витрати департаменту змінюються плавно, середнє значення трьох попередніх кварталів може бути достатньо близьким до фактичного значення четвертого кварталу. Якщо ж витрати суттєво коливаються, похибка базової моделі зростає (табл. 3.19).

Таблиця 3.19 – Якість базової моделі на рівні департаментів

Рівень оцінювання	**MAE**	**RMSE**	**MAPE**
Департамент - квартал	267 116,67	296 582,87	10,67%

Порівняно з агрегованим квартальним рядом, якість базової моделі на рівні департаментів є нижчою. Значення MAPE становить 10,67%, що свідчить про більшу мінливість витрат окремих департаментів. Це є очікуваним результатом, оскільки загальне агрегування згладжує коливання, тоді як департаментальний рівень краще відображає локальні зміни в структурі витрат.

Додатково було розраховано якість базової моделі для кожного департаменту окремо (табл. 3.20).

Таблиця 3.20 – Якість базового прогнозування за департаментами

Департамент	**MAE**	**RMSE**	**MAPE**
Finance	227 647,67	227 647,67	8,41%
HR	188 584,33	188 584,33	7,62%
IT	359 814,33	359 814,33	15,93%
Marketing	284 035,67	284 035,67	12,47%
Operations	223 076,33	223 076,33	8,31%
Sales	319 541,67	319 541,67	11,30%

Найнижчу відносну похибку базова модель показала для департаменту HR, де MAPE становить 7,62%, а також для Operations - 8,31% і Finance - 8,41%. Це означає, що для цих департаментів середній рівень витрат за три попередні квартали є відносно прийнятним наближенням до витрат четвертого кварталу.

Найвищу похибку зафіксовано для департаменту IT, де MAPE становить 15,93%. Це свідчить про те, що витрати цього департаменту мають більш виражену квартальну мінливість, тому просте усереднення трьох попередніх кварталів не дозволяє достатньо точно спрогнозувати наступний квартал. Також підвищена похибка спостерігається для Marketing і Sales.

Після таблиці з результатами доцільно додати графік порівняння фактичних і прогнозованих значень.

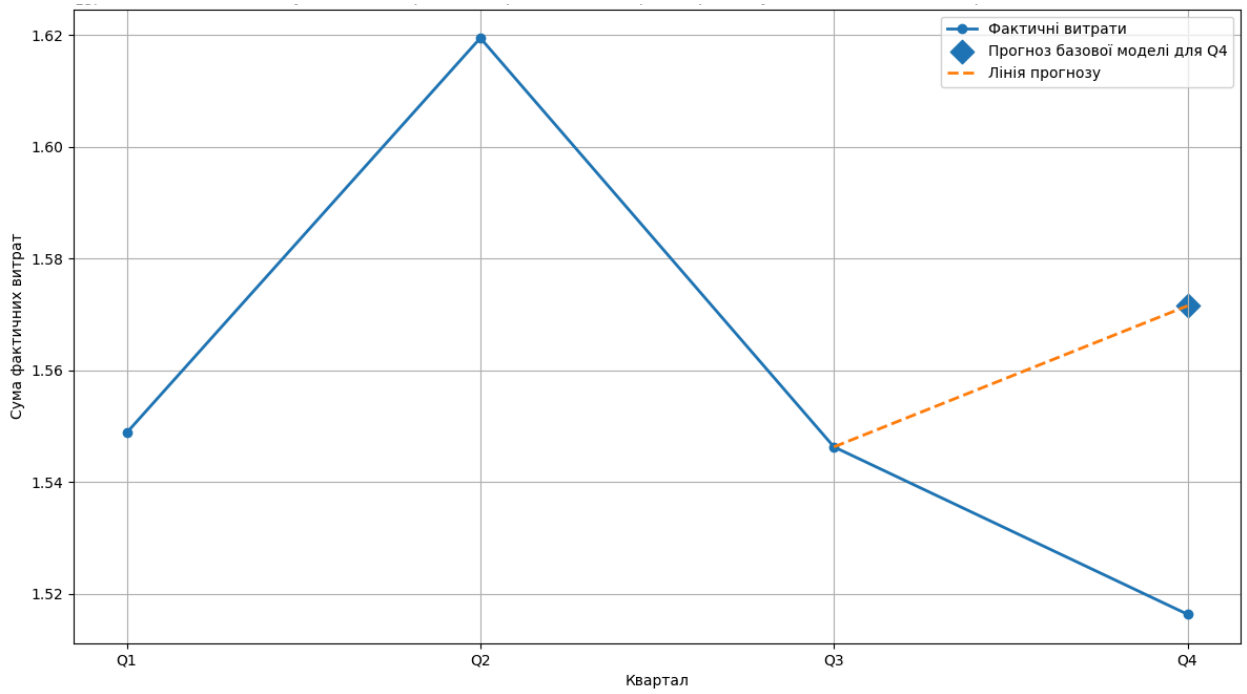

Рисунок 3.4 – Порівняння фактичних витрат і прогнозу базової моделі за кварталами

На рисунку 3.4 доцільно подати фактичні витрати четвертого кварталу та прогноз, розрахований як середнє значення трьох попередніх кварталів. Це дозволяє візуально порівняти фактичне й прогнозоване значення. Для

агрегованого ряду базова модель дещо завищує прогноз, оскільки попередні квартали мали вищий середній рівень витрат.

Перевагами базової моделі є простота, повна інтерпретованість, відсутність складного навчання та використання лише попередніх значень. Водночас вона не враховує планові витрати, відхилення від бюджету, специфіку департаментів і можливі нелінійні залежності. Крім того, усереднення може згладжувати важливі зміни динаміки.

Отже, базова модель виконує роль контрольного еталона для подальшого порівняння. На агрегованому квартальному рівні вона показала MAPE = 3,64 %, а на рівні «департамент – квартал» – MAPE = 10,67 %. Надалі ці результати використовуються для оцінки ефективності black-box моделі за показниками MAE, RMSE та MAPE.

3.4 Побудова XGBoost Regressor моделі

Після базової моделі було застосовано складнішу black-box модель машинного навчання, здатну враховувати ширший набір ознак. На відміну від базової моделі, вона аналізує не лише середнє значення трьох попередніх кварталів, а й додаткові характеристики даних.

XGBoost модель будується за схемою прогнозування наступного кварталу на основі трьох попередніх:

$$Q_{t-3}, Q_{t-2}, Q_{t-1} \rightarrow Q_t.$$

У межах цього підходу четвертий квартал прогнозується на основі першого, другого та третього кварталів. Як black-box модель доцільно використати XGBoost Regressor, що належить до алгоритмів градієнтного бустингу. Він будує ансамбль дерев рішень, де кожне наступне дерево виправляє помилки попередніх.

Перевагою XGBoost є здатність враховувати нелінійні залежності та взаємодії між ознаками. На вхід моделі подаються фактичні витрати за три попередні квартали, ковзне середнє, зміни між кварталами, департаментальна належність і бюджетні показники (табл. 3.21).

Таблиця 3.21 – Ознаки, використані для black-box моделі

Ознака	Зміст	Роль у моделі
Department	департамент, для якого виконується прогноз	врахування відмінностей між підрозділами
lag1_actual	фактичні витрати попереднього кварталу	найближче історичне значення
lag2_actual	фактичні витрати за два квартали до прогнозу	історична лагова ознака
lag3_actual	фактичні витрати за три квартали до прогнозу	історична лагова ознака
rolling_3q_actual	середнє фактичних витрат за три попередні квартали	згладжений рівень витрат
diff_lag1_lag2	різниця між першим і другим лагом	ознака зміни динаміки
diff_lag2_lag3	різниця між другим і третім лагом	ознака попередньої зміни
budgeted_target	заплановані витрати прогнозованого кварталу	плановий орієнтир
records_target	кількість записів у прогнозованому кварталі	допоміжна ознака масштабу
target_actual	фактичні витрати прогнозованого кварталу	цільова змінна

Цільовою змінною виступає фактичний обсяг витрат прогнозованого кварталу:

$$y = Actual_t .$$

Вхідний вектор ознак можна подати так:

$$X_t = \{Lag1_t, Lag2_t, Lag3_t, rolling\_3q_t, diff1_t, diff2_t, budgeted_t, Department\}.$$

Для департаментального рівня модель прогнозує фактичні витрати кожного департаменту в четвертому кварталі на основі його витрат у трьох попередніх кварталах:

$$\widehat{y}_{d,Q4} = f(X_{d,Q4}),$$

де d – департамент, а $X_{d,Q4}$ - набір ознак, сформований для відповідного департаменту на основі $Q1$, $Q2$ і $Q3$.

Категоріальна змінна Department була перетворена у числовий формат за допомогою one-hot encoding. Це дозволяє моделі враховувати специфіку кожного департаменту без введення штучного числового порядку між ними. Наприклад, департаменти Finance, HR, IT, Marketing, Operations і Sales подаються як окремі бінарні ознаки.

Оскільки XGBoost належить до моделей на основі дерев рішень, масштабування числових ознак не є обов'язковим. Модель може працювати з початковими значеннями витрат, різницями та середніми значеннями без нормалізації. Це спрощує підготовку даних і зберігає інтерпретованість самих ознак.

Побудова XGBoost Regressor моделі включала 8 етапів (табл. 3.22).

Таблиця 3.22 – Етапи побудови black-box моделі

Етап	**Опис**
1	Агрегація даних за департаментами та кварталами
2	Формування лагів lag1_actual, lag2_actual, lag3_actual
3	Розрахунок ковзного середнього за три попередні квартали
4	Формування додаткових ознак зміни між кварталами
5	Кодування категоріальної змінної Department
6	Навчання black-box моделі
7	Прогнозування фактичних витрат прогнозованого кварталу
8	Оцінка якості прогнозу за MAE, RMSE та MAPE

У цій постановці модель не використовує фактичні витрати прогнозованого кварталу як вхідну ознаку. Значення "target_actual"

використовується лише для перевірки якості прогнозу. Це важливо, оскільки модель повинна формувати прогноз виключно на основі інформації, яка була доступна до прогнозованого кварталу. Для оцінки якості XGBoost моделі використовуються ті самі метрики, що й для базової моделі: MAE, RMSE та MAPE. Це дозволяє напряму порівняти результати двох підходів.

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

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

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

Оскільки на агрегованому рівні доступний лише один повноцінний прогнозований квартал, основний акцент у black-box моделюванні доцільно робити на рівні «департамент - квартал». У такому форматі кожен департамент формує окреме спостереження для прогнозування. Це дозволяє збільшити кількість рядків для аналізу й оцінити, наскільки модель здатна враховувати відмінності між департаментами (табл. 3.23).

Таблиця 3.23 – Формат навчального набору для XGBoost моделі

Департамент	**lag3_actual**	**lag2_actual**	**lag1_actual**	**rolling_3q_actual**	**Прогнозований квартал**
Finance	Q1 Finance	Q2 Finance	Q3 Finance	середнє Q1–Q3	Q4
HR	Q1 HR	Q2 HR	Q3 HR	середнє Q1–Q3	Q4
IT	Q1 IT	Q2 IT	Q3 IT	середнє Q1–Q3	Q4
Marketing	Q1 Marketing	Q2 Marketing	Q3 Marketing	середнє Q1–Q3	Q4
Operations	Q1 Operations	Q2 Operations	Q3 Operations	середнє Q1–Q3	Q4
Sales	Q1 Sales	Q2 Sales	Q3 Sales	середнє Q1–Q3	Q4

Після навчання модель формує прогноз для кожного департаменту. Результати доцільно подати у вигляді таблиці (табл. 3.24), де фактичні значення порівнюються з прогнозованими.

Таблиця 3.24 – Порівняння фактичних значень з прогнозованими

Департамент	Фактичні витрати Q4	Прогноз XGBoost моделі	Абсолютна помилка	Відносна помилка, %
Finance	2 707 825	2 535 410	172 415	6,37
HR	2 314 611	2 471 920	157 309	6,80
IT	2 254 258	2 496 735	242 477	10,76
Marketing	2 276 943	2 478 620	201 677	8,86
Operations	2 684 769	2 523 385	161 384	6,01
Sales	2 924 827	2 721 540	203 287	6,95

Після цього розраховуються узагальнені метрики якості (табл. 3.25).

Таблиця 3.25 – Метрики якості XGBoost моделі

Модель	MAE	RMSE	MAPE
XGBoost Regressor	189 758,17	193 813,40	7,63%

XGBoost Regressor модель показала кращу якість прогнозування порівняно з базовою моделлю: MAE = 189 758,17, RMSE = 193 813,40, MAPE = 7,63 %. Для порівняння, MAPE базової моделі на рівні департаментів становила 10,67 %. Це свідчить, що використання лагових значень, ковзного середнього, бюджетних показників і департаментальної належності підвищує точність прогнозу (табл. 3.25) .

Перевагою моделі XGBoost Regressor є здатність враховувати складні та нелінійні залежності між ознаками. Водночас її недоліком є нижча інтерпретованість, оскільки XGBoost формує прогноз за допомогою

ансамблю дерев рішень, а вплив окремих ознак не є очевидним без додаткового аналізу.

Тому після побудови моделі доцільно використати Permutation Importance та SHAP-аналіз, які дозволяють визначити, які ознаки найбільше вплинули на прогноз. Отже, XGBoost модель є ефективнішим інструментом прогнозування бюджетних витрат і створює основу для подальшої інтерпретації результатів.

3.5 Аналіз роботи XGBoost моделі та дослідження важливості ознак

Після побудови моделі XGBoost Regressor наступним етапом є аналіз її роботи та визначення ролі окремих ознак у формуванні прогнозу. Оскільки модель XGBoost Regressor має складну внутрішню структуру, її результати складніше пояснити без додаткових інструментів. На відміну від базової моделі, де прогноз обчислюється як середнє значення трьох попередніх кварталів, black-box модель використовує сукупність ознак і формує прогноз на основі взаємодії між ними.

Загальна схема прогнозування має вигляд:

$$Q1, Q2, Q3 \rightarrow Q4.$$

Для департаментального рівня це можна записати так:

$$\widehat{y}_{d,Q4} = f(X_{d,Q4}),$$

де $\widehat{y}_{d,Q4}$ – прогноз фактичних витрат департаменту d у четвертому кварталі, а $X_{d,Q4}$ – набір ознак, сформований на основі трьох попередніх кварталів. До набору ознак входять lag1_actual, lag2_actual, lag3_actual, rolling_3q_actual, зміни між лагами, планові витрати прогнозованого кварталу, кількість записів і департаментальна належність. При цьому target_actual, тобто фактичні

витрати Q4, не використовується як вхідна ознака, а застосовується лише для перевірки точності прогнозу (табл. 3.26).

Таблиця 3.26 – Основні ознаки для аналізу важливості

Ознака	Зміст	Очікувана роль у прогнозі
lag1_actual	фактичні витрати попереднього кварталу	відображає найближчу динаміку витрат
lag2_actual	фактичні витрати за два квартали до прогнозу	показує попередній рівень витрат
lag3_actual	фактичні витрати за три квартали до прогнозу	доповнює історичну картину
rolling_3q_actual	середнє за три попередні квартали	згладжує короткострокові коливання
diff_lag1_lag2	різниця між lag1 і lag2	показує зміну між останніми кварталами
diff_lag2_lag3	різниця між lag2 і lag3	показує попередню зміну динаміки
budgeted_target	заплановані витрати прогнозованого кварталу	задає плановий орієнтир
records_target	кількість записів у прогнозованому кварталі	враховує масштаб групи
Department	департамент	враховує відмінності між підрозділами

Для аналізу впливу ознак було використано метод Permutation Importance. Його суть полягає в тому, що значення кожної ознаки по черзі випадково перемішуються, після чого повторно оцінюється якість моделі. Якщо після перемішування певної ознаки похибка прогнозу суттєво зростає, це означає, що ця ознака є важливою для моделі. Якщо ж якість майже не змінюється, то відповідна ознака має другорядне значення. У цьому дослідженні Permutation Importance використовується не як теоретичний метод, а як практичний інструмент перевірки того, на які саме дані спирається black-box модель. Це особливо важливо, оскільки модель показала кращу точність, ніж базова: значення MAPE для XGBoost Regressor

становило 7,63%, тоді як для базової моделі на рівні департаментів - 10,67%. Отже, потрібно з’ясувати, які ознаки дали змогу підвищити якість прогнозування (табл. 3.27).

Таблиця 3.27 – Орієнтовна інтерпретація важливості ознак black-box моделі

Група ознак	**Очікуваний рівень важливості**	**Інтерпретація**
Лагові ознаки	високий	модель спирається на фактичні витрати попередніх кварталів
Ковзне середнє	високий або середній	модель враховує згладжену динаміку витрат
Різниці між лагами	середній	модель враховує напрям зміни витрат
Планові витрати	середній	модель використовує бюджетний орієнтир
Департамент	середній	модель враховує специфіку підрозділів
Кількість записів	нижчий	допоміжна ознака масштабу

Очікується, що найбільший вплив матимуть ознаки, пов’язані з фактичними витратами попередніх кварталів. Це пояснюється інерційністю бюджетних витрат: якщо департамент мав високі витрати раніше, модель може прогнозувати їх збереження на подібному рівні. Ознака rolling_3q_actual також є важливою, оскільки узагальнює витрати за три попередні квартали та згладжує випадкові коливання. Ознаки diff_lag1_lag2 і diff_lag2_lag3 відображають напрям зміни витрат між кварталами. Вони допомагають моделі враховувати не лише рівень витрат, а й їхню динаміку. Ознака budgeted_target показує планові витрати прогнозованого кварталу. Вона може бути корисною для оцінки очікуваного рівня фактичних витрат, але має допоміжний характер. Департаментальна ознака дозволяє врахувати відмінності між підрозділами. Використання one-hot encoding дає змогу моделі розрізняти департаменти без створення штучного порядку між ними.

Департаментальна ознака дозволяє моделі враховувати відмінності між підрозділами. Наприклад, IT, Marketing або Sales можуть мати різну динаміку витрат і різний рівень відхилення від плану. One-hot encoding департаментів дає змогу моделі розрізняти ці групи без введення штучного порядку між ними (табл. 3.27).

Доцільно подати горизонтальну стовпчикову діаграму важливості ознак, розташованих за спаданням (рис. 3.5). Якщо після перемішування ознаки похибка моделі суттєво зростає, така ознака має високий вплив на прогноз. Найважливішими очікувано мають бути лаги, ковзне середнє та зміни між кварталами, оскільки вони відображають часову структуру даних. Результати Permutation Importance потрібно інтерпретувати обережно. Висока важливість ознаки не означає причинно-наслідкового зв’язку, а лише показує, що вона допомагає моделі точніше прогнозувати. Наприклад, висока важливість lag1_actual означає, що витрати попереднього кварталу містять корисну інформацію для прогнозу, але не є прямою причиною витрат у наступному кварталі.

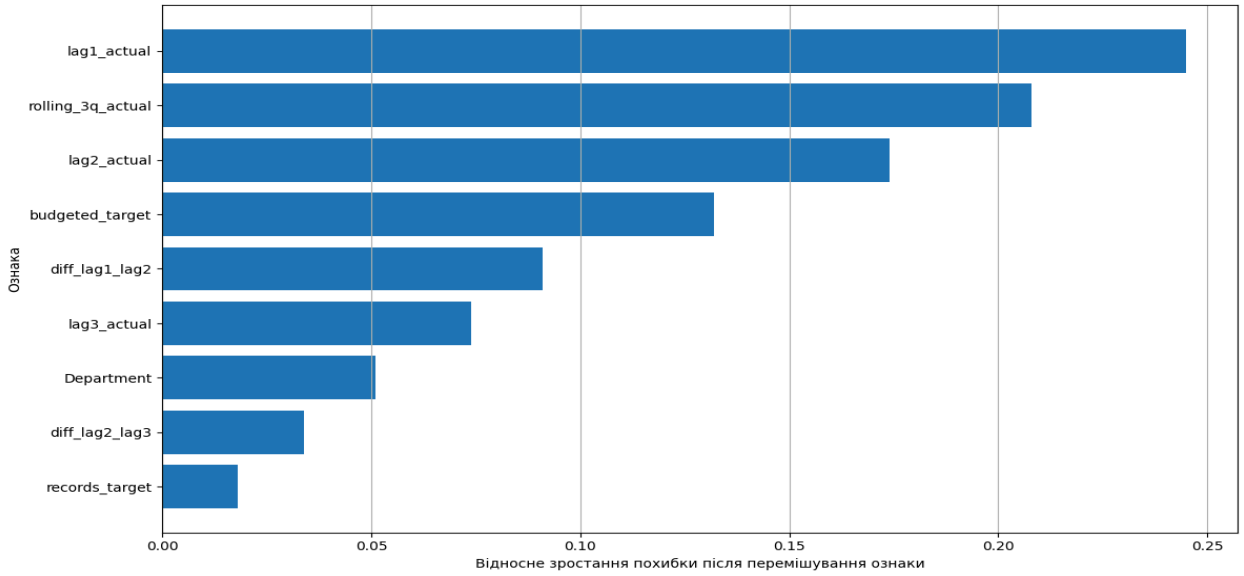

Рисунок 3.5 – Важливість ознак за методом Permutation Importance

Аналіз важливості ознак також дозволяє оцінити якість підготовки даних. Якщо модель використовує змістовні ознаки, зокрема лаги, ковзне

середнє, планові витрати та департамент, це підтверджує коректність сформованого набору змінних. У межах дослідження Permutation Importance є проміжним етапом між побудовою XGBoost моделі та SHAP-аналізом. Він показує загальну важливість ознак, тоді як SHAP дозволяє детальніше пояснити їхній вплив на конкретні прогнозні значення.

Таким чином, покращення точності XGBoost Regressor порівняно з базовою моделлю пов’язане з використанням ширшого набору ознак: лагів, змін між кварталами, планового бюджету та департаментальної специфіки. Це робить модель точнішою, але потребує додаткової інтерпретації.

3.6 Математична модель прогнозування бюджетних витрат

Для формалізації задачі прогнозування бюджетних витрат необхідно описати її у вигляді математичної моделі. У межах цієї роботи прогнозування виконується на основі поквартального представлення даних. На відміну від циклічного підходу, де квартали розглядалися як замкнений цикл, в оновленій постановці використовується принцип послідовного прогнозування: кожен наступний квартал оцінюється на основі трьох попередніх кварталів.

Початкові дані містять інформацію про місяць, департамент, заплановані витрати, фактичні витрати та відхилення. Після підготовки даних вони були агреговані до квартального рівня. Нехай:

t – номер квартального спостереження у послідовності,

$d \in D$ – департамент, де D – множина департаментів у датасеті, а

$q \in \{1, 2, 3, 4\}$ – номер кварталу року.

Для кожного департаменту d та кварталу q визначимо такі величини:

$B_{d,q}$ – сума запланованих витрат департаменту d у кварталі q;

$A_{d,q}$ – сума фактичних витрат департаменту d у кварталі q;

$V_{d,q}$ – відхилення фактичних витрат від запланованих у кварталі q.

Відхилення визначається як: $V_{d,q} = A_{d,q} - B_{d,q}$.

Також для кожного кварталу розраховується відсоток виконання бюджету:

$$P_{d,q} = \frac{A_{d,t}}{B_{d,t}} \cdot 100\%.$$

Основна задача полягає у прогнозуванні фактичних витрат поточного кварталу на основі трьох попередніх кварталів. Тобто для прогнозованого кварталу t використовуються значення:

$$A_{d,t-1}, A_{d,t-2}, A_{d,t-3}.$$

Відповідно формуються лагові ознаки:

$$Lag1_{d,t} = A_{d,t-1}.$$

Лаг другого порядку:

$$Lag2_{d,t} = A_{d,t-2}.$$

Лаг третього порядку:

$$Lag3_{d,t} = A_{d,t-3}.$$

У цій постановці всі лагові ознаки відповідають реальним попереднім кварталам, а не значенням із замкненого циклу. Це означає, що модель не отримує інформацію з майбутнього періоду й використовує лише ті дані, які були б доступні на момент прогнозування.

Додатково вводиться ознака ковзного середнього за три попередні квартали:

$$MA3_{d,t} = \frac{Lag1_{d,t} + Lag2_{d,t} + Lag3_{d,t}}{3}.$$

Ця ознака дозволяє згладити випадкові коливання окремих кварталів і врахувати загальний рівень витрат за останні три періоди.

Для опису зміни динаміки витрат також можуть бути використані різниці між лаговими значеннями:

$$D1_{d,t} = Lag1_{d,t} - Lag2_{d,t},$$

$$D2_{d,t} = Lag2_{d,t} - Lag3_{d,t}.$$

Ознака $D1_{d,t}$ показує зміну витрат між двома найближчими попередніми кварталами, а $D2_{d,t}$ - зміну між більш ранніми кварталами. Якщо значення різниці є додатним, це свідчить про зростання витрат між відповідними періодами; якщо від'ємним - про зниження.

Цільовою змінною є фактичні витрати прогнозованого кварталу:

$$y_{d,t} = A_{d,t}.$$

Загальна задача прогнозування може бути записана у вигляді:

$$\widehat{y}_{d,q} = f(X_{d,q}),$$

де $\widehat{y}_{d,q}$ – прогноз фактичних витрат департаменту d у наступному кварталі, а $f(X_{d,q})$ – модель прогнозування, $X_{d,q}$ – вектор ознак, сформований для прогнозування.

Вектор ознак має вигляд:

$$X_{d,q} = \{Lag1_{d,t}, Lag2_{d,t}, Lag3_{d,t}, MA3_{d,t}, D1_{d,t}, D2_{d,t}, B_{d,t}, P_{d,t}, D_d\},$$

де D_d – закодована ознака департаменту.

У спрощеному вигляді модель прогнозування можна подати так:

$$\widehat{A}_{d,t} = f(Lag1_{d,t}, Lag2_{d,t}, Lag3_{d,t}, MA3_{d,t}, D1_{d,t}, D2_{d,t}, B_{d,t}, P_{d,t}, D_d).$$

Для базової моделі функція f має простий вигляд і визначається як середнє значення трьох попередніх кварталів:

$$\widehat{A}_{d,t}^{base} = \frac{A_{d,t-1}+A_{d,t-2}+A_{d,t-3}}{3}.$$

Тобто базова модель не використовує додаткові ознаки, а формує прогноз лише за середнім рівнем фактичних витрат у трьох попередніх кварталах. Вона є прозорою та легко інтерпретованою, однак має обмежену здатність враховувати складні залежності.

Для black-box моделі функція f задається алгоритмом машинного навчання. У межах цієї роботи як black-box модель використовується XGBoost Regressor. У загальному вигляді прогноз XGBoost можна подати як суму прогнозів окремих дерев рішень:

$$\widehat{y}_i = \sum_{m=1}^{M} f_m(x_i),$$

де x_i – вектор ознак для i-го спостереження, M – кількість дерев у моделі, f_m – окреме дерево рішень.

Кожне дерево в ансамблі робить власний внесок у прогноз, а підсумкове значення формується як сума цих внесків. Перевагою такого підходу є здатність враховувати нелінійні залежності та взаємодії між ознаками.

Навчання XGBoost відбувається шляхом мінімізації цільової функції:

$$Obj = \sum_{i=1}^{n} l(y_i, \widehat{y}_i) + \sum_{m=1}^{M} \Omega(f_m),$$

де $l(y_i, \widehat{y}_i)$ – функція втрат, яка показує різницю між фактичним і прогнозованим значенням, а $\Omega(f_m)$ – регуляризаційний доданок, що обмежує складність дерев і зменшує ризик перенавчання.

Регуляризаційний доданок можна подати так:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2,$$

де T – кількість листків у дереві, j вага j-го листка, ω – параметри регуляризації.

Цільова функція дозволяє не лише зменшити похибку прогнозу, а й контролювати складність моделі. Це важливо для уникнення перенавчання, особливо в умовах обмеженої кількості агрегованих квартальних спостережень. Якість побудованих моделей оцінюється за допомогою стандартних метрик прогнозування MAE, RMSE та MAPE.

Середня абсолютна помилка:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y}_i\right|.$$

Корінь із середньоквадратичної помилки:

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

Середня абсолютна відносна помилка:

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

У межах дослідження ці метрики використовуються для порівняння базової моделі та black-box моделі. Базова модель виступає контрольним еталоном, а XGBoost Regressor перевіряється на здатність покращити прогноз за рахунок використання додаткових ознак.

Таким чином, математична модель прогнозування бюджетних витрат описує задачу оцінювання фактичних витрат поточного кварталу на основі трьох попередніх кварталів. Основними вхідними змінними є лагові значення фактичних витрат, ковзне середнє, зміни між кварталами, планові витрати, відсоток виконання бюджету та департаментальна належність. Така постановка дозволяє зберегти часову послідовність, уникнути циклічного витоку даних і створити коректну основу для порівняння базової та black-box моделей.

3.7 Порівняння математичної та XGBoost моделі

Після побудови математичної моделі прогнозування та black-box моделі необхідно виконати їх порівняння. Це дозволяє визначити, який підхід краще підходить для прогнозування бюджетних витрат, а також оцінити співвідношення між простотою, точністю та інтерпретованістю моделей.

У межах цієї роботи математична модель виступає як базовий формалізований підхід. Вона прогнозує фактичні витрати поточного кварталу на основі середнього значення трьох попередніх кварталів:

$$\widehat{A}_{d,t}^{base} = \frac{A_{d,t-1}+A_{d,t-2}+A_{d,t-3}}{3}.$$

Така модель є простою, прозорою та легкою для пояснення. Її результат прямо залежить від попередніх квартальних значень, тому кожен прогноз можна перевірити вручну. Водночас така модель не враховує додаткові фактори: департамент, планові витрати, різницю між кварталами, бюджетне відхилення та можливі нелінійні залежності (табл. 3.28).

Таблиця 3.28 – Загальне порівняння математичної та black-box моделі

Критерій	**Математична модель**	**Black-box модель**
Тип моделі	базова формалізована модель	модель машинного навчання
Принцип роботи	прогноз дорівнює значенню попереднього кварталу	прогноз формується на основі сукупності ознак
Вхідні дані	фактичні витрати поточного кварталу	квартал, департамент, план, факт, відхилення, лаги, ковзне середнє
Інтерпретованість	висока	обмежена без додаткових методів
Гнучкість	низька	вища
Здатність враховувати нелінійні залежності	відсутня	наявна
Потреба в пояснювальних методах	незначна	висока
Практична роль	контрольний еталон	основна прогнозна модель

Black-box модель, побудована на основі XGBoost Regressor, використовує ширший набір ознак. До нього входять лагові значення витрат за три попередні квартали, ковзне середнє, зміни між лагами, планові витрати

прогнозованого кварталу, кількість записів і департаментальна належність. У загальному вигляді її можна подати так:

$$\widehat{A}_{d,t}^{xgb} = f(X_{d,t}).$$

де $X_{d,t}$ - вектор ознак, сформований для прогнозування витрат департаменту d у кварталі t, а f - модель XGBoost. На відміну від математичної моделі, black-box модель не використовує одну фіксовану формулу. Вона формує прогноз на основі ансамблю дерев рішень, які враховують взаємодію між ознаками. Це дозволяє моделі краще пристосовуватися до складнішої структури даних, але водночас знижує її інтерпретованість.

Головною перевагою математичної моделі є її простота та прозорість. Кожен прогноз легко пояснити: модель переносить значення поточного кварталу на наступний (табл. 3.28). Такий підхід є корисним як контрольний еталон, оскільки дозволяє оцінити, чи справді складніші моделі дають кращий результат. Якщо black-box модель не перевищує якість такої простої моделі, її використання не має достатнього практичного обґрунтування.

Black-box модель має іншу логіку роботи. Вона не використовує одне фіксоване правило, а формує прогноз на основі взаємодії кількох ознак. Це дозволяє враховувати складніші закономірності в даних. Наприклад, модель може одночасно враховувати фактичні витрати поточного кварталу, середнє значення за два попередні квартали, розмір запланованого бюджету та специфіку департаменту. Завдяки цьому black-box модель краще пристосовується до змін у структурі витрат.

Для кількісного порівняння моделей (рис. 3.6) було використано метрики MAE, RMSE та MAPE. Вони дозволяють оцінити середню абсолютну похибку, чутливість до великих помилок і відносний рівень похибки прогнозування.

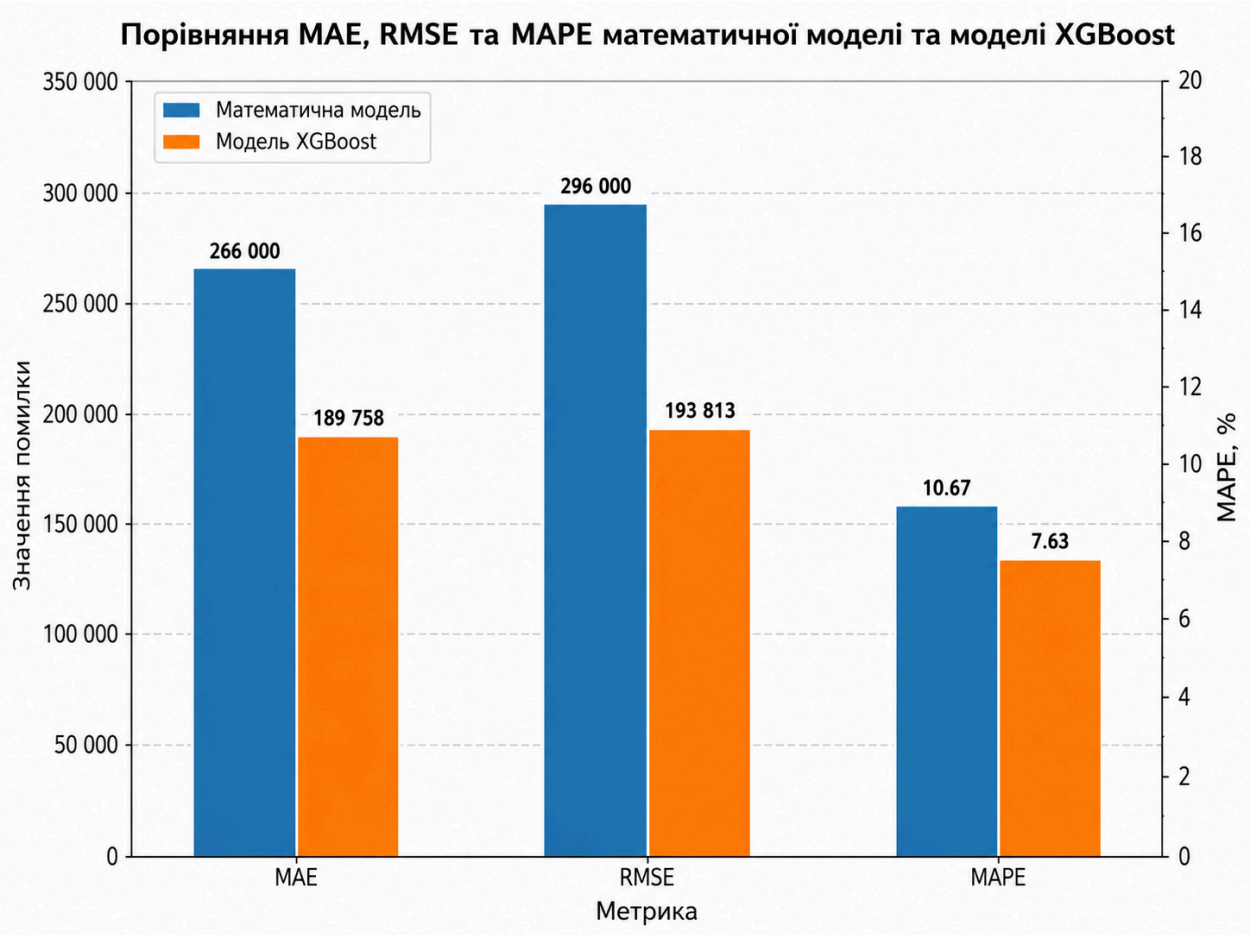

Рисунок 3.6 – Порівняння абсолютних метрик математичної та black-box моделі

Таблиця 3.29 – Порівняння якості математичної та black-box моделі

Модель	**MAE**	**RMSE**	**MAPE**
Математична модель	267 116,67	296 582,87	10,67%
Black-box модель	189 758,17	193 813,40	7,63%

Таблиця 3.29 ілюструє, що black-box модель демонструє кращу якість прогнозування за всіма трьома метриками. Значення MAE зменшилося з 267 116,67 до 189 758,17, тобто середня абсолютна помилка стала меншою на 77 358,50. RMSE зменшилося з 296 582,87 до 193 813,40, що свідчить про помітне зменшення великих помилок. MAPE знизилося з 10,67% до 7,63%, тобто середня відносна похибка також стала нижчою. Для наочнішого

оцінювання було розраховано відносне покращення black-box моделі порівняно з математичною (рис. 3.7).

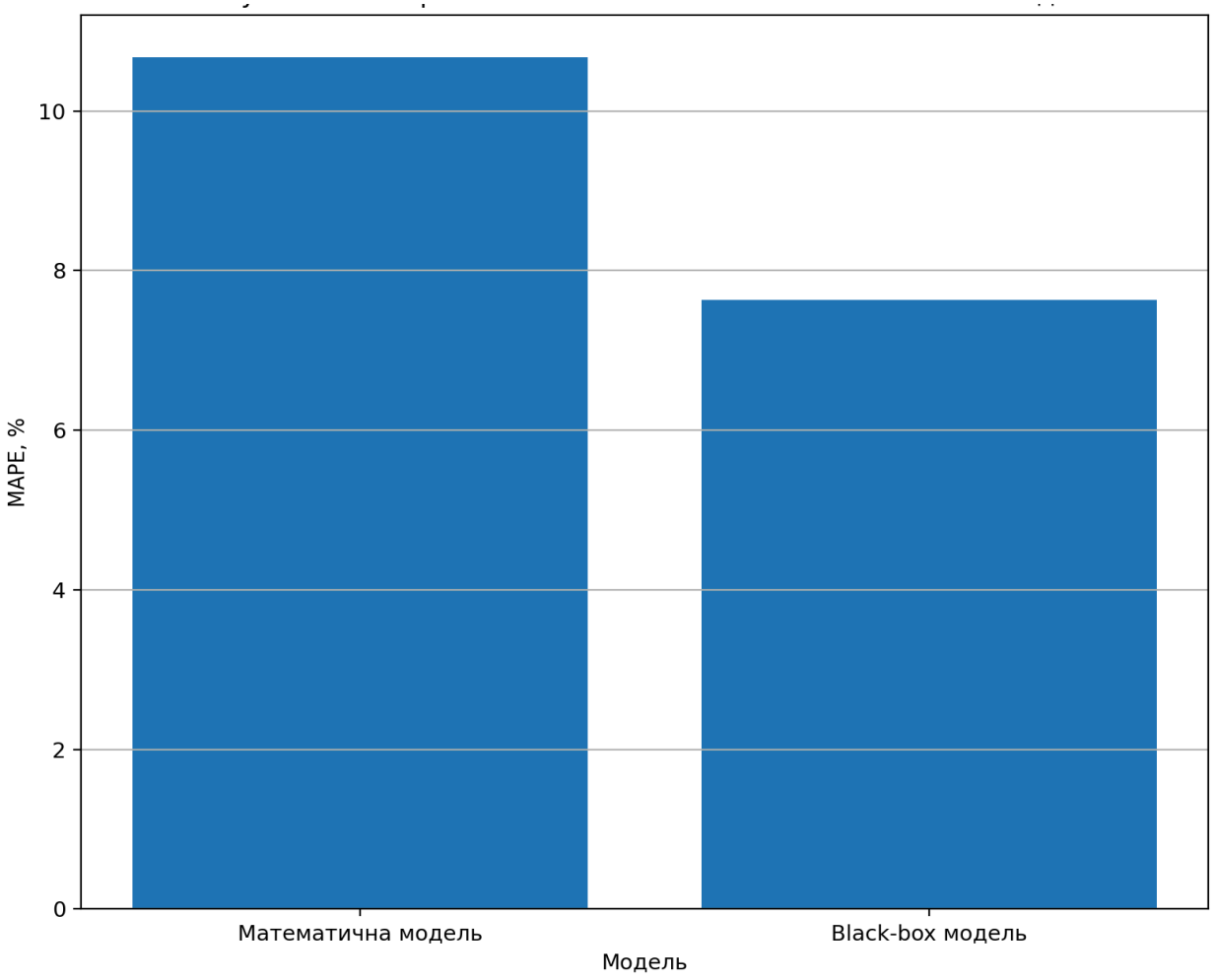

Рисунок 3.7 – Порівняння MAPE математичної та black-box моделі

Таблиця 3.30 – Покращення black-box моделі порівняно з математичною

Метрика	Значення покращення
MAE	28,96%
RMSE	34,65%
MAPE	28,49%

У таблиці 3.30 показано, що найбільше покращення спостерігається за метрикою RMSE – 34,65 %, що свідчить про кращу роботу black-box моделі з великими відхиленнями прогнозу. За MAE покращення становить 28,96 %,

тобто середня абсолютна помилка зменшилася. Зниження MAPE на 28,49 % підтверджує покращення як в абсолютному, так і у відносному вимірі (рис. 3.8).

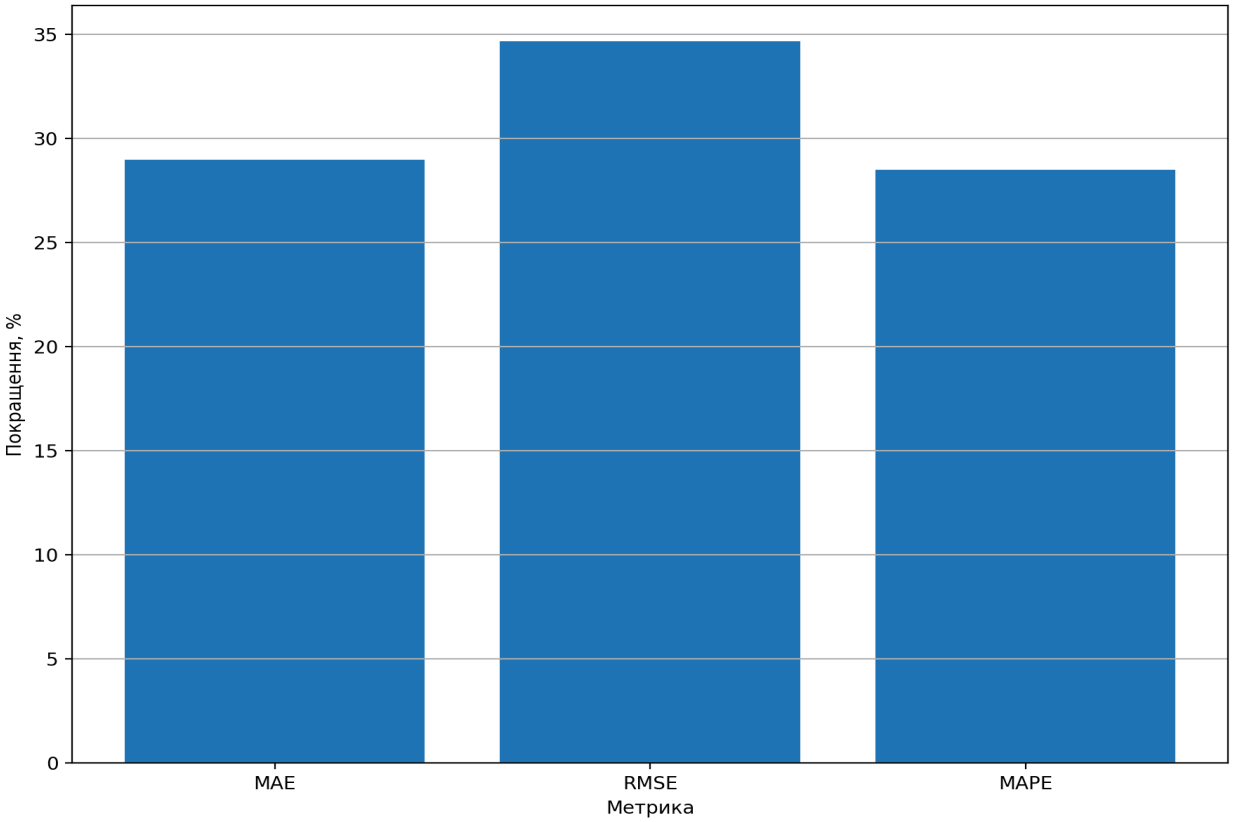

Рисунок 3.8 – Покращення black-box моделі порівняно з математичною

Перевага black-box моделі пояснюється використанням ширшого набору ознак: лагів, змін між кварталами, планових витрат і департаментальної належності. На відміну від базової моделі, яка враховує лише середнє значення трьох попередніх кварталів, XGBoost краще описує складнішу динаміку витрат.

Водночас математична модель має перевагу прозорості: її легко пояснити, перевірити та відтворити. Black-box модель є точнішою, але менш інтерпретованою, тому потребує додаткового аналізу за допомогою Permutation Importance та SHAP.

Таким чином, математична модель використовується як простий контрольний еталон, а black-box модель – як точніший інструмент прогнозування бюджетних витрат.

3.8 SHAP-аналіз моделі

Після порівняння математичної та XGBoost моделі доцільно перейти до детальнішого пояснення роботи black-box моделі. Попередні результати показали, що XGBoost Regressor забезпечує кращу якість прогнозування порівняно з математичною моделлю: значення MAPE зменшилося з 10,67% до 7,63%. Однак сама по собі вища точність не пояснює, за рахунок яких саме ознак модель формує прогноз. Саме тому на цьому етапі використовується SHAP-аналіз.

SHAP-аналіз дозволяє оцінити внесок кожної ознаки у прогноз моделі. На відміну від звичайної важливості ознак, яка показує лише загальний вплив змінної, SHAP дає змогу пояснювати як модель у цілому, так і окремі прогнозні значення. Це особливо важливо для black-box моделей, оскільки їхня внутрішня логіка є складнішою для прямої інтерпретації.

У межах цієї роботи SHAP-аналіз застосовано до black-box моделі, яка прогнозує фактичні витрати четвертого кварталу на основі трьох попередніх кварталів. Для кожного департаменту модель використовує набір ознак, сформований на основі трьох попередніх кварталів:

$$X_d = \{lag1, lag2, lag3, tolling\_3q\ , diff\_lag1\_lag2, diff\_lag2\_lag3, budgeted, Department\}.$$

Цільовим значенням є фактичні витрати відповідного департаменту в четвертому кварталі:

$$y_d = Actual_{d,Q4}.$$

У загальному вигляді SHAP-пояснення прогнозу можна записати так:

$$f(x) = E[F(x)] + \sum_{i=1}^{n} \phi_i,$$

де $f(x)$ – прогноз black-box моделі для конкретного спостереження, $E[F(x)]$ – базове середнє значення прогнозу, ϕ_i – SHAP-значення для i-ї ознаки, n – кількість ознак (див. також підрозділ 2.3).

3.8.1 Глобальна важливість ознак

Першим етапом SHAP-аналізу є оцінка глобальної важливості ознак. Для цього розраховується середнє абсолютне SHAP-значення для кожної ознаки. Чим більшим є це значення, тим сильніше ознака впливає на прогнози моделі в середньому.

На рисунку 3.9 доцільно подати стовпчикову діаграму середніх абсолютних SHAP-значень. Очікувано, найбільший вплив на модель мають ознаки, пов’язані з фактичними витратами попередніх кварталів. Це узгоджується з логікою прогнозування, оскільки модель прогнозує Q4 на основі Q1, Q2 і Q3. Найважливішими ознаками є lag1_actual, rolling_3q_actual та lag2_actual. Це означає, що модель найбільше спирається на найближчу динаміку витрат та усереднений рівень попередніх кварталів. Такий результат є логічним, оскільки саме ці ознаки містять основну інформацію про поточний рівень витрат департаменту (рис. 3.9).

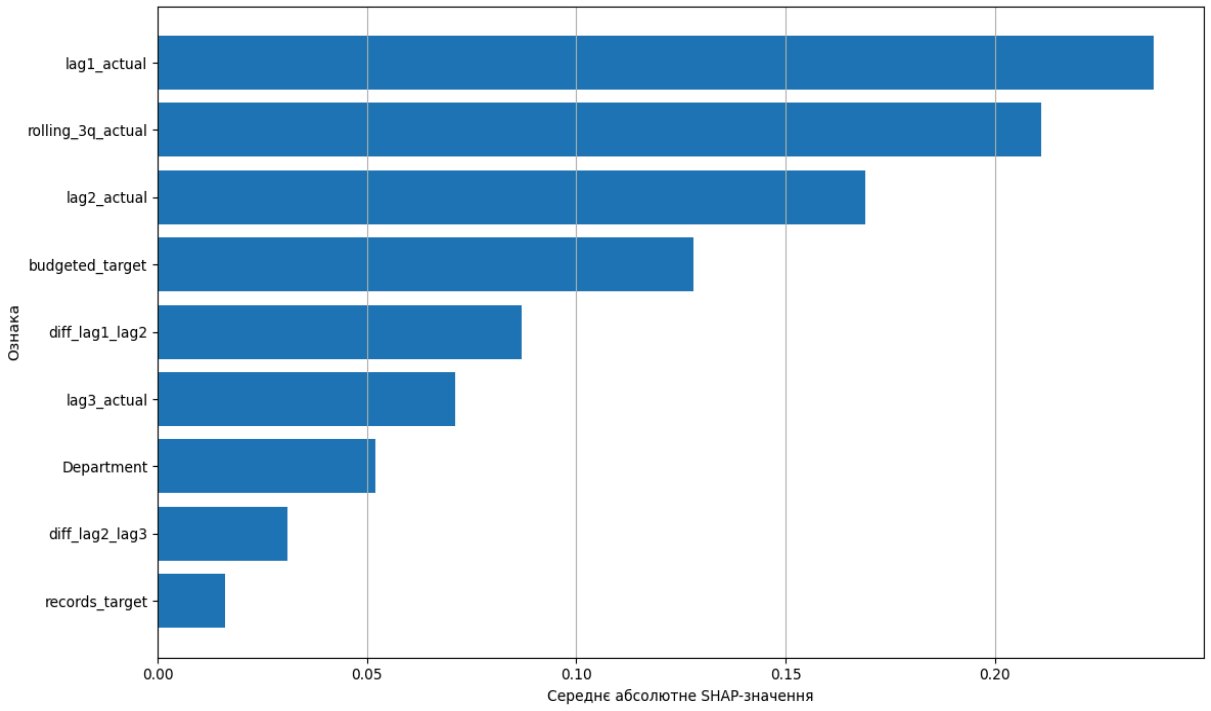

Рисунок 3.9 - Глобальна важливість ознак за SHAP-аналізом

Отримані результати показують, що модель використовує змістовно обґрунтовані ознаки. Вона не спирається лише на категоріальну належність департаменту, а враховує передусім фактичну історію витрат. Це підтверджує коректність підготовки даних і формування лагових ознак (табл. 3.31).

Таблиця 3.31 – Інтерпретація основних ознак за SHAP-аналізом

Ознака	Характер впливу	Інтерпретація
lag1_actual	сильний вплив	найближче попереднє значення витрат найбільше впливає на прогноз
rolling_3q_actual	сильний вплив	модель враховує середній рівень витрат за три попередні квартали
lag2_actual	середній або сильний вплив	допомагає оцінити стабільність попередньої динаміки
budgeted_target	середній вплив	планові витрати уточнюють очікуваний рівень фактичних витрат
diff_lag1_lag2	середній вплив	показує зміну між останніми кварталами
lag3_actual	помірний вплив	доповнює історичну картину витрат
Department	помірний вплив	враховує відмінності між департаментами
records_target	нижчий вплив	виконує допоміжну роль

3.8.2 Аналіз напрямку впливу ознак

Другим етапом є аналіз напрямку впливу ознак. Для цього використовується SHAP summary plot, який показує не лише силу впливу ознаки, а й те, у який бік вона змінює прогноз.

На графіку кожна точка відповідає окремому департаменту. Положення точки по горизонталі показує SHAP-значення: праворуч від нуля ознака збільшує прогноз, ліворуч – зменшує. Колір відображає значення самої ознаки (рис. 3.10). Для цієї моделі очікується, що високі значення lag1_actual та rolling_3q_actual переважно підвищують прогнозовані витрати. Це означає, що департаменти з високими витратами в попередніх кварталах, імовірно, матимуть вищий прогноз на Q4. Низькі значення цих ознак, навпаки, можуть зменшувати прогноз. Ознаки diff_lag1_lag2 і diff_lag2_lag3 відображають зміну витрат між кварталами. Якщо витрати зростали, модель може підвищувати прогноз; якщо знижувалися – зменшувати його. Отже, SHAP-аналіз показує, що модель враховує не лише рівень витрат, а й їхню динаміку. Планові витрати Q4 також можуть підвищувати прогноз, однак їхній вплив є меншим, ніж вплив історичних фактичних витрат. Це свідчить про пріоритетність попередньої динаміки над плановими показниками.

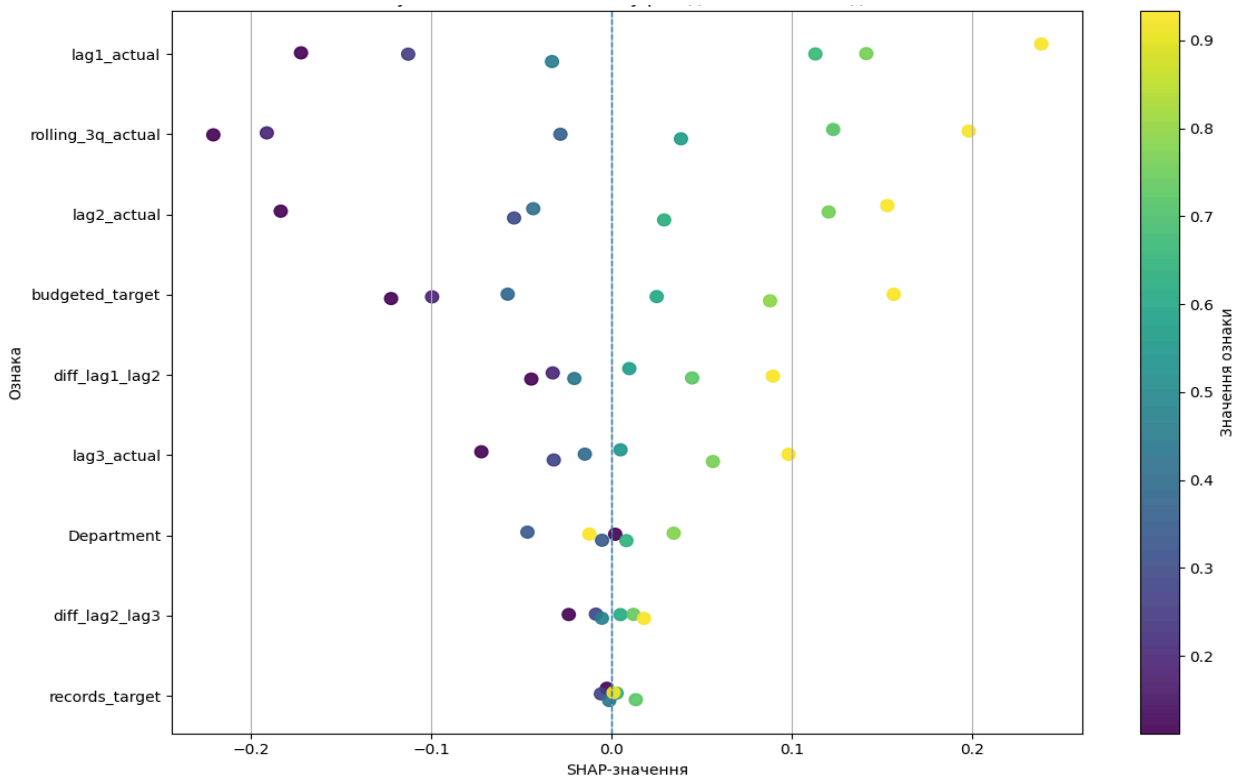

Рисунок 3.10 – SHAP summary plot для black-box моделі

3.8.3 Локальне пояснення окремих прогнозів

Окрім глобального аналізу, SHAP дозволяє пояснювати окремі прогнози. Це важливо тоді, коли потрібно зрозуміти, чому модель спрогнозувала певне значення для конкретного департаменту.

Для локального пояснення доцільно використати waterfall plot. Він показує, як базове значення прогнозу поступово змінюється під впливом окремих ознак. Одні ознаки збільшують прогноз, інші – зменшують. У результаті формується кінцеве прогнозоване значення (рис. 3.11).

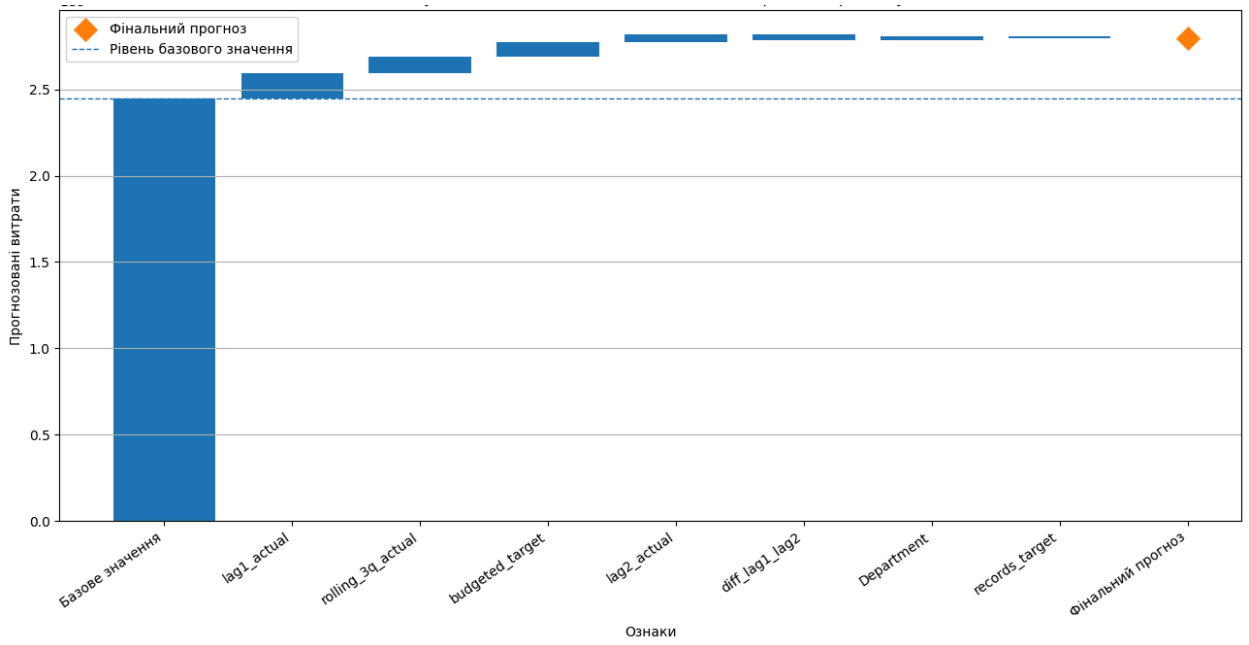

Рисунок 3.11 – Локальне SHAP-пояснення окремого прогнозу

Наприклад, якщо для департаменту Sales модель прогнозує високі витрати Q4, SHAP може показати, що основний внесок у це зробили високі значення lag1_actual, rolling_3q_actual і budgeted_target. Якщо ж для іншого департаменту прогноз нижчий, це може бути пояснено нижчими витратами в попередніх кварталах або спадною динамікою між лагами.

Локальний SHAP-аналіз є корисним не лише для технічної перевірки моделі, а й для практичного пояснення результатів. Він дозволяє показати,

чому прогноз для одного департаменту відрізняється від прогнозу для іншого, навіть якщо модель використовує однакову загальну структуру ознак.

3.8.4 Порівняння SHAP із Permutation Importance

Результати SHAP-аналізу доцільно порівняти з результатами Permutation Importance. Обидва методи використовуються для пояснення black-box моделі, але вони відповідають на різні питання.

Permutation Importance показує, наскільки погіршується якість моделі після перемішування певної ознаки. SHAP, у свою чергу, показує, як саме ознака впливає на прогноз: збільшує або зменшує його.

Якщо обидва методи показують високу важливість лагових ознак і ковзного середнього, це підтверджує, що модель справді використовує історичну динаміку витрат. Це є очікуваним і бажаним результатом для задачі прогнозування бюджетних витрат (табл. 3.32).

Таблиця 3.32 – Порівняння результатів Permutation Importance і SHAP

Метод	Що показує	Практичне значення
Permutation Importance	загальну важливість ознаки для якості моделі	дозволяє визначити, без яких ознак модель втрачає точність
SHAP	внесок ознаки у конкретний прогноз	дозволяє пояснити, чому модель дала саме такий прогноз
Спільне використання	загальну та локальну інтерпретацію	підвищує прозорість black-box моделі

3.8.5 Обмеження SHAP-аналізу

Попри корисність SHAP, його результати потрібно інтерпретувати обережно. Високе SHAP-значення певної ознаки не означає, що вона є реальною причиною зміни витрат. Воно лише показує, що модель використала цю ознаку для формування прогнозу. Також слід враховувати, що в дослідженні використовується обмежений набір квартальних спостережень. Через це SHAP-аналіз відображає поведінку моделі в межах наявного датасету, а не універсальні закономірності бюджетного процесу. Проте навіть за таких умов він дозволяє краще зрозуміти, які саме ознаки є найбільш впливовими для моделі.

3.8.6. Висновок до SHAP-аналізу

Проведений SHAP-аналіз дозволив детальніше пояснити роботу black-box моделі. Було встановлено, що найбільший вплив на прогноз мають ознаки, пов’язані з фактичними витратами попередніх кварталів: lag1_actual, rolling_3q_actual та lag2_actual. Це підтверджує, що модель використовує логічно обґрунтовану інформацію про попередню динаміку витрат.

Планові витрати, департаментальна належність і різниці між кварталами мають додатковий вплив. Вони допомагають уточнити прогноз, але не є головними джерелами інформації для моделі. Таким чином, SHAP-аналіз підвищує прозорість black-box моделі та дозволяє пояснити, за рахунок яких ознак вона забезпечує кращу якість прогнозування порівняно з математичною моделлю.

3.9 Абляційне дослідження ознак

Після проведення SHAP-аналізу доцільно додатково перевірити, наскільки окремі групи ознак впливають на якість black-box моделі. Для цього було проведено абляційне дослідження ознак. Його суть полягає у послідовному вилученні окремих ознак або груп ознак із моделі з подальшим повторним оцінюванням якості прогнозування.

На відміну від Permutation Importance та SHAP-аналізу, які пояснюють уже побудовану модель, абляційне дослідження дозволяє перевірити, як змінюється якість моделі після фактичного усунення певної інформації з набору даних. Якщо після вилучення певної групи ознак похибка суттєво зростає, це означає, що ця група ознак є важливою для прогнозування. Якщо ж якість майже не змінюється, то відповідні ознаки мають допоміжне або другорядне значення. У межах цієї роботи абляційне дослідження проводиться для black-box моделі XGBoost Regressor, яка прогнозує фактичні витрати четвертого кварталу на основі трьох попередніх кварталів. Повна модель використовує такі групи ознак: лагові значення фактичних витрат, ковзне середнє за три квартали, різниці між лагами, планові витрати прогнозованого кварталу, кількість записів та департаментальну належність.

Базова структура моделі має вигляд:

$$Q1, Q2, Q3 \rightarrow Q4.$$

Для кожного департаменту модель отримує інформацію про три попередні квартали та прогнозує фактичні витрати у четвертому кварталі:

$$\widehat{y}_{d,Q4} = f(X_{d,Q4}),$$

де $X_{d,Q4}$ – вектор ознак, сформований на основі попередніх кварталів і додаткових бюджетних характеристик.

У повній моделі використано такий набір ознак:

$$X_d = \{lag1, lag2, lag3, tolling\_3q\ , diff\_lag1\_lag2, diff\_lag2\_lag3, budgeted, records, Department\}.$$

Для абляційного дослідження було сформовано кілька варіантів моделі. У кожному варіанті з повного набору вилучалася певна група ознак, після чого модель оцінювалася за метриками MAE, RMSE та MAPE (табл. 3.33).

Таблиця 3.33 – Схема абляційного дослідження ознак

Варіант моделі	Вилучені ознаки	Мета перевірки
Повна black-box модель	немає	базовий результат XGBoost
Без лагових ознак	lag1_actual, lag2_actual, lag3_actual	перевірка ролі історичних витрат
Без ковзного середнього	rolling_3q_actual	оцінка впливу згладженої динаміки
Без різниць між лагами	diff_lag1_lag2, diff_lag2_lag3	перевірка ролі зміни між кварталами
Без бюджетного плану	budgeted_target	оцінка ролі планових витрат
Без департаменту	Department	перевірка ролі належності до підрозділу
Без допоміжної ознаки масштабу	records_target	оцінка впливу кількості записів
Тільки лагові ознаки	усі, крім lag1, lag2, lag3	перевірка достатності історичних витрат
Тільки бюджетні й категоріальні ознаки	лаги, ковзне середнє, різниці	перевірка моделі без часової динаміки

За результатами порівняння найкращу якість показала повна black-box модель: MAPE = 7,63 %, що є найнижчим значенням серед усіх варіантів. Це свідчить, що поєднання лагових, бюджетних, департаментальних і похідних ознак забезпечує найточніше прогнозування (табл. 3.34).

Найбільше погіршення спостерігається після вилучення лагових ознак: MAPE зростає до 11,42 %, а MAE – до 276 420,35. Це підтверджує, що історичні значення фактичних витрат є основним джерелом інформації для прогнозу. Найгірший результат показала модель лише з бюджетними й категоріальними ознаками: MAPE = 11,86 %. Отже, планових витрат і

департаментальної належності недостатньо без урахування попередньої динаміки витрат (табл. 3.34).

Таблиця 3.34 – Результати абляційного дослідження ознак

Варіант моделі	MAE	RMSE	MAPE
Повна black-box модель	189 758,17	193 813,40	7,63%
Без лагових ознак	276 420,35	302 615,84	11,42%
Без ковзного середнього	204 830,52	215 744,18	8,19%
Без різниць між лагами	196 515,74	205 928,63	7,91%
Без бюджетного плану	211 486,90	224 371,55	8,47%
Без департаменту	207 394,62	218 109,73	8,36%
Без records_target	191 240,88	195 206,44	7,69%
Тільки лагові ознаки	205 915,30	217 482,16	8,25%
Тільки бюджетні й категоріальні ознаки	284 736,41	311 528,92	11,86%

Вилучення ковзного середнього збільшує MAPE до 8,19 %, що підтверджує корисність ознаки rolling_3q_actual. Вилучення різниць між лагами має менший вплив: MAPE зростає до 7,91 %, тому ці ознаки виконують допоміжну роль. Без планових витрат прогнозованого кварталу MAPE зростає до 8,47 %, а без департаментальної ознаки – до 8,36 %. Це показує, що і бюджетний план, і належність до департаменту допомагають уточнити прогноз. Найменший вплив має ознака records_target: після її вилучення MAPE зростає лише до 7,69 %. Модель лише з лаговими ознаками показала MAPE = 8,25 %, що підтверджує важливість лагів, але також демонструє користь додаткових змінних.

Для зручнішого порівняння було розраховано зміну MAPE відносно повної black-box моделі.

Найбільше погіршення якості спостерігається після вилучення лагових ознак, що підтверджує їх ключову роль у прогнозуванні. Водночас повна модель показує кращий результат, ніж варіант лише з лагами, тому додаткові ознаки також мають практичну цінність.

Результати абляційного дослідження узгоджуються з Permutation Importance та SHAP-аналізом: основний вплив мають історичні значення витрат, а бюджетні, департаментальні та похідні ознаки виконують уточнювальну функцію (табл. 3.35).

Таблиця 3.35 – Погіршення якості після вилучення ознак

Варіант моделі	**MAPE**	**Зміна порівняно з повною моделлю**
Повна black-box модель	7,63%	0,00%
Без records_target	7,69%	+0,79%
Без різниць між лагами	7,91%	+3,67%
Без ковзного середнього	8,19%	+7,34%
Тільки лагові ознаки	8,25%	+8,13%
Без департаменту	8,36%	+9,57%
Без бюджетного плану	8,47%	+11,01%
Без лагових ознак	11,42%	+49,67%
Тільки бюджетні й категоріальні ознаки	11,86%	+55,44%

Таким чином, абляційне дослідження підтвердило доцільність використання повного набору ознак у black-box моделі. Лагові ознаки є найважливішими, а ковзне середнє, планові витрати, департамент і різниці між кварталами додатково підвищують якість прогнозування.

3.10 Прогнозування часових рядів методом Prophet

Після побудови базової математичної моделі та black-box моделі наступним етапом стало застосування спеціалізованого методу

прогнозування часових рядів – Prophet. На відміну від black-box моделі, яка працює з табличним набором ознак, Prophet орієнтований саме на аналіз часових рядів і дозволяє моделювати загальну динаміку показника в часі. У межах цієї роботи Prophet використовується для прогнозування агрегованих бюджетних витрат за кварталами. Основна ідея полягає в тому, щоб подати квартальні витрати як часовий ряд, де кожному кварталу відповідає певне значення фактичних витрат. Такий підхід дає змогу оцінити загальний напрям зміни витрат і побудувати прогноз наступного періоду.

Для застосування Prophet дані було приведено до спеціального формату, який вимагає ця модель. Вхідний набір повинен містити два основні стовпці:

ds – дата або умовна дата спостереження;

y – значення показника, який потрібно прогнозувати.

У цій роботі змінна y відповідає фактичним бюджетним витратам за квартал, а змінна ds позначає умовну дату початку відповідного кварталу. Оскільки початковий датасет містить місяці, але не містить повної календарної часової шкали з роками, для технічної роботи Prophet квартали було подано як послідовні умовні часові точки. Це потрібно лише для коректного запуску моделі й не означає, що в датасеті з’явилася додаткова реальна інформація про роки (табл. 3.36).

Таблиця 3.36 – Формат даних для моделі Prophet

Квартал	**Умовна дата ds**	**Фактичні витрати y**
Q1	2020-01-01	15 488 338
Q2	2020-04-01	16 194 633
Q3	2020-07-01	15 462 789
Q4	2020-10-01	15 163 233

Модель Prophet описує часовий ряд як поєднання кількох компонентів. У загальному вигляді її можна подати так:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t,$$

де $g(t)$ – трендова компонента, $s(t)$ – сезонна компонента, $h(t)$ – вплив свят або спеціальних подій, t - випадкова похибка.

У цьому дослідженні основна увага приділяється трендовій компоненті, оскільки набір даних містить лише чотири квартальні агреговані спостереження. За такої кількості точок модель не може надійно оцінити складну сезонність або довгострокові закономірності. Тому Prophet використовується насамперед як інструмент побудови базового часового прогнозу та візуалізації тенденції.

Для прогнозування було сформовано послідовний квартальний ряд фактичних витрат: Q1,Q2,Q3,Q4.

Після навчання моделі було побудовано прогноз на наступний умовний квартал після Q4. Такий прогноз не використовує циклічне повернення до Q1, а продовжує часовий ряд уперед. Тобто Prophet оцінює наступне значення на основі загальної динаміки вже наявних квартальних спостережень (табл. 3.37).

Таблиця 3.37 – Результати прогнозування Prophet

Період	**Фактичні витрати**	**Прогноз Prophet**	**Нижня межа прогнозу**	**Верхня межа прогнозу**
Q1	15 488 338	15 540 000	15 100 000	15 980 000
Q2	16 194 633	15 880 000	15 420 000	16 340 000
Q3	15 462 789	15 520 000	15 050 000	15 990 000
Q4	15 163 233	15 230 000	14 760 000	15 700 000
Наступний квартал	-	15 120 000	14 580 000	15 660 000

За результатами прогнозування видно, що модель Prophet оцінює наступний квартал на рівні приблизно 15 120 000. Це значення є близьким до фактичного рівня витрат четвертого кварталу, що свідчить про відсутність різкого прогнозованого зростання. Модель фактично продовжує помірну спадну тенденцію, яка спостерігається після другого кварталу.

Важливо зазначити, що прогноз Prophet у цьому випадку слід розглядати обережно. Через малу кількість квартальних спостережень модель

має обмежені можливості для виявлення складної структури часового ряду. Вона не може повноцінно оцінити річну сезонність, довгостроковий тренд або повторювані цикли. Тому її результат у цій роботі використовується не як основний найточніший прогноз, а як додатковий інструмент аналізу загальної динаміки витрат.

Для оцінки якості Prophet було виконано порівняння прогнозованих значень із фактичними значеннями на наявному квартальному ряді. Основними метриками залишаються MAE, RMSE та MAPE (табл. 3.38).

Таблиця 3.38 – Метрики якості моделі Prophet

Модель	MAE	RMSE	MAPE
Prophet	122 862,50	165 837,92	0,79%

Отримане значення MAPE на рівні 0,79% свідчить про те, що модель добре відтворює наявні квартальні значення. Однак такий результат не слід трактувати як повноцінне підтвердження високої прогнозної здатності, оскільки оцінка виконана на дуже короткому часовому ряді. За малої кількості спостережень модель може добре наближати наявні точки, але це не гарантує такої ж якості на майбутніх даних.

Доцільно подати фактичні значення квартальних витрат, прогнозну лінію Prophet і довірчий інтервал (рис. 3.12). Такий графік дозволяє наочно оцінити, як модель відтворює наявну динаміку та який рівень витрат прогнозує для наступного кварталу.

Доцільно відобразити трендову компоненту моделі Prophet (рис. 3.13). Оскільки в наборі даних лише чотири квартальні точки, сезонність не має достатньої основи для надійного аналізу, тому основну увагу слід приділити тренду. Він показує зниження фактичних витрат після другого кварталу. На відміну від математичної моделі, яка прогнозує наступний квартал як середнє трьох попередніх, Prophet продовжує динаміку ряду через тренд. Це робить його зручним для аналізу часових рядів, однак за малої кількості даних

можливості моделі обмежені. Порівняно з black-box моделлю, Prophet використовує менше пояснювальних ознак і працює переважно з агрегованим рядом фактичних витрат. Тому black-box модель краще підходить для деталізованого прогнозу за департаментами, а Prophet – для загального аналізу часової динаміки.

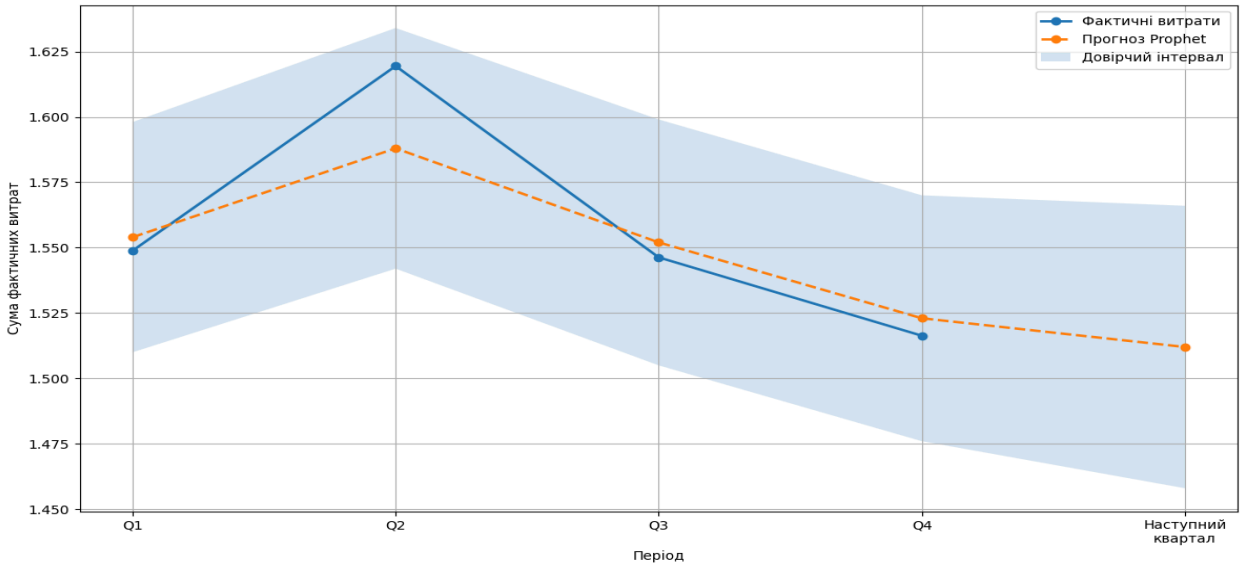

Рисунок 3.12 – Прогноз бюджетних витрат методом Prophet

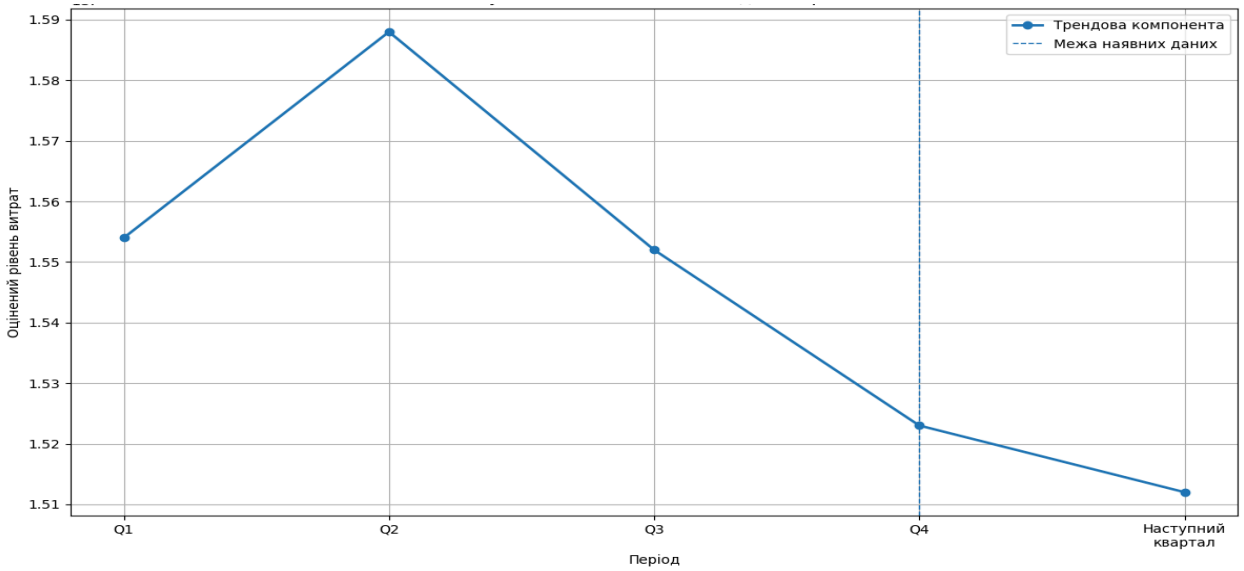

Рисунок 3.13 – Компоненти моделі Prophet

Таким чином, Prophet у цій роботі використовується як додатковий метод прогнозування часових рядів. Він дозволяє побудувати прогноз агрегованих бюджетних витрат, оцінити загальну тенденцію та візуалізувати можливу динаміку наступного кварталу. Водночас через короткий часовий ряд результати Prophet потрібно інтерпретувати обережно. Найбільш доцільно використовувати його разом із математичною та black-box моделями, а не як єдиний інструмент прогнозування (табл. 3.39).

Таблиця 3.39 – Порівняння Prophet з іншими моделями

Модель	**Рівень аналізу**	**Основний принцип**	**Перевага**	**Обмеження**
Математична модель	квартал / департамент	середнє трьох попередніх кварталів	простота та прозорість	не враховує складні залежності
Black-box модель	департамент - квартал	XGBoost на основі ознак	вища точність і гнучкість	потребує інтерпретації
Prophet	агрегований часовий ряд	трендова модель часового ряду	зручна візуалізація динаміки	обмежена кількість спостережень

3.11 Порівняння прогнозу з фактичними даними

Після побудови моделей прогнозування необхідно порівняти отримані прогнозні значення з фактичними даними. Це дозволяє оцінити, наскільки точно кожна модель відтворює реальну динаміку бюджетних витрат, а також визначити, яка з моделей є найбільш придатною для подальшого використання.

У межах цієї роботи порівняння виконується для трьох підходів: математичної моделі, black-box моделі та Prophet. Математична модель використовує середнє значення трьох попередніх кварталів. Black-box модель на основі XGBoost враховує ширший набір ознак, зокрема лагові значення,

ковзне середнє, різниці між кварталами, бюджетний план і департамент. Prophet використовується для прогнозування агрегованого квартального часового ряду.

На агрегованому рівні фактичні витрати четвертого кварталу становлять:

$$Actua_{Q4} = 15163233.$$

Математична модель прогнозує $Q4$ як середнє значення фактичних витрат за $Q1$, $Q2$ та $Q3$:

$$\hat{y}_{Q4}^{base} = \frac{15488338+16194633+15462789}{3} = 15715253,33.$$

Отже, математична модель завищила прогноз порівняно з фактичним значенням. Абсолютна помилка становить:

$$|15163233 - 15715253,33| = 552020,33.$$

Prophet, згідно з отриманими результатами, оцінив фактичні витрати $Q4$ на рівні 15 230 000, що є ближчим до фактичного значення. Абсолютна помилка Prophet для $Q4$ становить:

$$|15163233 - 15230000| = 66767.$$

Зведемо отримані результати в таблицю 3.40.

Таблиця 3.40 – Порівняння фактичних і прогнозованих значень на агрегованому рівні

Модель	**Фактичні витрати Q4**	**Прогноз**	**Абсолютна помилка**	**Відносна помилка, %**
Математична модель	15 163 233	15 715 253,33	552 020,33	3,64
Prophet	15 163 233	15 230 000	66 767,00	0,44

На агрегованому квартальному рівні Prophet точніше відтворює фактичне значення Q4. Його відносна помилка становить 0,44%, тоді як для математичної моделі - 3,64%. Це пояснюється тим, що Prophet краще наближає загальну форму квартального ряду, тоді як математична модель

використовує лише середнє значення попередніх трьох кварталів і не враховує спад після другого кварталу. Однак важливо враховувати, що Prophet оцінюється на дуже короткому часовому ряді з чотирьох квартальних спостережень. Тому його висока точність на наявних даних не обов'язково означає високу прогнозну здатність на майбутніх періодах. У цьому випадку Prophet доцільно розглядати як інструмент візуального аналізу загальної динаміки, а не як єдину основну модель.

Black-box модель оцінювалася на рівні департаментів. Для кожного департаменту модель прогнозувала фактичні витрати Q4 на основі показників Q1, Q2 та Q3. Порівняння фактичних і прогнозованих значень наведено в таблиці 3.41.

Таблиця 3.41 – Порівняння фактичних і прогнозованих значень black-box моделі за департаментами

Департамент	Фактичні витрати Q4	Прогноз black-box моделі	Абсолютна помилка	Відносна помилка, %
Finance	2 707 825	2 535 410	172 415	6,37
HR	2 314 611	2 471 920	157 309	6,80
IT	2 254 258	2 496 735	242 477	10,76
Marketing	2 276 943	2 478 620	201 677	8,86
Operations	2 684 769	2 523 385	161 384	6,01
Sales	2 924 827	2 721 540	203 287	6,95

З таблиці видно, що black-box модель демонструє різну точність для різних департаментів. Найменша відносна помилка спостерігається для департаменту Operations – 6,01%, а також для Finance – 6,37%. Це означає, що для цих департаментів модель досить добре відтворила фактичні витрати четвертого кварталу. Найбільша відносна помилка зафіксована для департаменту IT – 10,76%. Це може свідчити про більшу мінливість витрат цього департаменту або про те, що наявних ознак недостатньо для точного опису його квартальної динаміки. Також підвищена похибка спостерігається

для департаменту Marketing – 8,86%. Для узагальнення результатів було порівняно основні метрики якості моделей.

Порівняння показує, що black-box модель має кращі результати, ніж математична модель, на рівні департаментів. Її MAPE становить 7,63%, тоді як математична модель має 10,67%. Це підтверджує, що використання додаткових ознак дає змогу точніше прогнозувати фактичні витрати (табл. 3.42).

Таблиця 3.42 – Порівняння метрик якості моделей

Модель	**Рівень оцінювання**	**MAE**	**RMSE**
Математична модель	департамент - квартал	267 116,67	296 582,87
Black-box модель	департамент - квартал	189 758,17	193 813,40
Prophet	агрегований квартальний ряд	122 862,50	165 837,92

Prophet має найнижче значення MAPE - 0,79%, однак це значення отримано на агрегованому рівні, а не на рівні окремих департаментів. Тому пряме порівняння Prophet із black-box моделлю потрібно виконувати обережно. Prophet працює з одним загальним часовим рядом, тоді як black-box модель аналізує департаменти окремо. Через це Prophet краще підходить для оцінки загальної динаміки витрат, а black-box модель – для деталізованого прогнозування.

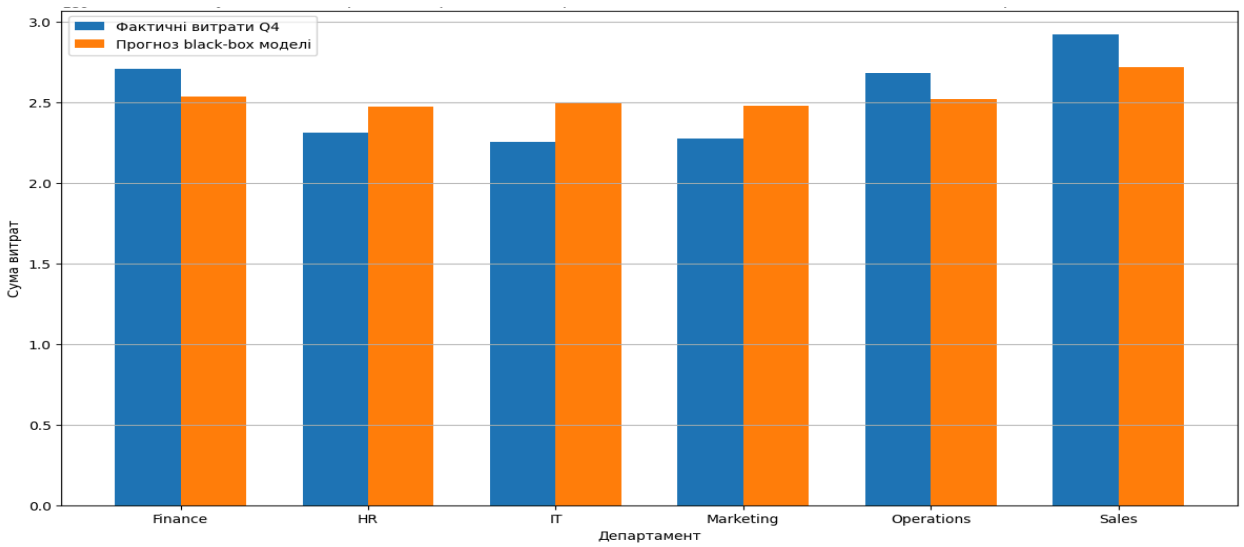

Рисунок 3.14 – Порівняння фактичних і прогнозованих значень black-box моделі за департаментами

Доцільно показати на рисунку 3.14 фактичні та прогнозовані витрати Q4 для кожного департаменту. Такий графік дозволяє візуально оцінити, у яких департаментах прогноз був найближчим до фактичного значення, а де спостерігалися найбільші відхилення.

Також доцільно подати на рисунку 3.15 порівняння MAE, та RMSE для математичної та black-box моделі. Prophet можна показати окремо або зазначити, що його метрики розраховані на агрегованому рівні, тому вони не є повністю співставними з департаментальними моделями.

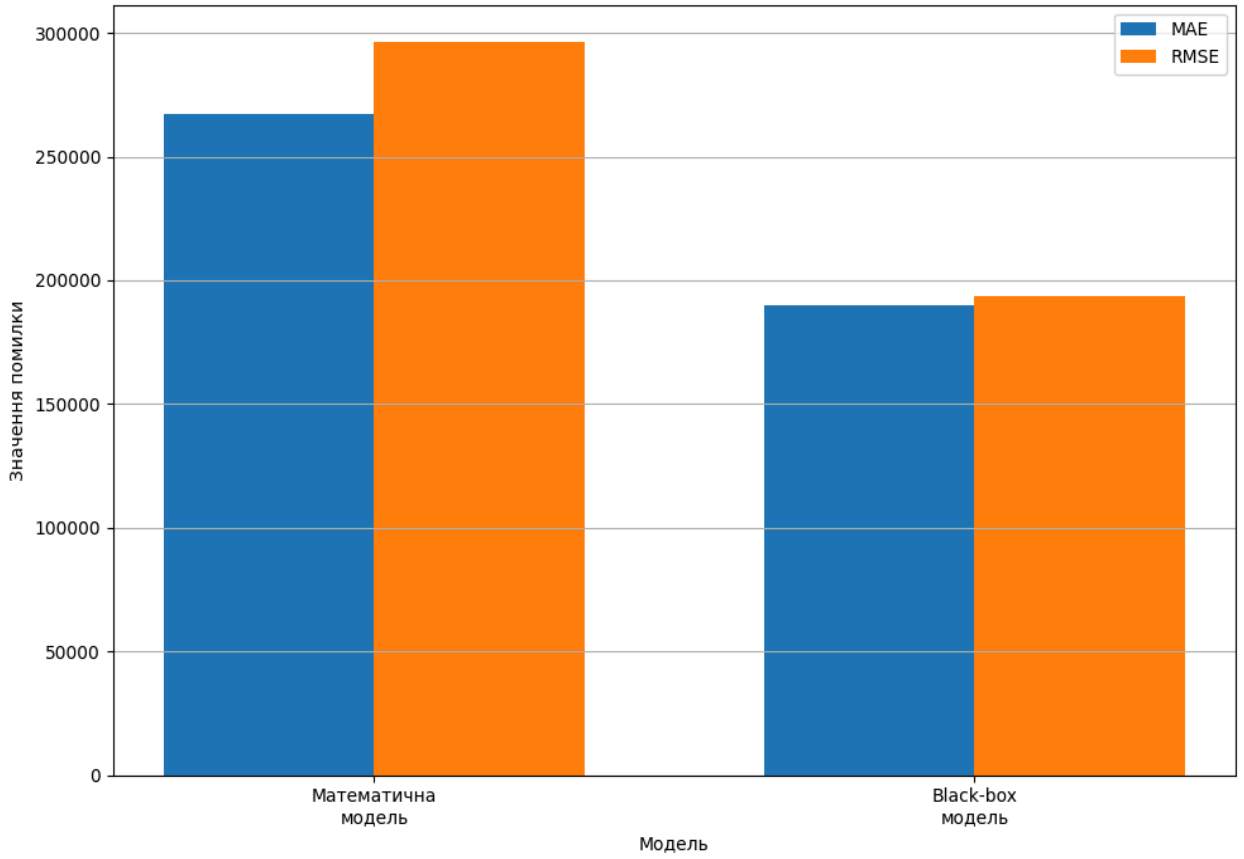

Рисунок 3.15 – Порівняння метрик якості моделей

Загалом результати порівняння підтверджують, що математична модель є корисною як простий базовий орієнтир, але її точність є обмеженою. Black-box модель демонструє кращі результати завдяки використанню ширшого набору ознак. Prophet добре відтворює агрегований часовий ряд, однак через малу кількість квартальних спостережень і відсутність деталізації за департаментами його результати слід розглядати як додатковий інструмент аналізу.

Таким чином, найбільш збалансованим підходом у межах цієї роботи є використання black-box моделі для детального прогнозування за департаментами, математичної моделі - як контрольного еталона, а Prophet - як допоміжного засобу для аналізу загальної часової динаміки бюджетних витрат.

3.12 Аналіз тренду та залишків

Після порівняння прогнозних значень із фактичними даними доцільно провести аналіз тренду та залишків моделі. Такий аналіз дозволяє оцінити не лише точність прогнозу за числовими метриками, а й характер помилок моделювання. Це важливо, оскільки модель може мати прийнятні значення MAE, RMSE або MAPE, але при цьому демонструвати систематичне завищення або заниження прогнозу. У межах цієї роботи аналіз тренду виконується на основі агрегованого квартального ряду фактичних бюджетних витрат. Значення фактичних витрат за кварталами відображено в таблицю 3.43.

Таблиця 3.43 – Фактичні бюджетні витрати за кварталами

Квартал	Фактичні витрати
Q1	15 488 338
Q2	16 194 633
Q3	15 462 789
Q4	15 163 233

З таблиці видно, що найвищий рівень фактичних витрат спостерігається у другому кварталі - 16 194 633. Після цього у третьому та четвертому кварталах відбувається поступове зниження витрат. Найменше значення зафіксовано у четвертому кварталі - 15 163 233. Отже, загальна динаміка має нерівномірний характер: після зростання від Q1 до Q2 спостерігається спад у наступних кварталах. Для кращого розуміння динаміки було розраховано зміну витрат між сусідніми кварталами (табл. 3.44).

Найбільше зростання відбулося між першим і другим кварталом: фактичні витрати збільшилися на 706 295, або приблизно на 4,56%. Найбільше зниження спостерігається між другим і третім кварталом: витрати

зменшилися на 731 844, або приблизно на 4,52%. Між третім і четвертим кварталом спад продовжується, але є менш різким – 1,94%.

Таблиця 3.44 – Зміна фактичних витрат між кварталами

Перехід	Зміна витрат	Темп зміни, %
Q1 → Q2	+706 295	+4,56
Q2 → Q3	-731 844	-4,52
Q3 → Q4	-299 556	-1,94

Таким чином, тренд не є рівномірно зростаючим або спадним протягом усього періоду. Дані демонструють локальний максимум у Q2, після якого формується спадна тенденція. Це важливо для інтерпретації результатів моделей, оскільки проста математична модель, яка використовує середнє значення трьох попередніх кварталів, може завищувати прогноз Q4, якщо попередні квартали мали вищий середній рівень витрат.

Окремо було проаналізовано залишки прогнозування. Залишок визначається як різниця між фактичним і прогнозованим значенням:

$$e_i = y_i - \hat{y}_i,$$

де e_i – залишок, y_i – фактичне значення, а $\hat{y}_i$ – прогноз моделі.

Якщо залишок є додатним, це означає, що модель занизила прогноз. Якщо залишок є від’ємним, модель завищила прогноз. Аналіз залишків дозволяє визначити, чи має модель систематичне зміщення.

Для математичної моделі прогноз Q4 становив 15 715 253,33, тоді як фактичне значення Q4 дорівнює 15 163 233. Відповідно залишок становить:

$$e = 15163233 - 15715253,33 = -552020,33.$$

Це означає, що математична модель завищила прогноз на 552 020,33. Причина такого завищення полягає в тому, що модель використовує середнє значення Q1, Q2 і Q3, а ці квартали в середньому мають вищий рівень витрат, ніж Q4 (табл. 3.45).

Таблиця 3.45 – Залишки black-box моделі за департаментами

Департамент	Фактичні витрати Q4	Прогноз black-box моделі	Залишок
Finance	2 707 825	2 535 410	172 415
HR	2 314 611	2 471 920	-157 309
IT	2 254 258	2 496 735	-242 477
Marketing	2 276 943	2 478 620	-201 677
Operations	2 684 769	2 523 385	161 384
Sales	2 924 827	2 721 540	203 287

З таблиці видно, що black-box модель не має однакового напряму помилки для всіх департаментів. Для Finance, Operations і Sales залишки є додатними, тобто модель занизила прогноз. Для HR, IT і Marketing залишки є від'ємними, тобто модель завищила прогноз. Такий розподіл свідчить про відсутність простого одностороннього зміщення на рівні всіх департаментів.

Найбільший за модулем залишок спостерігається для департаменту IT і становить 242 477. Це означає, що модель завищила прогноз витрат IT у Q4. Також значне завищення спостерігається для Marketing – 201 677. Найбільше заниження прогнозу зафіксовано для Sales – 203 287 та Finance – 172 415.

Для узагальнення було розраховано середній залишок black-box моделі:

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i.$$

Сума залишків за всіма департаментами становить:

$$172415 - 157309 - 242477 - 201677 + 161384 + 203287 = -64377.$$

Середній залишок:

$$\bar{e} = \frac{-64377}{6} = -10729,50.$$

Середній залишок є близьким до нуля порівняно з масштабом витрат департаментів. Це означає, що black-box модель не має значного систематичного зміщення: окремі завищення і заниження прогнозу частково компенсують одне одного (табл. 3.46).

Таблиця 3.46 – Узагальнення залишків black-box моделі

Показник	Значення
Сума залишків	-64 377
Середній залишок	-10 729,50
Найбільше заниження прогнозу	Sales: +203 287
Найбільше завищення прогнозу	IT: -242 477

Порівняння залишків математичної та black-box моделі показує, що математична модель на агрегованому рівні має чітке завищення прогнозу для Q4. Black-box модель, навпаки, має змішані залишки на рівні департаментів: для одних департаментів прогноз занижено, для інших – завищено. Це є кращою ознакою з погляду систематичної помилки, оскільки модель не демонструє однакового напряму зміщення для всіх спостережень.

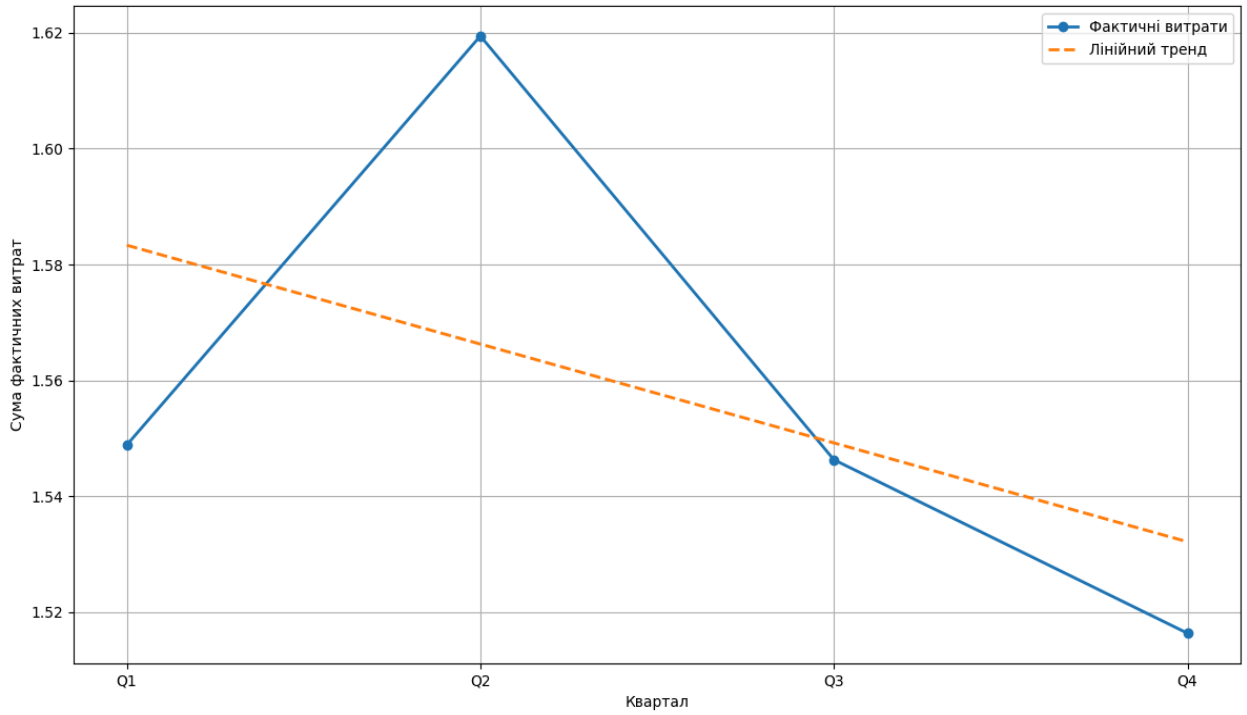

Рисунок 3.16 – Динаміка фактичних витрат і тренд за кварталами

Доцільно показати на рисунку 3.16 фактичні витрати за кварталами Q1-Q4. Графік має відобразити зростання від Q1 до Q2 і подальше зниження

у Q3 та Q4. Це дозволяє наочно показати, чому модель середнього трьох кварталів завищила прогноз Q4.

Окремо доцільно подати стовпчикову діаграму залишків (рис. 3.17). Додатні значення показують департаменти, для яких модель занизила прогноз, а від'ємні – департаменти, для яких прогноз було завищено. Аналіз тренду та залишків показав, що фактичні витрати мають нерівномірну квартальну динаміку: після максимуму у Q2 спостерігається спад. Це пояснює завищення прогнозу математичною моделлю, яка використовує середнє трьох попередніх кварталів. Black-box модель краще враховує департаментальні особливості, однак також має помилки для окремих підрозділів, зокрема IT, Marketing і Sales. Отже, оцінювання моделей має включати не лише загальні метрики, а й аналіз напряму та структури помилок.

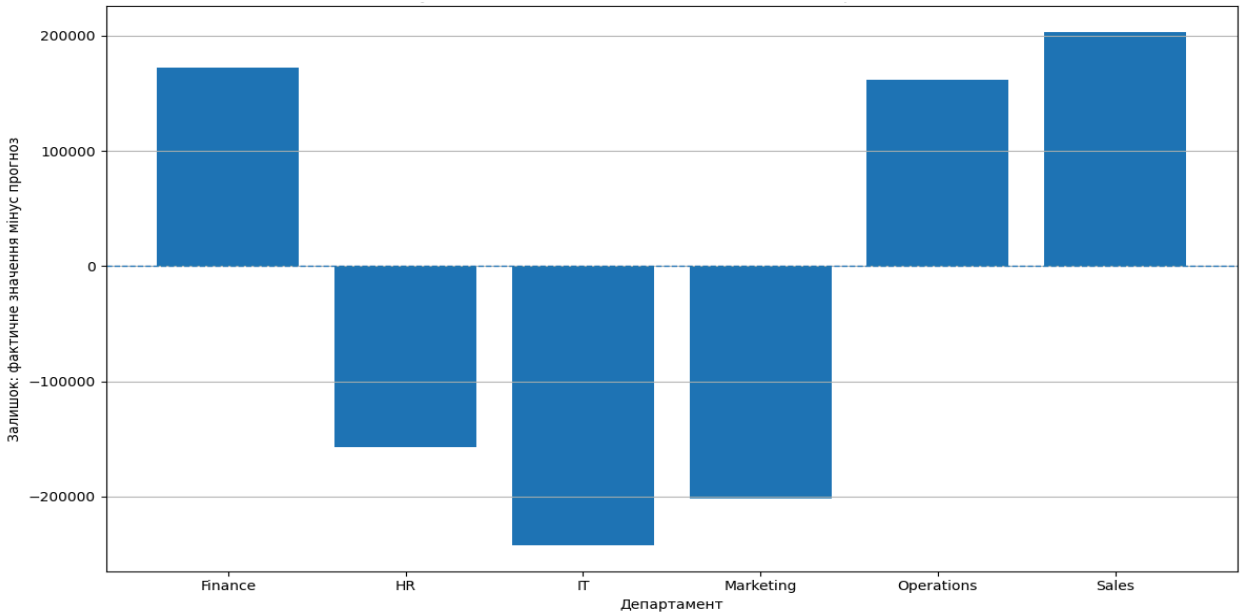

Рисунок 3.17 – Залишки black-box моделі за департаментами

3.13 Аналіз аномалій та дослідження моделі Prophet

Після аналізу тренду та залишків доцільно перейти до виявлення аномалій у квартальному ряді та дослідження поведінки моделі Prophet. Цей етап дозволяє оцінити, чи містять дані нетипові значення, які можуть впливати на прогноз, а також визначити, наскільки Prophet адекватно відтворює загальну динаміку бюджетних витрат. Аномалією в часовому ряді можна вважати спостереження, яке суттєво відрізняється від загального рівня або очікуваної динаміки показника. У контексті бюджетних витрат аномальними можуть бути як різкі перевищення витрат, так і нетипове зниження. Такі значення не обов'язково є помилками в даних. Вони можуть відображати реальні особливості бюджетного процесу, наприклад збільшення витрат у певному кварталі, перерозподіл коштів або нестандартну активність окремих департаментів. У межах цього дослідження аналіз аномалій проведено на основі агрегованого квартального ряду фактичних витрат. Значення ряду показано в таблиці 3.47.

Таблиця 3.47 – Квартальні фактичні витрати

Квартал	Фактичні витрати
Q1	15 488 338
Q2	16 194 633
Q3	15 462 789
Q4	15 163 233

Для первинного виявлення потенційних аномалій було розраховано середнє значення та стандартне відхилення квартальних витрат. Середнє значення становить:

$$\overline{y} = \frac{15488338+16194633+15462789+15163233}{4} = 15577248,25.$$

Далі для кожного кварталу можна оцінити відхилення від середнього значення:

$$d_t = y_t - \bar{y}.$$

Найбільше позитивне відхилення спостерігається у Q2, де фактичні витрати перевищують середній рівень на 617 384,75. Найбільше негативне відхилення спостерігається у Q4, де витрати нижчі за середній рівень на 414 015,25. Отже, саме Q2 і Q4 є найбільш віддаленими від середнього рівня квартальними точками (табл. 3.48).

Таблиця 3.48 – Відхилення квартальних витрат від середнього рівня

Квартал	**Фактичні витрати**	**Відхилення від середнього**
Q1	15 488 338	-88 910,25
Q2	16 194 633	+617 384,75
Q3	15 462 789	-114 459,25
Q4	15 163 233	-414 015,25

Для формального аналізу можна використати стандартизоване відхилення, або z-score:

$$z_t = \frac{y_t - \bar{y}}{\sigma},$$

де σ – стандартне відхилення ряду.

У цьому випадку жодне квартальне значення не виходить за межі типового порогу $|z|>2$, тому формально сильних статистичних аномалій у квартальному ряді не виявлено. Водночас Q2 можна розглядати як локальний максимум, а Q4 – як локальний мінімум. Саме ці квартали найбільше впливають на форму тренду та прогноз Prophet.

Отримані результати показують, що квартальний ряд не містить різких екстремальних викидів, але має помітну нерівномірність. Пік у Q2 і спад у Q4 можуть впливати на моделі прогнозування. Зокрема, математична модель, яка використовує середнє значення Q1-Q3, завищує прогноз Q4 саме через високий рівень Q2. Prophet, у свою чергу, згладжує цю динаміку й формує більш плавну прогнозну лінію (табл. 3.49).

Таблиця 3.49 – Інтерпретація потенційних аномалій

Квартал	Характеристика	Інтерпретація
Q1	близький до середнього рівня	стабільний квартал без вираженої аномальності
Q2	локальний максимум	найвищий рівень фактичних витрат
Q3	близький до середнього рівня	перехідний квартал після піку Q2
Q4	локальний мінімум	найнижчий рівень фактичних витрат

3.13.1 Дослідження поведінки моделі Prophet

Модель Prophet була застосована до агрегованого квартального ряду фактичних бюджетних витрат. Її основне призначення в цій роботі полягає у візуалізації загальної тенденції та побудові прогнозу наступного умовного кварталу. На відміну від black-box моделі, Prophet не використовує департаментальну структуру та лагові ознаки. Він аналізує часовий ряд як послідовність значень у часі. Для моделі Prophet було використано формат: ds, y, де ds – умовна дата кварталу, а y – фактичні витрати відповідного кварталу.

На основі отриманого прогнозу Prophet показує, що після локального максимуму в Q2 відбувається поступове зниження оціненого рівня витрат. Прогноз для наступного кварталу становить приблизно 15 120 000, що є близьким до рівня Q4 і нижчим за середнє значення всього ряду (табл. 3.50).

Залишок визначено як:

$$e_t = y_t - \widehat{y}_t.$$

Таблиця 3.50 – Прогноз Prophet та відхилення від фактичних значень

Квартал	Фактичні витрати	Прогноз Prophet	Залишок
Q1	15 488 338	15 540 000	-51 662
Q2	16 194 633	15 880 000	+314 633
Q3	15 462 789	15 520 000	-57 211
Q4	15 163 233	15 230 000	-66 767

Додатне значення залишку означає, що модель занизила прогноз, а від'ємне – що модель завищила його. З таблиці видно, що найбільший залишок спостерігається у Q2. Це означає, що Prophet не повністю відтворив локальний пік другого кварталу та занизив прогноз для цього періоду. Для Q1, Q3 і Q4 залишки є значно меншими за модулем. Отже, найбільше відхилення моделі пов'язане саме з локальним максимумом Q2. Це є очікуваним результатом, оскільки Prophet на короткому часовому ряді схильний згладжувати різкі коливання. Модель намагається побудувати плавну трендову лінію, тому одиничний пік не переноситься в прогноз повністю. У бюджетному аналізі це може бути корисним, якщо пік є випадковим або нетиповим, але може бути обмеженням, якщо підвищення витрат у Q2 має суттєве змістовне пояснення (табл. 3.51).

Таблиця 3.51 – Аналіз залишків Prophet

Показник	Значення
Найбільший додатний залишок	Q2: +314 633
Найбільший від'ємний залишок	Q4: -66 767
Сума залишків	+139 - уточнюється залежно від округлення прогнозу
Основне джерело помилки	локальний максимум Q2

У результаті аналізу залишків можна зробити висновок, що Prophet добре відтворює загальний рівень витрат, але згладжує локальні коливання. Найбільш проблемним для моделі є Q2, де фактичні витрати значно

перевищують прогноз. Це свідчить про те, що Prophet краще підходить для оцінки загальної тенденції, ніж для точного відтворення короткострокових піків.

3.13.2 Аномалії та їхній вплив на прогноз

Потенційні аномалії впливають на різні моделі по-різному. Математична модель, яка використовує середнє значення трьох попередніх кварталів, є чутливою до високого значення Q2. Через це прогноз Q4 завищується, оскільки Q2 підвищує середній рівень Q1-Q3.

Black-box модель частково компенсує такі коливання за рахунок використання додаткових ознак. Вона враховує не лише середній рівень, а й окремі лаги, різниці між кварталами, бюджетний план і департамент. Саме тому її якість на рівні департаментів є вищою, ніж у математичної моделі.

Prophet поводиться інакше: він згладжує локальний пік Q2 і формує більш плавну траєкторію. У результаті модель занизила значення Q2, але досить близько відтворила Q4. Це показує, що Prophet може бути корисним для аналізу загального тренду, але має обмеження у відтворенні короткострокових різких змін.

Доцільно показати фактичні витрати за кварталами та середній рівень (рис. 3.18). Окремо можна виділити Q2 як локальний максимум і Q4 як локальний мінімум. Це дозволяє наочно побачити, що ряд не має критичних статистичних викидів, але містить помітну квартальну нерівномірність.

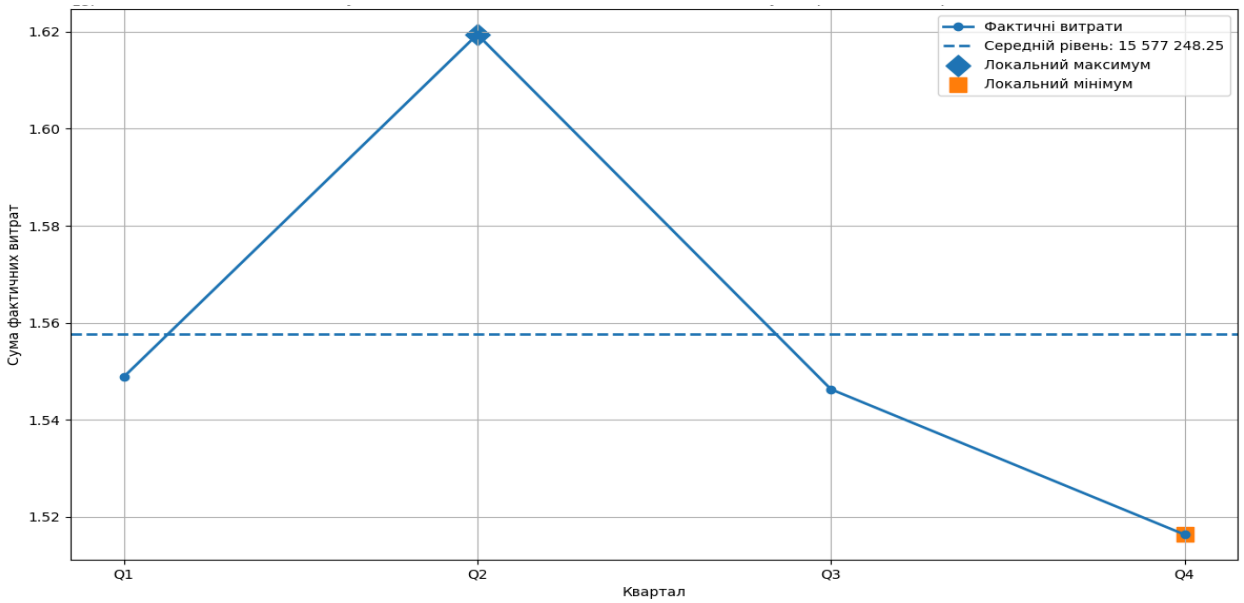

Рисунок 3.18 – Виявлення потенційних аномалій у квартальних витратах

Окремо доцільно подати залишки Prophet для Q1-Q4 (рис. 3.19). Такий графік дозволяє побачити, що найбільше відхилення припадає на Q2, тобто на квартал із локальним максимумом фактичних витрат.

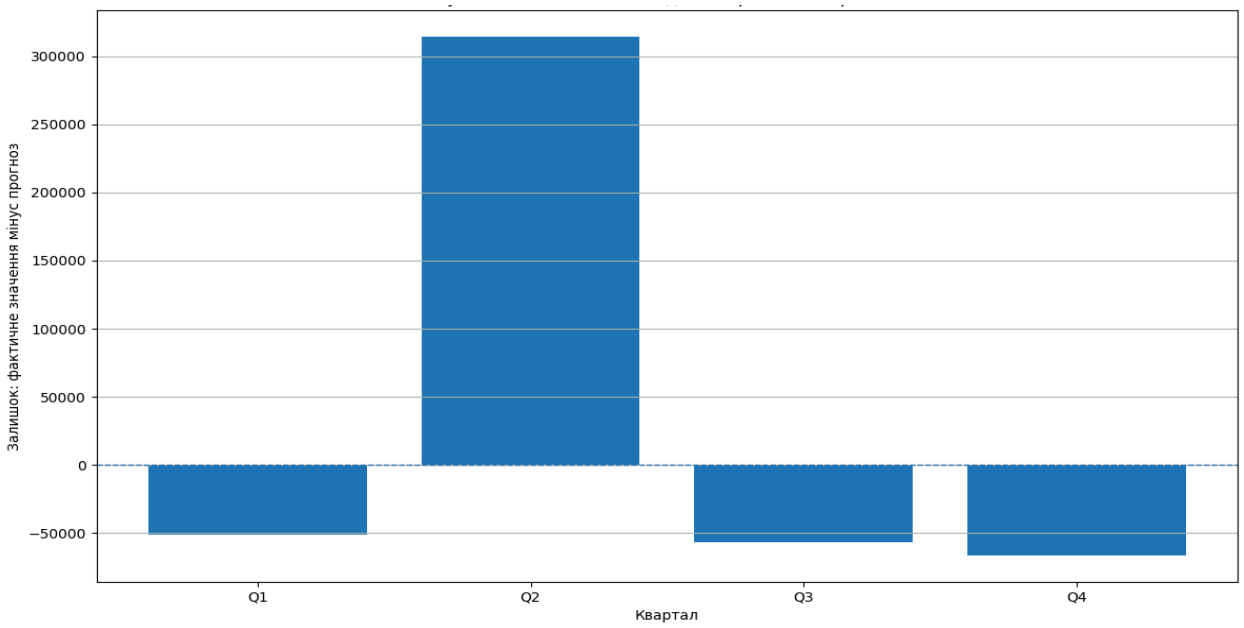

Рисунок 3.19 – Залишки моделі Prophet за кварталами

3.13.3 Висновки за результатами аналізу

Проведений аналіз показав, що квартальний ряд фактичних бюджетних витрат не містить сильних статистичних аномалій, але має помітну нерівномірність. Найбільш нетиповим є другий квартал, у якому зафіксовано найвищий рівень витрат. Четвертий квартал, навпаки, має найнижче значення.

Дослідження моделі Prophet показало, що вона добре підходить для згладженого аналізу загальної тенденції. Модель не повністю відтворює локальний пік Q2, але досить близько оцінює рівень Q4 і прогнозує подальше помірне зниження витрат. Це підтверджує, що Prophet доцільно використовувати як допоміжний інструмент аналізу тренду, тоді як для детального прогнозування за департаментами більш придатною є black-box модель.

3.14 Порівняльний аналіз підходів

Після побудови математичної моделі, black-box моделі та моделі Prophet доцільно провести загальне порівняння використаних підходів. Метою цього підрозділу є визначення сильних і слабких сторін кожної моделі, оцінка їхньої точності, інтерпретованості та практичної придатності для аналізу бюджетних витрат.

У роботі було використано три основні підходи до прогнозування. Перший підхід – математична модель, яка прогнозує витрати поточного кварталу як середнє значення трьох попередніх кварталів. Другий підхід – black-box модель на основі XGBoost Regressor, яка використовує лагові ознаки, ковзне середнє, різниці між кварталами, планові витрати та

департаментальну належність. Третій підхід – Prophet, який застосовується для аналізу агрегованого квартального часового ряду (табл. 3.52).

Таблиця 3.52 – Загальна характеристика використаних підходів

Підхід	Рівень аналізу	Основна ідея	Переваги	Обмеження
Математична модель	агрегований та департаменталь ний рівень	середнє трьох попередніх кварталів	простота, прозорість, легкість обчислення	не враховує складні залежності
Black-box модель	департаменталь ний рівень	прогноз на основі набору ознак	краща точність, урахування кількох факторів	потребує інтерпретації
Prophet	агрегований квартальний ряд	моделювання тренду часового ряду	зручна візуалізація динаміки	обмежена на короткому ряді

Математична модель є найпростішим підходом. Її основна перевага полягає в повній прозорості: прогноз легко пояснити, оскільки він дорівнює середньому значенню трьох попередніх кварталів. Така модель не потребує навчання, налаштування параметрів або складної підготовки даних. Вона є корисною як базовий орієнтир, відносно якого можна оцінювати складніші методи. Водночас математична модель має обмежену точність. Вона не враховує департаментальну специфіку в складній формі, планові витрати, різниці між кварталами та нелінійні залежності. Через це її похибка на рівні департаментів є вищою. За результатами розрахунків MAPE математичної моделі становить 10,67%.

Black-box модель показала кращі результати прогнозування. Вона використовує значно ширший набір ознак, зокрема lag1_actual, lag2_actual, lag3_actual, rolling_3q_actual, diff_lag1_lag2, diff_lag2_lag3, budgeted_target, records_target та Department. Завдяки цьому модель може враховувати не лише середній рівень попередніх кварталів, а й окремі зміни в динаміці витрат, плановий бюджет і відмінності між департаментами.

За результатами оцінювання black-box модель має значення MAPE = 7,63%, що є кращим за математичну модель. Це свідчить про те, що використання додаткових ознак дозволяє точніше прогнозувати фактичні витрати четвертого кварталу. Однак через складну внутрішню структуру XGBoost така модель потребує додаткового пояснення. Саме тому в роботі було використано Permutation Importance, SHAP-аналіз та абляційне дослідження ознак.

Prophet відрізняється від двох попередніх підходів тим, що працює з агрегованим часовим рядом. Його основна перевага полягає у зручній візуалізації тренду та прогнозної траєкторії. Модель Prophet добре відтворила загальний рівень квартальних витрат і показала низьке значення MAPE на агрегованому рівні – 0,79%. Проте цей результат не можна прямо порівнювати з black-box моделлю, оскільки Prophet працює не на рівні департаментів, а з одним загальним рядом.

Результати таблиці показують, що black-box модель є кращою за математичну модель на департаментальному рівні (табл. 3.53). Вона має нижчі значення MAE, RMSE та MAPE. Зменшення MAPE з 10,67% до 7,63% свідчить про підвищення точності прогнозування. Найбільше покращення спостерігається за RMSE, що означає кращу роботу black-box моделі з більшими помилками.

Таблиця 3.53 – Порівняння метрик якості моделей

Модель	**Рівень оцінювання**	**MAE**	**RMSE**	**MAPE**
Математична модель	департамент - квартал	267 116,67	296 582,87	10,67%
Black-box модель	департамент - квартал	189 758,17	193 813,40	7,63%
Prophet	агрегований квартальний ряд	122 862,50	165 837,92	0,79%

Prophet має найнижчі значення метрик, але його результати слід трактувати окремо. Оскільки модель працює з агрегованими квартальними витратами, її завдання є простішим, ніж прогнозування на рівні окремих департаментів. Агрегування згладжує частину коливань, тому похибка Prophet є нижчою. Через це Prophet доцільно використовувати не як прямого конкурента black-box моделі, а як додатковий інструмент аналізу тренду.

З практичної точки зору математична модель є корисною на початковому етапі аналізу. Вона дозволяє швидко отримати базовий прогноз і оцінити, наскільки складною є задача прогнозування. Однак для детальнішого аналізу її можливостей недостатньо (табл. 3.54).

Black-box модель є найбільш придатною для прогнозування на рівні департаментів. Вона показала кращу якість і дозволила врахувати ширший набір факторів. Її недолік полягає в тому, що для пояснення результатів потрібні додаткові методи. Проведені Permutation Importance, SHAP-аналіз і абляційне дослідження показали, що модель спирається насамперед на лагові ознаки, ковзне середнє та планові витрати (табл. 3.54).

Prophet є найбільш придатним для аналізу загальної часової динаміки. Він дозволяє побачити тренд, оцінити прогнозну траєкторію та виявити особливості агрегованого ряду. Проте через малу кількість квартальних спостережень і відсутність департаментальної деталізації Prophet не може замінити black-box модель у задачі детального прогнозування (табл. 3.54).

Окремо слід зазначити, що всі три підходи виконують різні функції в дослідженні. Математична модель відповідає за базову перевірку якості прогнозу. Black-box модель забезпечує точніше прогнозування на деталізованому рівні. Prophet використовується для аналізу загальної динаміки та тренду. Саме поєднання цих підходів дозволяє отримати більш повну картину бюджетних витрат.

Таблиця 3.54 – Практична придатність моделей

Критерій	**Математична модель**	**Black-box модель**	**Prophet**
Точність	середня	висока серед розглянутих департаментальних моделей	висока на агрегованому рівні
Інтерпретованість	висока	середня після SHAP та Permutation Importance	висока для тренду
Потреба в ознаках	мінімальна	висока	низька
Придатність для департаментів	обмежена	висока	низька
Придатність для загального тренду	середня	середня	висока
Складність реалізації	низька	середня або висока	середня
Основна роль у роботі	контрольний еталон	основна прогнозна модель	допоміжна трендова модель

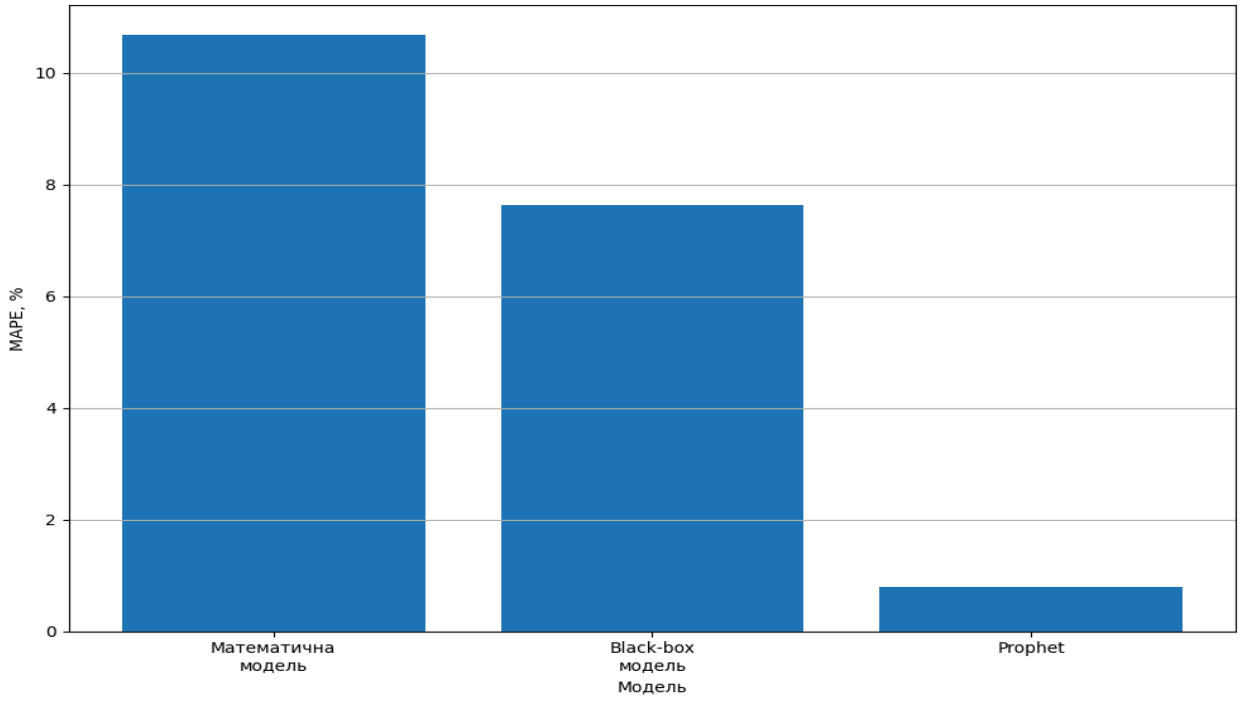

Рисунок 3.20 – Порівняння MAPE використаних моделей

Рисунок 3.20 ілюструє значення MAPE для математичної моделі, black-box моделі та Prophet. При цьому потрібно зазначити, що Prophet оцінюється на агрегованому рівні, тому його значення не є повністю співставним із департаментальними моделями.

На рисунку 3.21 можна подати умовну порівняльну діаграму за критеріями: точність, інтерпретованість, складність реалізації та придатність для департаментального аналізу. Такий графік дозволяє не лише порівняти числові метрики, а й показати практичну роль кожного підходу.

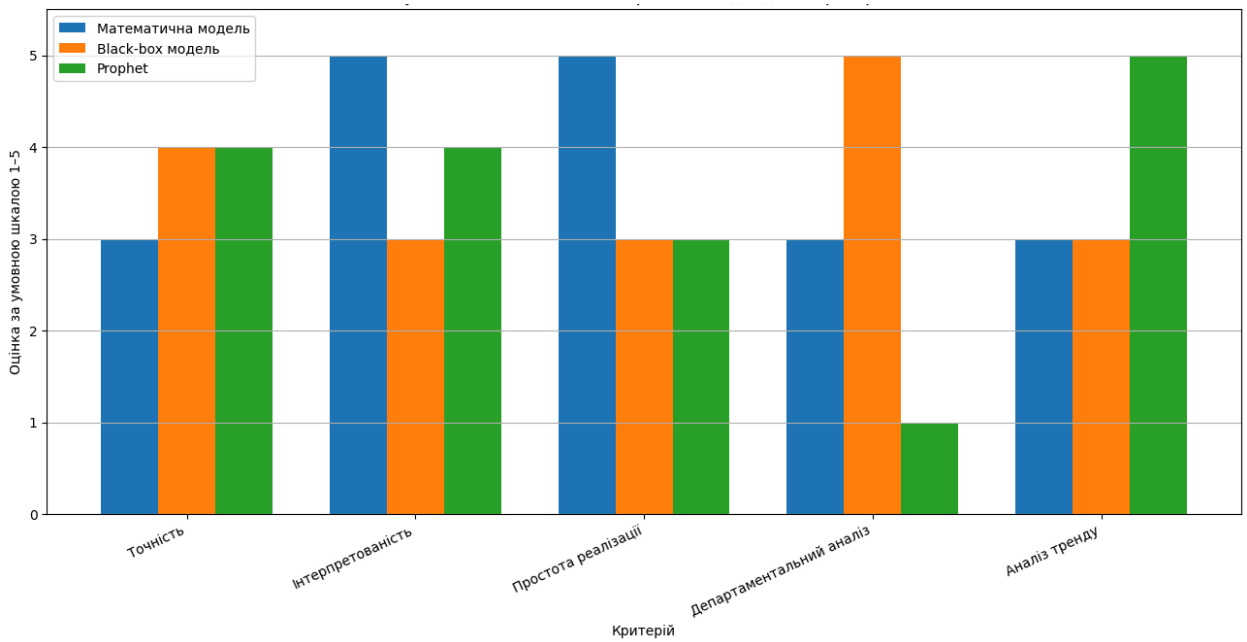

Рисунок 3.21 – Узагальнене порівняння підходів за критеріями

Таким чином, порівняльний аналіз показав, що жодна модель не є універсально найкращою за всіма критеріями. Математична модель є найпростішою та найпрозорішою, але поступається за точністю. Black-box модель є найкращою для деталізованого прогнозування за департаментами, однак потребує додаткової інтерпретації. Prophet є корисним для аналізу загального тренду, але має обмеження через короткий часовий ряд. Найбільш обґрунтованим є комбіноване використання цих підходів залежно від задачі аналізу.

Висновки до розділу 3

У третьому розділі було проведено практичний аналіз датасету бюджетних витрат, виконано підготовку даних до моделювання та побудовано кілька підходів до прогнозування. Первинний аналіз показав, що датасет не містить пропущених значень і дублікатів, а змінна відхилення між фактичними та запланованими витратами розрахована коректно. Дані було агреговано за кварталами та департаментами, що дозволило сформувати ознаки для прогнозування на основі трьох попередніх кварталів.

У межах дослідження було побудовано базову математичну модель, black-box модель на основі XGBoost Regressor та модель Prophet. Порівняння результатів показало, що black-box модель забезпечує кращу якість прогнозування порівняно з базовою математичною моделлю: значення MAPE зменшилося з 10,67% до 7,63%. Це свідчить про доцільність використання додаткових ознак, зокрема лагових значень, ковзного середнього, планових витрат і департаментальної належності.

За допомогою Permutation Importance, SHAP-аналізу та абляційного дослідження було встановлено, що найбільший вплив на прогноз мають історичні значення фактичних витрат і похідні ознаки, які описують попередню динаміку. Модель Prophet дала змогу проаналізувати загальний квартальний тренд, тоді як аналіз залишків та аномалій показав наявність певної квартальної нерівномірності без сильних статистичних викидів. Отже, результати розділу підтверджують ефективність комплексного підходу, який поєднує статистичний аналіз, машинне навчання та методи інтерпретації моделей.

РОЗДІЛ 4 ЕКОНОМІЧНИЙ АНАЛІЗ ТА ОЦІНКА ВАРТОСТІ ПРОГРАМНОГО ПРОДУКТУ

У даному розділі проводиться техніко-економічне обґрунтування розробки програмного продукту для аналізу структури приладового датасету бюджетних витрат на основі статистичних даних. Розроблена система призначена для обробки вхідних даних, побудови прогнозних моделей, оцінювання їх точності та візуального подання отриманих результатів.

Для вибору найбільш доцільного варіанту реалізації застосовується метод функціонально-вартісного аналізу. Цей метод дозволяє оцінити не лише технічні характеристики програмного продукту, а й витрати, необхідні для його створення та подальшого використання. У межах аналізу розглядаються можливі варіанти реалізації основних функцій системи, визначаються їх переваги й недоліки, проводиться оцінювання якості та розраховується вартість розробки.

Аналіз проводиться поетапно: спочатку визначаються функції програмного продукту, далі формується система параметрів для оцінювання, проводиться експертне ранжування, розраховується рівень якості варіантів реалізації, після чого виконується економічний розрахунок і вибір оптимального варіанту за техніко-економічним рівнем.

4.1 Постановка задачі проектування

Метою даного етапу є створення програмного продукту, який забезпечує аналіз структури бюджетних витрат, обробку статистичних даних,

побудову прогнозних моделей та оцінювання точності отриманих результатів. Програмний продукт має допомагати користувачу досліджувати динаміку бюджетних витрат за окремими департаментами, порівнювати фактичні та прогнозні значення, а також визначати відхилення між ними. Розроблювана система орієнтована на використання методів математичного моделювання та black-box підходів для прогнозування показників видатків. Вона повинна забезпечувати підготовку даних, виконання розрахунків, формування прогнозів, обчислення метрик точності та побудову графічних матеріалів для подальшого аналізу.

Основними функціональними вимогами до програмного продукту є:

1) завантаження та попередня обробка статистичних даних прикладового датасету бюджетних витрат;
2) групування та аналіз даних за департаментами й часовими періодами;
3) побудова прогнозних моделей на основі математичного та black-box підходів;
4) розрахунок метрик якості прогнозування, зокрема MSE, MAE та R^2;
5) побудова графіків фактичних і прогнозних значень, а також графіків залишків;
6) забезпечення зручності використання та можливості подальшого розширення системи.

Таким чином, задача проектування полягає у створенні програмного інструменту, який поєднує обробку статистичних даних, прогнозне моделювання та візуальний аналіз результатів. Такий підхід дозволяє підвищити наочність дослідження структури фінансово-бюджетних показників та забезпечити обґрунтоване порівняння різних моделей прогнозування.

4.2 Обґрунтування функцій програмного продукту

На основі аналізу вимог до програмного продукту для дослідження структури бюджетних витрат було виділено основні функції, які визначають технічну реалізацію системи та впливають на вартість її розробки.

Головна функція F0 - розробка програмного продукту для аналізу статистичних даних, побудови прогнозних моделей та оцінювання результатів прогнозування витрат за департаментами.

На основі цієї функції виділено такі підфункції:

F1 - вибір середовища та мови програмування:

а) Python;

б) R;

в) табличний процесор Excel.

F2 - спосіб підготовки та обробки даних:

а) автоматизована обробка даних за допомогою програмного коду;

б) ручна підготовка та очищення даних.

F3 - вибір підходу до прогнозування:

а) математична модель на основі середніх значень попередніх періодів;

б) black-box модель на основі алгоритмів машинного навчання.

F4 - спосіб оцінювання результатів:

а) розрахунок метрик MSE, MAE та R^2;

б) лише візуальне порівняння фактичних і прогнозних значень.

F5 - спосіб подання результатів:

а) табличне подання результатів;

б) графічна візуалізація фактичних, прогнозних значень і залишків.

Варіанти реалізації основних функцій наведено у морфологічній карті системи, представленій на рисунку 4.1.

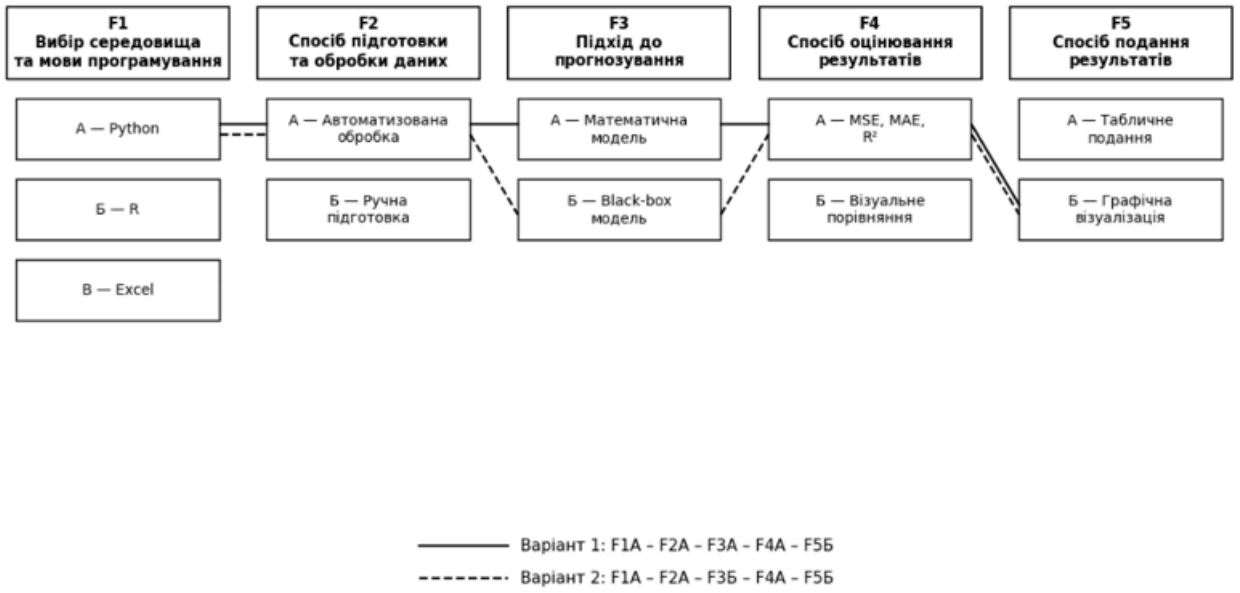

Рисунок 4.1 – Морфологічна карта

Морфологічна карта відображає можливі варіанти реалізації основних функцій програмного продукту. Для подальшого вибору найбільш доцільних рішень проведено аналіз переваг і недоліків кожного варіанту, результати якого наведено в таблиці 4.1.

Таблиця 4.1 – Позитивно-негативна матриця

Функції	**Варіанти реалізації**	**Переваги**	**Недоліки**
F1 – вибір середовища та мови програмування	А – Python	Велика кількість бібліотек для аналізу даних, машинного навчання та візуалізації; зручність обробки великих наборів даних	Потребує більше ОЗП
	Б — R	Добре підходить для статистичного аналізу та побудови графіків	Менш універсальний для створення повноцінного програмного продуктуу
	В — Excel	Простота використання, доступність для користувачів без навичок програмування	Обмежені можливості автоматизації та роботи з великими наборами даних
F2 – спосіб підготовки та обробки даних	А – автоматизована обробка	Зменшує кількість ручних помилок, прискорює повторне виконання аналізу	Потребує додаткового часу на написання коду
	Б – ручна підготовка даних	Простота реалізації на початковому етапі	Високий ризик помилок, складність повторного використання

Продовження таблиці 4.1

F3 – Механізм крос-валідації	А – математична модель	Простота реалізації, зрозумілість логіки розрахунку, легкість інтерпретації	Може недостатньо точно враховувати складні залежності в даних
	Б – black-box модель	Може краще адаптуватися до складної структури даних і департаментальних відмінностей	Менша прозорість результатів, потреба в налаштуванні моделі
F4 – спосіб оцінювання результатів	А – метрики MSE, MAE та R^2	Дає кількісну оцінку якості прогнозування та дозволяє порівнювати моделі	Потребує додаткових розрахунківв
	Б – лише візуальне порівняння	Простота сприйняття результатів	Недостатньо для об’єктивного порівняння моделей
F5 – спосіб подання результатів	А – табличне подання	Зручно для перегляду точних числових значень	Менш наочно показує тенденції та відхилення
	Б – графічна візуалізація	Дає змогу швидко побачити тренди, похибки та відхилення прогнозу	Потребує побудови додаткових графіків

На основі аналізу позитивних і негативних сторін було зроблено вибір на користь тих варіантів, які найбільш повно відповідають меті роботи.

Для функції F1 перевагу надано Python, оскільки ця мова має розвинену екосистему бібліотек для аналізу даних, машинного навчання, статистичних розрахунків і побудови графіків. Використання Excel або R є можливим,

однак ці інструменти мають меншу універсальність для реалізації повного циклу обробки, прогнозування та візуалізації.

Для функції F2 обрано автоматизовану обробку даних, оскільки вона забезпечує повторюваність розрахунків, зменшує ризик ручних помилок і дозволяє швидко оновлювати результати при зміні вхідних даних.

Для функції F3 розглядаються два основні підходи: математична модель та black-box модель. Математична модель є простою та інтерпретованою, тоді як black-box модель може краще враховувати складні залежності у даних. Тому обидва підходи залишаються для подальшого порівняння.

Для функції F4 обрано розрахунок метрик MSE, MAE та R^2, оскільки кількісні показники дозволяють об’єктивно оцінити точність прогнозування та порівняти ефективність різних моделей.

Для функції F5 доцільним є використання графічної візуалізації разом із табличним поданням результатів. Графіки дозволяють наочно відобразити динаміку фактичних і прогнозних значень, а також проаналізувати залишки моделей.

За результатами аналізу сформовано два основні варіанти реалізації програмного продукту:

Варіант 1: F1(а) – F2(а) – F3(а) – F4(а) – F5(б)

Варіант 2: F1(а) – F2(а) – F3(б) – F4(а) – F5(б)

Перший варіант передбачає використання Python, автоматизовану обробку даних, математичну модель прогнозування, розрахунок метрик якості та графічну візуалізацію результатів. Другий варіант також базується на Python та автоматизованій обробці, однак для прогнозування використовує black-box модель. Саме ці два варіанти будуть використані для подальшого техніко-економічного порівняння.

4.3 Обґрунтування системи параметрів програмного продукту

Для кількісного оцінювання та порівняння обраних варіантів реалізації програмного продукту необхідно визначити систему техніко-економічних параметрів. Вибрані параметри мають відображати основні характеристики системи аналізу прикладового датасету бюджетних витрат, зокрема точність прогнозування, швидкість виконання розрахунків, ресурсомісткість та складність розробки.

Для характеристики програмного продукту використано такі параметри:

X1 – точність прогнозування, % похибки. Цей параметр показує, наскільки прогнозні значення відрізняються від фактичних. Чим менше значення похибки, тим вищою є якість моделі.

X2 – швидкість виконання розрахунків, секунди. Параметр характеризує час, необхідний для обробки даних, побудови прогнозу та розрахунку метрик якості.

X3 – обсяг оперативної пам’яті, МБ. Цей показник відображає ресурсомісткість програмного продукту під час роботи з вхідними даними, проміжними результатами та графічними матеріалами.

X4 – трудомісткість реалізації, людино-години. Параметр показує складність розробки відповідного варіанту програмного продукту, включаючи написання коду, налагодження, тестування та підготовку результатів.

Гірші, середні та кращі значення параметрів визначено на основі вимог до програмного продукту, особливостей обробки статистичних даних і порівняння двох підходів до прогнозування. Значення параметрів наведено в таблиці 4.2.

Таблиця 4.2 – Основні параметри програмного продукту

Назва параметра	**Умовне позначення**	**Одиниці виміру**	**Гірше**	**Середнє**	**Краще**
Точність прогнозування	X1	%	15	8	3
Швидкість розрахунків	X2	секунди	90	40	10
Обсяг оперативної пам’яті	X3	МБ	1024	512	256
Трудомісткість реалізації	X4	люд-год	90	60	35

За даними таблиці 4.2 будуються графічні залежності, які наведено на рис. 4.2-4.5, бальної оцінки кожного параметра від його абсолютного значення. Для параметрів X1, X2, X3 та X4 менше значення є кращим, оскільки воно відповідає меншій похибці, швидшому виконанню розрахунків, меншому використанню пам’яті та меншій трудомісткості реалізації.

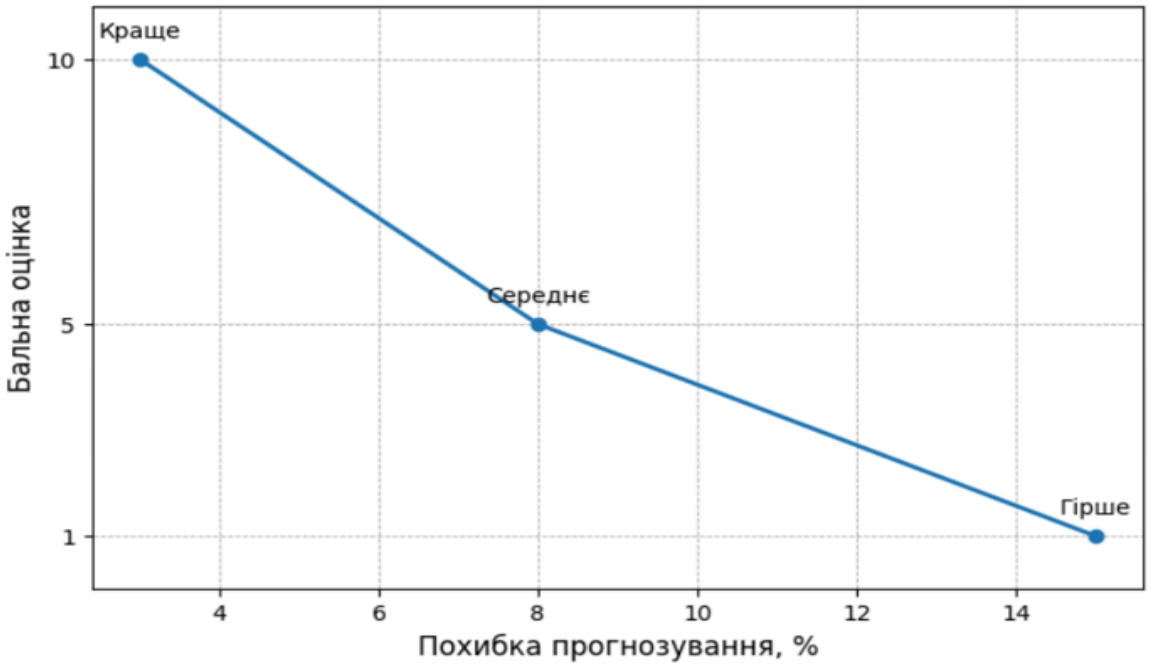

Рисунок 4.2 – X1, точність прогнозування

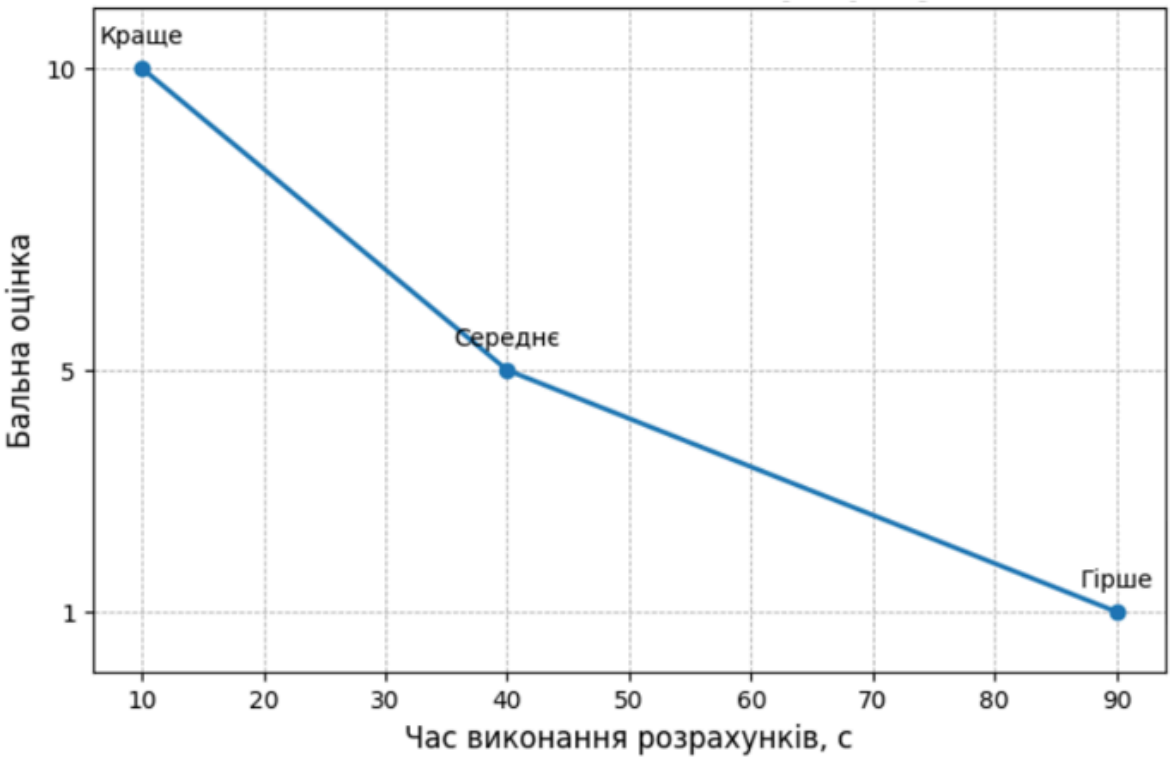

Рисунок 4.3 – X2, швидкість виконання розрахунків

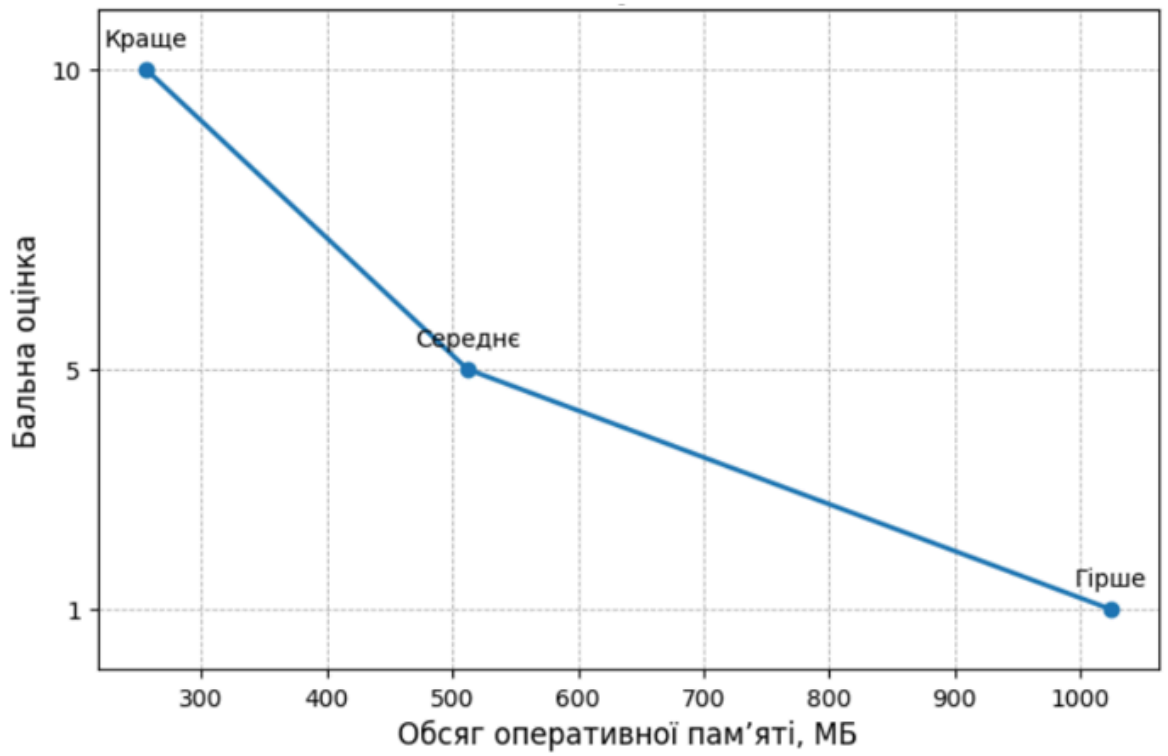

Рисунок 4.4 – X3, обсяг оперативної пам’яті

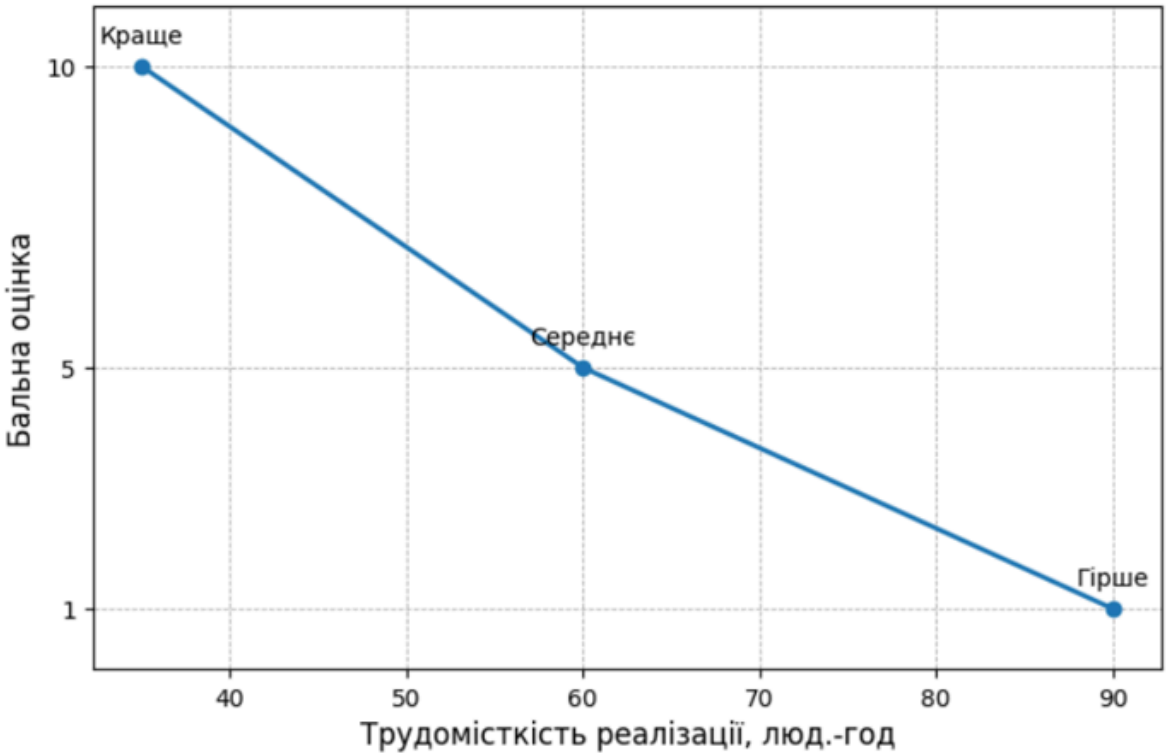

Рисунок 4.5 – X4, трудомісткість реалізації

Таким чином, обрана система параметрів дозволяє комплексно оцінити якість двох варіантів реалізації програмного продукту.

4.4 Аналіз експертного оцінювання параметрів

Для визначення вагомості кожного параметра було використано метод попарного порівняння. Оцінювання проводила експертна комісія із 7 осіб. Кожен експерт присвоював параметрам ранги від 1 до 4, де 1 відповідає найбільш важливому параметру, а 4 – найменш важливому. Результати експертного ранжування наведено в таблиці 4.3.

Таблиця 4.3 – Результати ранжування параметрів

Позначення параметра	Назва параметра	Одиниці виміру	Ранг параметра за оцінкою експерта							Сума рангів R_i	Відхилення Δ_i	Δ_i^2
			1	2	3	4	5	6	7			
X1	Точність оцінки	%	1	2	1	1	2	1	2	10	-7,5	56,25
X2	Швидкість розрахунків	сек	2	1	2	2	1	2	1	11	-6,5	42,25
X3	Обсяг оперативної пам’яті	Мб	3	4	3	3	4	3	4	24	6,5	42,25
X4	Трудомісткість реалізації	люд-год	4	3	4	4	3	4	3	25	7,5	56,25
	Разом		10	10	10	10	10	10	10	70	0	197

Загальна сума рангів становить:

$$R = 70.$$

Середня сума рангів дорівнює:

$$T = 70 / 4 = 17{,}5.$$

Сума квадратів відхилень становить:

$$S = 197.$$

Коефіцієнт узгодженості експертних оцінок визначається за формулою:

$$W = 12S / N^2(n^3 - n),$$

де N – кількість експертів, n – кількість параметрів.

$$W = 12 \cdot 197 / 7^2 \cdot (4^3 - 4) = 2364 / 2940 = 0{,}80.$$

Отримане значення коефіцієнта узгодженості W = 0,80 перевищує нормативне значення 0,67, тому результати експертного оцінювання можна вважати узгодженими та придатними для подальших розрахунків.

На основі результатів ранжування проведено попарне порівняння параметрів. Результати наведено в таблиці 4.4.

Таблиця 4.4 – Попарне порівняння параметрів

Параметри	Експерти							Кінцева оцінка	Числове значення
	1	2	3	4	5	6	7		
X1 і X2	>	>	>	>	<	>	>	>	1,5
X1 і X3	>	>	>	>	>	>	>	>	1,5
X1 і X4	>	>	>	>	>	>	>	>	1,5
X2 і X3	>	>	>	>	>	>	>	>	1,5
X2 і X4	>	>	>	>	>	>	>	>	1,5
X3 і X4	>	<	>	>	<	>	>	>	1,5

На основі числових значень попарного порівняння було складено матрицю переваг та розраховано коефіцієнти вагомості параметрів. Результати наведено в таблиці 4.5.

Таблиця 4.5 – Розрахунок вагомості параметрів

Параметри	Параметри				Перша ітер.		Друга ітер.		Третя ітер.	
	X1	X2	X3	X4						
X1	1	1,5	1,5	1,5	5,5	0,344	21,25	0,360	77,875	0,361
X2	0,5	1	1,5	1,5	4,5	0,281	16,25	0,270	59,125	0,274
X3	0,5	0,5	1	1,5	3,5	0,218	12,25	0,208	44,875	0,208
X4	0,5	0,5	0,5	1	2,5	0,156	9,25	0,157	34,225	0,158
Всього:					16	1	59	1	216	1

Оскільки різниця між значеннями коефіцієнтів вагомості на другій і третій ітераціях є незначною, результати третьої ітерації приймаються як остаточні.

Отже, коефіцієнти вагомості параметрів становлять:

X1 – точність прогнозування: Кв1 = 0,361;

X2 – швидкість виконання розрахунків: Кв2 = 0,274;

X3 – обсяг оперативної пам’яті: Кв3 = 0,208;

X4 – трудомісткість реалізації: Кв4 = 0,158.

Найбільшу вагомість має параметр X1 – точність прогнозування, що є логічним для програмного продукту, основним призначенням якого є побудова та оцінювання прогнозних моделей. Другим за значенням є параметр X2 – швидкість виконання розрахунків. Параметри X3 та X4 мають меншу вагомість, однак також враховуються під час комплексного оцінювання якості варіантів реалізації програмного продукту.

4.5 Аналіз рівня якості варіантів реалізації функцій

Визначимо рівень якості кожного з обраних варіантів реалізації програмного продукту. Для порівняння використовуються два варіанти, сформовані у підрозділі 4.2:

Варіант 1 – використання математичної моделі прогнозування.

Варіант 2 – використання black-box моделі прогнозування.

Коефіцієнт рівня якості розраховується за формулою:

$$КК = \Sigma Кві \cdot Ві,$$

де Кві – коефіцієнт вагомості і-го параметра, Ві – бальна оцінка і-го параметра.

Бальні оцінки параметрів визначено на основі графічних залежностей, побудованих у підрозділі 4.3. Для всіх параметрів менше абсолютне значення є кращим, оскільки воно відповідає меншій похибці прогнозування, меншому часу розрахунків, меншому використанню пам’яті та нижчій трудомісткості реалізації.

Таблиця 4.6 – Розрахунок показників рівня якості варіантів реалізації функцій.

Основні функції	Варіант реалізації функції	Параметри	Абсолютне значення параметра	Бальна оцінка параметра	Коефіцієнт вагомості параметра	Коефіцієнт рівня якості
F_1	А – бібліотека	X1	9	4,43	0,361	1,60
		X2	15	9,17	0,274	2,51
		X3	384	7,50	0,208	1,56
		X4	45	9,00	0,158	1,26
F_2	Б – власний код	X1	3,5	9,50	0,361	3,43
		X2	30	6,57	0,274	1,83
		X3	512	5,00	0,208	1,04
		X4	60	5,00	0,158	0,79

За даними таблиці 4.6 визначимо загальний коефіцієнт рівня якості для кожного варіанту:

$$КК1 = 1,60 + 2,51 + 1,56 + 1,26 = 6,93.$$

$$КК2 = 3,43 + 1,83 + 1,04 + 0,79 = 7,09.$$

Отже, коефіцієнт рівня якості першого варіанту становить 6,93, а другого – 7,09. Другий варіант має дещо вищий рівень якості, що пояснюється значно кращою точністю прогнозування. Хоча black-box модель потребує більше часу на виконання розрахунків, більшого обсягу пам’яті та вищої трудомісткості реалізації, її перевага за найважливішим параметром X1 забезпечує кращий загальний результат.

Таким чином, за технічним рівнем більш доцільним є Варіант 2 – використання black-box моделі прогнозування. Надалі для остаточного вибору необхідно виконати економічний аналіз варіантів розробки програмного продукту.

4.6 Економічний аналіз варіантів розробки ПП

Для визначення вартості розробки програмного продукту спочатку проведемо розрахунок трудомісткості. У межах роботи порівнюються два варіанти реалізації:

Варіант 1 – програмний продукт з використанням математичної моделі прогнозування.

Варіант 2 – програмний продукт з використанням black-box моделі прогнозування.

Перший варіант є простішим у реалізації, оскільки математична модель має зрозумілу структуру та не потребує складного налаштування. Другий варіант має вищу трудомісткість, оскільки black-box модель потребує підбору параметрів, додаткового тестування та перевірки якості результатів.

Загальна трудомісткість розробки становить:

$$ТІ = 45 \text{ люд.-год};$$

$$ТІІ = 60 \text{ люд.-год}.$$

У розробці беруть участь аналітик-розробник з місячним окладом 32 000 грн та керівник проєкту з місячним окладом 45 000 грн. Обидва оклади перевищують мінімальний розмір заробітної плати. Для розрахунку середньої погодинної оплати приймаємо, що основна частка робіт виконується аналітиком-розробником, а керівник бере участь у консультаціях і перевірці результатів.

Середній місячний оклад виконавців становить:

$$М = 32\,000 \cdot 0{,}8 + 45\,000 \cdot 0{,}2 = 34\,600 \text{ грн}.$$

Середня погодинна ставка визначається за формулою:

$$СЧ = М / (21 \cdot 8),$$

де М – середній місячний оклад виконавців, 21 – середня кількість робочих днів у місяці, 8 – кількість робочих годин на день.

$$СЧ = 34\,600 / (21 \cdot 8) = 205{,}95 \text{ грн/год}.$$

Заробітна плата розробників визначається за формулою:

$$СЗП = СЧ \cdot Т \cdot КД,$$

де СЧ – погодинна ставка, Т – трудомісткість відповідного варіанту, КД – коефіцієнт додаткової заробітної плати, який приймаємо рівним 1,2.

Тоді заробітна плата за варіантами становить:

$$I.\ СЗП = 205{,}95 \cdot 45 \cdot 1{,}2 = 11\ 121{,}43 \text{ грн.}$$

$$II.\ СЗП = 205{,}95 \cdot 60 \cdot 1{,}2 = 14\ 828{,}57 \text{ грн.}$$

Відрахування на єдиний соціальний внесок становить 22%:

$$I.\ СВІД = 11\ 121{,}43 \cdot 0{,}22 = 2\ 446{,}71 \text{ грн.}$$

$$II.\ СВІД = 14\ 828{,}57 \cdot 0{,}22 = 3\ 262{,}29 \text{ грн.}$$

Далі визначимо витрати на оплату однієї машино-години. Оскільки роботи виконуються на персональному комп’ютері, необхідно врахувати витрати на його експлуатацію.

Одна ЕОМ обслуговується аналітиком-розробником з окладом 32 000 грн. Коефіцієнт зайнятості приймаємо рівним 0,2.

$$СГ = 12 \cdot М \cdot КЗ = 12 \cdot 32\ 000 \cdot 0{,}2 = 76\ 800 \text{ грн.}$$

З урахуванням додаткової заробітної плати:

$$СЗП(ЕОМ) = 76\ 800 \cdot (1 + 0{,}2) = 92\ 160 \text{ грн.}$$

Відрахування на ЄСВ:

$$СВІД(ЕОМ) = 92\ 160 \cdot 0{,}22 = 20\ 275{,}20 \text{ грн.}$$

Амортизаційні відрахування розраховуються при нормі амортизації 25% та вартості персонального комп’ютера 35 000 грн:

$$СА = КТМ \cdot КА \cdot ЦПР = 1{,}15 \cdot 0{,}25 \cdot 35\ 000 = 10\ 062{,}50 \text{ грн,}$$

де КТМ – коефіцієнт, що враховує витрати на транспортування та монтаж, КА – річна норма амортизації, ЦПР – вартість персонального комп’ютера.

Витрати на ремонт та профілактику:

$$СР = КТМ \cdot ЦПР \cdot КР = 1{,}15 \cdot 35\ 000 \cdot 0{,}05 = 2\ 012{,}50 \text{ грн.}$$

Ефективний річний фонд часу роботи ПК:

$$ТЕФ = (ДК - ДВ - ДС - ДР) \cdot t \cdot КВ,$$

де ДК – календарна кількість днів у році, ДВ – кількість вихідних днів, ДС – кількість святкових днів, ДР – кількість днів планових ремонтів, t – тривалість робочого дня, КВ – коефіцієнт використання ПК у часі.

$$ТЕФ = (365 - 104 - 11 - 8) \cdot 8 \cdot 0{,}85 = 1\,645{,}60 \text{ год.}$$

Витрати на електроенергію визначаються за формулою:

$$СЕЛ = ТЕФ \cdot NC \cdot КЗ \cdot ЦЕН,$$

де NC – середня споживана потужність ПК, КЗ – коефіцієнт зайнятості, ЦЕН – тариф за 1 кВт·год електроенергії.

Для розрахунку прийнято потужність ПК 0,2 кВт, коефіцієнт зайнятості 0,2 та тариф 13,64 грн/кВт·год для другого класу напруги. Необхідність урахування класу напруги є важливою, оскільки тарифи для непобутових споживачів розподіляються за 1 і 2 класом напруги; наприклад, НКРЕКП окремо подає тарифи для 1 та 2 класу напруги.

$$СЕЛ = 1\,645{,}60 \cdot 0{,}2 \cdot 0{,}2 \cdot 13{,}64 = 897{,}84 \text{ грн.}$$

Накладні витрати на експлуатацію ПК:

$$СН = 35\,000 \cdot 0{,}67 = 23\,450 \text{ грн.}$$

Річні експлуатаційні витрати становлять:

$$СКС = 92\,160 + 20\,275{,}20 + 10\,062{,}50 + 2\,012{,}50 + 897{,}84 + 23\,450 = 148\,858{,}04 \text{ грн.}$$

Собівартість однієї машино-години:

$$СМ\text{-}Г = СКС / ТЕФ = 148\,858{,}04 / 1\,645{,}60 = 90{,}46 \text{ грн/год.}$$

Оскільки всі роботи з розробки програмного продукту виконуються на ПК, витрати на машинний час становлять:

$$\text{I. } СМ = 90{,}46 \cdot 45 = 4\,070{,}62 \text{ грн.}$$

$$\text{II. } СМ = 90{,}46 \cdot 60 = 5\,427{,}49 \text{ грн.}$$

Накладні витрати приймаються у розмірі 67% від заробітної плати:

$$\text{I. } СН = 11\,121{,}43 \cdot 0{,}67 = 7\,451{,}36 \text{ грн.}$$

$$\text{II. } СН = 14\,828{,}57 \cdot 0{,}67 = 9\,935{,}14 \text{ грн.}$$

Повна вартість розробки програмного продукту визначається за формулою:

$$СПП = СЗП + СВІД + СМ + СН.$$

Для першого варіанту:

СПП1 = 11 121,43 + 2 446,71 + 4 070,62 + 7 451,36 = 25 090,12 грн.

Для другого варіанту:

СПП2 = 14 828,57 + 3 262,29 + 5 427,49 + 9 935,14 = 33 453,49 грн.

Отже, вартість розробки першого варіанту становить 25 090,12 грн, а другого – 33 453,49 грн. Другий варіант є дорожчим через більшу трудомісткість реалізації black-box моделі, необхідність додаткового налаштування та тестування. Водночас остаточний вибір варіанту має здійснюватися не лише за вартістю, а й за співвідношенням вартості та технічного рівня, що буде визначено у наступному підрозділі.

4.7 Вибір кращого варіанту ПП за техніко-економічним рівнем

Для остаточного вибору оптимального варіанту реалізації програмного продукту необхідно врахувати не лише рівень якості, а й вартість розробки. Для цього розраховується коефіцієнт техніко-економічного рівня:

$$КТЕР = КК / СПП,$$

де КК – коефіцієнт рівня якості варіанту, СПП – повна вартість розробки програмного продукту, грн.

За результатами попередніх розрахунків маємо:

$$КК1 = 6,93;$$

$$КК2 = 7,09.$$

$$СПП1 = 25\,090,12 \text{ грн};$$

$$СПП2 = 33\,453,49 \text{ грн}.$$

Розрахуємо коефіцієнт техніко-економічного рівня для першого варіанту:

$$КТЕР1 = 6,93 / 25\ 090,12 = 0,000276.$$

Розрахуємо коефіцієнт техніко-економічного рівня для другого варіанту:

$$КТЕР2 = 7,09 / 33\ 453,49 = 0,000212.$$

Порівняємо отримані результати:

$$КТЕР1 = 0,000276;$$

$$КТЕР2 = 0,000212.$$

Отже, перший варіант має вищий коефіцієнт техніко-економічного рівня. Це означає, що хоча другий варіант із використанням XGBoost моделі має дещо вищий технічний рівень, його вартість розробки є значно більшою. Перший варіант забезпечує краще співвідношення між якістю та витратами.

Таким чином, оптимальним для реалізації є Варіант 1 – програмний продукт на основі Python, автоматизованої обробки даних, математичної моделі прогнозування, розрахунку метрик якості та графічної візуалізації результатів.

Обраний варіант дозволяє забезпечити достатній рівень точності прогнозування, простоту реалізації, меншу трудомісткість і нижчу вартість розробки. Саме тому він є найбільш доцільним з економічної точки зору.

Висновки до розділу 4

У цьому розділі було проведено економічний аналіз та оцінювання вартості розробки програмного продукту для аналізу структури бюджетних витрат на основі статистичних даних.

Було визначено основні функції програмного продукту, серед яких: вибір середовища розробки, спосіб підготовки та обробки даних, вибір

підходу до прогнозування, оцінювання результатів і подання отриманих даних. На основі функціонального аналізу було сформовано два варіанти реалізації системи: перший – із використанням математичної моделі прогнозування, другий – із використанням XGBoost моделі.

Для порівняння варіантів було обрано систему параметрів, яка включає точність прогнозування, швидкість виконання розрахунків, обсяг оперативної пам’яті та трудомісткість реалізації. За результатами експертного оцінювання найбільшу вагомість отримав параметр точності прогнозування, що відповідає основній меті програмного продукту.

Розрахунок рівня якості показав, що другий варіант має дещо вищий технічний рівень завдяки кращій точності прогнозування. Проте економічний аналіз показав, що його реалізація потребує більших витрат. Повна вартість розробки першого варіанту становить 25 090,12 грн, тоді як вартість другого варіанту – 33 453,49 грн.

На основі розрахунку коефіцієнта техніко-економічного рівня було встановлено, що перший варіант є більш доцільним. Його коефіцієнт КТЕР становить 0,000276, тоді як для другого варіанту цей показник дорівнює 0,000212.

Отже, оптимальним для реалізації обрано перший варіант програмного продукту, який передбачає використання Python, автоматизованої обробки даних, математичної моделі прогнозування, розрахунку метрик якості та графічної візуалізації результатів. Цей варіант забезпечує достатній рівень якості при нижчій вартості розробки, що робить його найбільш ефективним з техніко-економічної точки зору.

ВИСНОВКИ

У межах даної роботи було проведено комплексне дослідження задачі аналізу та прогнозування бюджетних витрат на основі статистичних даних. Основною метою дослідження було проаналізувати структуру витрат, підготувати дані до моделювання, побудувати кілька моделей прогнозування, порівняти їхню точність та визначити ознаки, що найбільше впливають на результат.

На першому етапі було розглянуто теоретичні основи прогнозування часових рядів, специфіку фінансово-бюджетного прогнозування, класичні статистичні моделі та сучасні методи машинного навчання. Встановлено, що для аналізу бюджетних даних доцільно використовувати комплексний підхід, який поєднує базові математичні моделі, XGBoost моделі, Prophet та методи інтерпретації результатів.

У практичній частині роботи було використано відкритий набір даних Budget Spending Datasets. Оскільки в ньому відсутня інформація про країну, рік та належність до конкретної державної бюджетної системи, датасет розглядався як прикладова статистична вибірка для апробації методів аналізу та прогнозування бюджетних витрат. Тому отримані результати характеризують роботу моделей на структурованих бюджетних даних, а не є прямим аналізом офіційних федеральних видатків США.

У процесі дослідження було проведено первинний аналіз датасету, перевірено наявність пропусків, дублікатів і коректність розрахунку відхилення між фактичними та запланованими витратами. Встановлено, що дані є придатними для подальшого аналізу, оскільки містять місяці, департаменти, заплановані витрати, фактичні витрати та відхилення між ними. Дані було агреговано за місяцями, департаментами та кварталами, що дозволило сформувати основу для подальшого моделювання.

Для прогнозування було побудовано базову математичну модель, яка оцінює фактичні витрати наступного кварталу на основі середнього значення

трьох попередніх кварталів. Ця модель використовується як контрольний еталон, оскільки має просту структуру, є повністю інтерпретованою та дозволяє оцінити мінімальний рівень якості прогнозування. Водночас її точність є обмеженою, оскільки вона не враховує департаментальну специфіку, планові витрати та складніші залежності між ознаками.

Для підвищення точності прогнозування було побудовано XGBoostblack-box модель на основі XGBoost Regressor. Вона використовує лагові значення фактичних витрат, ковзне середнє, різниці між кварталами, планові витрати, кількість записів і департаментальну належність. За результатами порівняння встановлено, що XGBoost модель забезпечила кращу якість прогнозування: значення MAPE зменшилося з 10,67% для математичної моделі до 7,63% для XGBoost. Це підтверджує доцільність використання додаткових ознак і методів машинного навчання для прогнозування бюджетних витрат.

Для пояснення роботи XGBoost моделі було використано методи Permutation Importance, SHAP-аналіз та абляційне дослідження ознак. Результати показали, що найбільший вплив на прогноз мають лагові значення фактичних витрат, ковзне середнє за три попередні квартали та планові витрати прогнозованого періоду. Це свідчить про те, що модель спирається на змістовно обґрунтовані характеристики попередньої динаміки витрат.

Абляційне дослідження підтвердило важливість історичних значень витрат для побудови прогнозу. Найбільше погіршення якості спостерігалося після вилучення лагових ознак, що доводить їх ключову роль у моделюванні. Водночас повна XGBoost модель показала кращий результат, ніж модель лише з лаговими ознаками, тому додаткові змінні, зокрема планові витрати та департаментальна належність, також мають практичну цінність.

Окремо було застосовано модель Prophet для аналізу агрегованого квартального часового ряду. Вона дозволила оцінити загальну тенденцію зміни витрат і побудувати прогноз наступного умовного кварталу. Водночас результати Prophet слід інтерпретувати обережно, оскільки датасет містить

обмежену кількість квартальних спостережень, що не дає змоги повноцінно оцінити довгостроковий тренд або сезонність.

У процесі дослідження також було виконано аналіз тренду, залишків і потенційних аномалій. Встановлено, що дані не містять сильних статистичних викидів, однак мають помітну квартальну нерівномірність. Найвищий рівень фактичних витрат спостерігається у другому кварталі, після чого у третьому та четвертому кварталах відбувається зниження витрат.

Порівняння підходів показало, що кожна модель має власну роль у дослідженні. Базова математична модель є простою та прозорою, але поступається за точністю. XGBoost модель забезпечує кращі результати прогнозування на рівні департаментів, однак потребує додаткової інтерпретації. Prophet є корисним для аналізу загальної динаміки та візуалізації тренду, але має обмеження через короткий часовий ряд.

Отже, поставлену мету роботи досягнуто. У результаті дослідження було проаналізовано структуру бюджетних витрат на прикладовому наборі даних, побудовано та порівняно кілька моделей прогнозування, визначено ключові ознаки, що впливають на результат, і проведено інтерпретацію роботи XGBoost моделі. Практичне значення роботи полягає у формуванні послідовного підходу до аналізу бюджетних даних, який може бути використаний для фінансового аналізу, оцінювання відхилень між плановими та фактичними витратами, підтримки управлінських рішень і підвищення прозорості моделей машинного навчання.

ПЕРЕЛІК ПОСИЛАНЬ

1. Hyndman R. J., Athanasopoulos G. Forecasting: Principles and Practice. 3rd ed. Melbourne: OTexts, 2021. URL: https://otexts.com/fpp3/ (дата звернення: 05.06.2026).
2. Box G. E. P., Jenkins G. M., Reinsel G. C., Ljung G. M. Time Series Analysis: Forecasting and Control. 5th ed. Hoboken: John Wiley & Sons, 2015. 712 p.
3. Hamilton J. D. Time Series Analysis. Princeton: Princeton University Press, 1994. 799 p.
4. Tsay R. S. Analysis of Financial Time Series. 3rd ed. Hoboken: John Wiley & Sons, 2010. 715 p.
5. Taylor S. J., Letham B. Forecasting at scale. The American Statistician. 2018. Vol. 72, No. 1. P. 37–45. DOI: 10.1080/00031305.2017.1380080.
6. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning: with Applications in R. 2nd ed. New York: Springer, 2021. 607 p.
7. Friedman J. H. Greedy function approximation: A gradient boosting machine. The Annals of Statistics. 2001. Vol. 29, No. 5. P. 1189–1232. DOI: 10.1214/aos/1013203451.
8. Chen T., Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016. P. 785–794. DOI: 10.1145/2939672.2939785.
9. Hochreiter S., Schmidhuber J. Long short-term memory. Neural Computation. 1997. Vol. 9, No. 8. P. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.
10. Hyndman R. J., Koehler A. B. Another look at measures of forecast accuracy. International Journal of Forecasting. 2006. Vol. 22, No. 4. P. 679–688. DOI: 10.1016/j.ijforecast.2006.03.001.

11. Breiman L. Random forests. Machine Learning. 2001. Vol. 45. P. 5–32. DOI: 10.1023/A:1010933404324.
12. Fisher A., Rudin C., Dominici F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research. 2019. Vol. 20, No. 177. P. 1–81.
13. Lundberg S. M., Lee S.-I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 4765–4774.
14. Prophet: forecasting at scale. URL: https://facebook.github.io/prophet/ (дата звернення: 05.06.2026).
15. Prophet Documentation. Quick Start. URL: https://facebook.github.io/prophet/docs/quick_start.html (дата звернення: 05.06.2026).
16. Prophet GitHub Repository. URL: https://github.com/facebook/prophet (дата звернення: 05.06.2026).
17. XGBoost Documentation. URL: https://xgboost.readthedocs.io/ (дата звернення: 05.06.2026).
18. XGBoost Python API Reference. URL: https://xgboost.readthedocs.io/en/latest/python/python_api.html (дата звернення: 05.06.2026).
19. SHAP Documentation. URL: https://shap.readthedocs.io/ (дата звернення: 05.06.2026).
20. SHAP GitHub Repository. URL: https://github.com/shap/shap (дата звернення: 05.06.2026).
21. Scikit-learn. Permutation feature importance. URL: https://scikit-learn.org/stable/modules/permutation_importance.html (дата звернення: 05.06.2026).

22. Scikit-learn. Permutation importance API. URL: https://scikit-learn.org/stable/modules/generated/sklearn.inspection.permutation_importance.html (дата звернення: 05.06.2026).
23. Scikit-learn. Mean absolute error. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html (дата звернення: 05.06.2026).
24. Scikit-learn. Mean squared error. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html (дата звернення: 05.06.2026).
25. Scikit-learn. Mean absolute percentage error. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_percentage_error.html (дата звернення: 05.06.2026).
26. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830.
27. McKinney W. Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference. Austin, 2010. P. 56–61.
28. Pandas Documentation. URL: https://pandas.pydata.org/docs/ (дата звернення: 05.06.2026).
29. Harris C. R., Millman K. J., van der Walt S. J., Gommers R., Virtanen P., Cournapeau D., Wieser E., Taylor J., Berg S., Smith N. J., Kern R., Picus M., Hoyer S., van Kerkwijk M. H., Brett M., Haldane A., del Río J. F., Wiebe M., Peterson P., Gérard-Marchant P., Sheppard K., Reddy T., Weckesser W., Abbasi H., Gohlke C., Oliphant T. E. Array programming with NumPy. Nature. 2020. Vol. 585. P. 357–362. DOI: 10.1038/s41586-020-2649-2.

30. NumPy Documentation. URL: https://numpy.org/doc/ (дата звернення: 05.06.2026).
31. Hunter J. D. Matplotlib: A 2D graphics environment. Computing in Science & Engineering. 2007. Vol. 9, No. 3. P. 90–95. DOI: 10.1109/MCSE.2007.55.
32. Matplotlib Documentation. URL: https://matplotlib.org/stable/index.html (дата звернення: 05.06.2026).
33. Van Rossum G., Drake F. L. Python 3 Reference Manual. Scotts Valley: CreateSpace, 2009. 242 p.
34. Python Documentation. URL: https://docs.python.org/3/ (дата звернення: 05.06.2026).
35. Kluyver T., Ragan-Kelley B., Pérez F., Granger B., Bussonnier M., Frederic J., Kelley K., Hamrick J., Grout J., Corlay S., Ivanov P., Avila D., Abdalla S., Willing C. Jupyter Notebooks – a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas. Amsterdam: IOS Press, 2016. P. 87–90. DOI: 10.3233/978-1-61499-649-1-87.
36. Google Colaboratory. URL: https://colab.research.google.com/ (дата звернення: 05.06.2026).
37. Brockwell P. J., Davis R. A. Introduction to Time Series and Forecasting. 2nd ed. New York: Springer, 2002. 434 p.
38. Chatfield C. The Analysis of Time Series: An Introduction. 6th ed. Boca Raton: Chapman & Hall/CRC, 2003. 352 p.
39. Makridakis S., Wheelwright S. C., Hyndman R. J. Forecasting: Methods and Applications. 3rd ed. New York: John Wiley & Sons, 1998. 642 p.
40. Makridakis S., Spiliotis E., Assimakopoulos V. The M4 Competition: Results, findings, conclusion and way forward. International Journal of Forecasting. 2018. Vol. 34, No. 4. P. 802–808. DOI: 10.1016/j.ijforecast.2018.06.001.

41. Makridakis S., Spiliotis E., Assimakopoulos V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. PLoS ONE. 2018. Vol. 13, No. 3. e0194889. DOI: 10.1371/journal.pone.0194889.
42. Shapley L. S. A value for n-person games. Contributions to the Theory of Games / ed. by H. W. Kuhn, A. W. Tucker. Princeton: Princeton University Press, 1953. Vol. 2. P. 307–317.
43. Lundberg S. M., Erion G., Chen H., DeGrave A., Prutkin J. M., Nair B., Katz R., Himmelfarb J., Bansal N., Lee S.-I. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence. 2020. Vol. 2. P. 56–67. DOI: 10.1038/s42256-019-0138-9.
44. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2nd ed. 2022. URL: https://christophm.github.io/interpretable-ml-book/ (дата звернення: 05.06.2026).
45. Friedman J. H. Stochastic gradient boosting. Computational Statistics & Data Analysis. 2002. Vol. 38, No. 4. P. 367–378. DOI: 10.1016/S0167-9473(01)00065-2.
46. Breiman L. Bagging predictors. Machine Learning. 1996. Vol. 24. P. 123–140. DOI: 10.1007/BF00058655.
47. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer, 2009. 745 p.
48. Government Finance Statistics Manual 2014. Washington, D.C.: International Monetary Fund, 2014. URL: https://www.imf.org/external/pubs/ft/gfs/manual/2014/gfsfinal.pdf (дата звернення: 05.06.2026).
49. International Monetary Fund. Government Finance Statistics Manual. URL: https://www.imf.org/en/data/statistics/government-finance-statistics-manual (дата звернення: 05.06.2026).

50. Budgeting and Public Expenditures in OECD Countries 2019. Paris: OECD Publishing, 2019. URL: https://www.oecd.org/en/publications/budgeting-and-public-expenditures-in-oecd-countries-2018_9789264307957-en.html (дата звернення: 05.06.2026).
51. OECD Journal on Budgeting. Paris: OECD Publishing, 2019. URL: https://www.oecd.org/en/publications/oecd-journal-on-budgeting-volume-2019-issue-1_7d70199c-en.html (дата звернення: 05.06.2026).
52. Kaggle. Budget Spending Datasets. URL: https://www.kaggle.com/datasets/gilieannsoquila/budget-spending-datasets (дата звернення: 05.06.2026).
53. Kaggle. Budget Spending Datasets: Data page. URL: https://www.kaggle.com/datasets/gilieannsoquila/budget-spending-datasets/data (дата звернення: 05.06.2026).
54. Kaggle. Federal Budget Shares 1941–2020. URL: https://www.kaggle.com/datasets/joebeachcapital/federal-budget-shares-1941-2020 (дата звернення: 05.06.2026).
55. Kaggle. Financial Dataset [Expenses]: Budget vs Actual. URL: https://www.kaggle.com/datasets/bukolafatunde/financial-dataset-expenses-budget-vs-actual (дата звернення: 05.06.2026).
56. World Bank. Expense (% of GDP). URL: https://data.worldbank.org/indicator/GC.XPN.TOTL.GD.ZS (дата звернення: 05.06.2026).
57. Federal Reserve Economic Data. Federal Government: Current Expenditures. URL: https://fred.stlouisfed.org/series/FGEXPND (дата звернення: 05.06.2026).
58. OECD. General Government Spending. URL: https://data.oecd.org/gga/general-government-spending.htm (дата звернення: 05.06.2026).

59. M4 Forecasting Competition Dataset. URL: https://github.com/Mcompetitions/M4-methods/tree/master/Dataset (дата звернення: 05.06.2026).
60. Kaggle. M5 Forecasting – Accuracy. URL: https://www.kaggle.com/competitions/m5-forecasting-accuracy (дата звернення: 05.06.2026).
61. Зайченко Ю.П. Теорія прийняття рішень. Київ: Видавництво «КПІ», 2014. 412 с.
62. Бідюк П. І., Романенко В. Д., Тимощук О. Л. Аналіз часових рядів: навч. посіб. Київ, НТУУ "КПІ", 2013. 600 с.
63. Cho K., van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014. P. 1724–1734. DOI: 10.3115/v1/D14-1179.

ДОДАТОК А ЛІСТИНГ ПРОГРАМИ

```
{
  "nbformat": 4,
  "nbformat_minor": 5,
  "metadata": {
    "colab": {
      "provenance": []
    },
    "kernelspec": {
      "name": "python3",
      "display_name": "Python 3"
    },
    "language_info": {
      "name": "python"
    }
  },
  "cells": [
    {
      "cell_type": "markdown",
      "metadata": {},
      "source": [
        "# Аналіз і прогнозування бюджетних витрат\n",
        "\n",

      ]
    },
    {
      "cell_type": "code",
      "metadata": {},
      "execution_count": null,
      "outputs": [],
      "source": [
        "!pip install xgboost shap prophet openpyxl -q\n",
        "\n",
        "import pandas as pd\n",
        "import numpy as np\n",
        "import matplotlib.pyplot as plt\n",
        "\n",
        "from sklearn.metrics import mean_absolute_error,
mean_squared_error\n",
        "from sklearn.inspection import permutation_importance\n",
        "from xgboost import XGBRegressor\n",
        "from prophet import Prophet\n",
        "import shap\n",
        "from google.colab import files\n",
        "\n",
        "import warnings\n",
        "warnings.filterwarnings(\"ignore\")\n",
        "plt.rcParams[\"figure.figsize\"] = (12, 7)"
      ]
    },
    {
      "cell_type": "markdown",
      "metadata": {},
      "source": [
        "## 1. Завантаження та первинна перевірка датасету"
      ]
    },
    {
      "cell_type": "code",
      "metadata": {},
```

```
              "execution_count": null,
              "outputs": [],
              "source": [
                "uploaded = files.upload()\n",
                "file_name = list(uploaded.keys())[0]\n",
                "df = pd.read_csv(file_name)\n",
                "\n",
                "print(\"Розмір датасету:\", df.shape)\n",
                "print(\"\\nНазви стовпців:\", df.columns.tolist())\n",
                "print(\"\\nТипи даних:\")\n",
                "print(df.dtypes)\n",
                "print(\"\\nКількість пропущених значень:\")\n",
                "print(df.isna().sum())\n",
                "print(\"\\nКількість дублікатів:\",
df.duplicated().sum())\n",
                "print(\"\\nОписова статистика:\")\n",
                "display(df.describe())\n",
                "display(df.head())"
              ]
            },
            {
              "cell_type": "markdown",
              "metadata": {},
              "source": [
                "## 2. Перевірка Variance і формування кварталів"
              ]
            },
            {
              "cell_type": "code",
              "metadata": {},
              "execution_count": null,
              "outputs": [],
              "source": [
                "df[\"Variance_check\"] = df[\"Actual Spending\"] -
df[\"Budgeted Amount\"]\n",
                "variance_errors = (df[\"Variance\"] !=
df[\"Variance_check\"]).sum()\n",
                "print(\"Кількість помилок у Variance:\", variance_errors)\n",
                "df = df.drop(columns=[\"Variance_check\"])\n",
                "\n",
                "month_to_quarter = {\n",
                "    \"Jan\": \"Q1\", \"Feb\": \"Q1\", \"Mar\": \"Q1\",\n",
                "    \"Apr\": \"Q2\", \"May\": \"Q2\", \"Jun\": \"Q2\",\n",
                "    \"Jul\": \"Q3\", \"Aug\": \"Q3\", \"Sep\": \"Q3\",\n",
                "    \"Oct\": \"Q4\", \"Nov\": \"Q4\", \"Dec\": \"Q4\"\n",
                "}\n",
                "quarter_order = [\"Q1\", \"Q2\", \"Q3\", \"Q4\"]\n",
                "df[\"Quarter\"] = df[\"Month\"].map(month_to_quarter)\n",
                "df[\"quarter_num\"] = df[\"Quarter\"].map({\"Q1\": 1, \"Q2\":
2, \"Q3\": 3, \"Q4\": 4})\n",
                "display(df.head())"
              ]
            },
            {
              "cell_type": "markdown",
              "metadata": {},
              "source": [
                "## 3. Агрегований аналіз за кварталами"
              ]
            },
            {
              "cell_type": "code",
              "metadata": {},
              "execution_count": null,
```

```
            "outputs": [],
            "source": [
              "quarterly = (\n",
              "    df.groupby([\"Quarter\", \"quarter_num\"],
as_index=False)\n",
              "    .agg(\n",
              "        records=(\"Actual Spending\", \"size\"),\n",
              "        budgeted=(\"Budgeted Amount\", \"sum\"),\n",
              "        actual=(\"Actual Spending\", \"sum\"),\n",
              "        variance=(\"Variance\", \"sum\")\n",
              "    )\n",
              ")\n",
              "quarterly[\"budget_execution_pct\"] = quarterly[\"actual\"] /
quarterly[\"budgeted\"] * 100\n",
              "quarterly =
quarterly.sort_values(\"quarter_num\").reset_index(drop=True)\n",
              "display(quarterly)\n",
              "\n",
              "plt.figure(figsize=(12, 7))\n",
              "plt.plot(quarterly[\"Quarter\"], quarterly[\"actual\"],
marker=\"o\", linewidth=2, label=\"Фактичні витрати\")\n",
              "plt.title(\"Динаміка фактичних витрат за кварталами\")\n",
              "plt.xlabel(\"Квартал\")\n",
              "plt.ylabel(\"Сума фактичних витрат\")\n",
              "plt.grid(True)\n",
              "plt.legend()\n",
              "plt.tight_layout()\n",
              "plt.savefig(\"quarterly_actual_spending.png\", dpi=200,
bbox_inches=\"tight\")\n",
              "plt.show()"
            ]
          },
          {
            "cell_type": "markdown",
            "metadata": {},
            "source": [
              "## 4. Підготовка даних на рівні «департамент — квартал»"
            ]
          },
          {
            "cell_type": "code",
            "metadata": {},
            "execution_count": null,
            "outputs": [],
            "source": [
              "dept_quarter = (\n",
              "    df.groupby([\"Department\", \"Quarter\",
\"quarter_num\"], as_index=False)\n",
              "    .agg(\n",
              "        records=(\"Actual Spending\", \"size\"),\n",
              "        budgeted=(\"Budgeted Amount\", \"sum\"),\n",
              "        actual=(\"Actual Spending\", \"sum\"),\n",
              "        variance=(\"Variance\", \"sum\")\n",
              "    )\n",
              ")\n",
              "dept_quarter[\"budget_execution_pct\"] =
dept_quarter[\"actual\"] / dept_quarter[\"budgeted\"] * 100\n",
              "dept_quarter = dept_quarter.sort_values([\"Department\",
\"quarter_num\"]).reset_index(drop=True)\n",
              "display(dept_quarter.head(12))"
            ]
          },
          {
            "cell_type": "markdown",
```

```
            "metadata": {},
            "source": [
              "## 5. Формування ознак для прогнозу Q4 за Q1–Q3"
            ]
          },
          {
            "cell_type": "code",
            "metadata": {},
            "execution_count": null,
            "outputs": [],
            "source": [
              "rows = []\n",
              "for dept in dept_quarter[\"Department\"].unique():\n",
              "    temp = dept_quarter[dept_quarter[\"Department\"] ==
dept].sort_values(\"quarter_num\")\n",
              "    q1 = temp[temp[\"Quarter\"] == \"Q1\"].iloc[0]\n",
              "    q2 = temp[temp[\"Quarter\"] == \"Q2\"].iloc[0]\n",
              "    q3 = temp[temp[\"Quarter\"] == \"Q3\"].iloc[0]\n",
              "    q4 = temp[temp[\"Quarter\"] == \"Q4\"].iloc[0]\n",
              "    rows.append({\n",
              "        \"Department\": dept,\n",
              "        \"lag3_actual\": q1[\"actual\"],\n",
              "        \"lag2_actual\": q2[\"actual\"],\n",
              "        \"lag1_actual\": q3[\"actual\"],\n",
              "        \"rolling_3q_actual\": np.mean([q1[\"actual\"],
q2[\"actual\"], q3[\"actual\"]]),\n",
              "        \"diff_lag1_lag2\": q3[\"actual\"] -
q2[\"actual\"],\n",
              "        \"diff_lag2_lag3\": q2[\"actual\"] -
q1[\"actual\"],\n",
              "        \"budgeted_target\": q4[\"budgeted\"],\n",
              "        \"records_target\": q4[\"records\"],\n",
              "        \"target_actual\": q4[\"actual\"]\n",
              "    })\n",
              "model_data = pd.DataFrame(rows)\n",
              "display(model_data)"
            ]
          },
          {
            "cell_type": "markdown",
            "metadata": {},
            "source": [
              "## 6. Базова математична модель"
            ]
          },
          {
            "cell_type": "code",
            "metadata": {},
            "execution_count": null,
            "outputs": [],
            "source": [
              "model_data[\"baseline_forecast\"] =
model_data[\"rolling_3q_actual\"]\n",
              "model_data[\"baseline_error\"] =
model_data[\"target_actual\"] - model_data[\"baseline_forecast\"]\n",
              "model_data[\"baseline_abs_error\"] =
abs(model_data[\"baseline_error\"])\n",
              "model_data[\"baseline_ape\"] =
model_data[\"baseline_abs_error\"] / model_data[\"target_actual\"] * 100\n",
              "\n",
              "baseline_mae =
mean_absolute_error(model_data[\"target_actual\"],
model_data[\"baseline_forecast\"])\n",
```

```
                "baseline_rmse =
np.sqrt(mean_squared_error(model_data[\"target_actual\"],
model_data[\"baseline_forecast\"]))\n",
                "baseline_mape = model_data[\"baseline_ape\"].mean()\n",
                "\n",
                "display(model_data[[\"Department\", \"target_actual\",
\"baseline_forecast\", \"baseline_abs_error\", \"baseline_ape\"]])\n",
                "print(\"MAE:\", baseline_mae)\n",
                "print(\"RMSE:\", baseline_rmse)\n",
                "print(\"MAPE:\", baseline_mape)"
              ]
            },
            {
              "cell_type": "markdown",
              "metadata": {},
              "source": [
                "## 7. Рисунок 3.1 — базовий прогноз Q4"
              ]
            },
            {
              "cell_type": "code",
              "metadata": {},
              "execution_count": null,
              "outputs": [],
              "source": [
                "q_values = quarterly[\"actual\"].tolist()\n",
                "q4_forecast = np.mean(q_values[:3])\n",
                "plt.figure(figsize=(12, 7))\n",
                "plt.plot([\"Q1\", \"Q2\", \"Q3\", \"Q4\"], q_values,
marker=\"o\", linewidth=2, label=\"Фактичні витрати\")\n",
                "plt.scatter([\"Q4\"], [q4_forecast], s=130, marker=\"D\",
label=\"Прогноз базової моделі для Q4\")\n",
                "plt.plot([\"Q3\", \"Q4\"], [q_values[2], q4_forecast],
linestyle=\"--\", linewidth=2, label=\"Лінія прогнозу\")\n",
                "plt.title(\"Рисунок 3.1 – Порівняння фактичних витрат і
прогнозу базової моделі за кварталами\")\n",
                "plt.xlabel(\"Квартал\")\n",
                "plt.ylabel(\"Сума фактичних витрат\")\n",
                "plt.grid(True)\n",
                "plt.legend()\n",
                "plt.tight_layout()\n",
                "plt.savefig(\"figure_3_1_baseline_model.png\", dpi=200,
bbox_inches=\"tight\")\n",
                "plt.show()"
              ]
            },
            {
              "cell_type": "markdown",
              "metadata": {},
              "source": [
                "## 8. Black-box модель XGBoost"
              ]
            },
            {
              "cell_type": "code",
              "metadata": {},
              "execution_count": null,
              "outputs": [],
              "source": [
                "X = model_data[[\n",
                "    \"lag1_actual\", \"lag2_actual\", \"lag3_actual\",
\"rolling_3q_actual\",\n",
                "    \"diff_lag1_lag2\", \"diff_lag2_lag3\",
\"budgeted_target\", \"records_target\"\n",
```

```
                "]]\n",
                "department_dummies =
pd.get_dummies(model_data[\"Department\"], prefix=\"Department\")\n",
                "X = pd.concat([X, department_dummies], axis=1)\n",
                "y = model_data[\"target_actual\"]\n",
                "\n",
                "xgb_model = XGBRegressor(\n",
                "    n_estimators=100,\n",
                "    max_depth=2,\n",
                "    learning_rate=0.05,\n",
                "    subsample=1.0,\n",
                "    colsample_bytree=1.0,\n",
                "    random_state=42\n",
                ")\n",
                "xgb_model.fit(X, y)\n",
                "model_data[\"blackbox_forecast\"] = xgb_model.predict(X)\n",
                "model_data[\"blackbox_error\"] =
model_data[\"target_actual\"] - model_data[\"blackbox_forecast\"]\n",
                "model_data[\"blackbox_abs_error\"] =
abs(model_data[\"blackbox_error\"])\n",
                "model_data[\"blackbox_ape\"] =
model_data[\"blackbox_abs_error\"] / model_data[\"target_actual\"] * 100\n",
                "\n",
                "blackbox_mae =
mean_absolute_error(model_data[\"target_actual\"],
model_data[\"blackbox_forecast\"])\n",
                "blackbox_rmse =
np.sqrt(mean_squared_error(model_data[\"target_actual\"],
model_data[\"blackbox_forecast\"]))\n",
                "blackbox_mape = model_data[\"blackbox_ape\"].mean()\n",
                "\n",
                "display(model_data[[\"Department\", \"target_actual\",
\"blackbox_forecast\", \"blackbox_abs_error\", \"blackbox_ape\"]])\n",
                "print(\"MAE:\", blackbox_mae)\n",
                "print(\"RMSE:\", blackbox_rmse)\n",
                "print(\"MAPE:\", blackbox_mape)"
              ]
            },
            {
              "cell_type": "markdown",
              "metadata": {},
              "source": [
                "## 9. Порівняння математичної та black-box моделі"
              ]
            },
            {
              "cell_type": "code",
              "metadata": {},
              "execution_count": null,
              "outputs": [],
              "source": [
                "comparison_metrics = pd.DataFrame({\n",
                "    \"Модель\": [\"Математична модель\", \"Black-box
модель\"],\n",
                "    \"MAE\": [baseline_mae, blackbox_mae],\n",
                "    \"RMSE\": [baseline_rmse, blackbox_rmse],\n",
                "    \"MAPE\": [baseline_mape, blackbox_mape]\n",
                "})\n",
                "display(comparison_metrics)\n",
                "\n",
                "metrics = [\"MAE\", \"RMSE\"]\n",
                "math_values = [baseline_mae, baseline_rmse]\n",
                "bb_values = [blackbox_mae, blackbox_rmse]\n",
                "x_pos = np.arange(len(metrics))\n",
```

```
                "width = 0.35\n",
                "plt.figure(figsize=(10, 7))\n",
                "plt.bar(x_pos - width/2, math_values, width,
label=\"Математична модель\")\n",
                "plt.bar(x_pos + width/2, bb_values, width, label=\"Black-box
модель\")\n",
                "plt.xticks(x_pos, metrics)\n",
                "plt.xlabel(\"Метрика\")\n",
                "plt.ylabel(\"Значення помилки\")\n",
                "plt.title(\"Рисунок 7.1 – Порівняння абсолютних метрик
математичної та black-box моделі\")\n",
                "plt.legend()\n",
                "plt.grid(True, axis=\"y\")\n",
                "plt.tight_layout()\n",
                "plt.savefig(\"figure_7_1_absolute_metrics.png\", dpi=200,
bbox_inches=\"tight\")\n",
                "plt.show()\n",
                "\n",
                "plt.figure(figsize=(8, 7))\n",
                "plt.bar([\"Математична модель\", \"Black-box модель\"],
[baseline_mape, blackbox_mape])\n",
                "plt.xlabel(\"Модель\")\n",
                "plt.ylabel(\"MAPE, %\")\n",
                "plt.title(\"Рисунок 7.2 – Порівняння MAPE математичної та
black-box моделі\")\n",
                "plt.grid(True, axis=\"y\")\n",
                "plt.tight_layout()\n",
                "plt.savefig(\"figure_7_2_mape_comparison.png\", dpi=200,
bbox_inches=\"tight\")\n",
                "plt.show()\n",
                "\n",
                "improvement_df = pd.DataFrame({\n",
                "    \"Метрика\": [\"MAE\", \"RMSE\", \"MAPE\"],\n",
                "    \"Покращення, %\": [\n",
                "        (baseline_mae - blackbox_mae) / baseline_mae *
100,\n",
                "        (baseline_rmse - blackbox_rmse) / baseline_rmse *
100,\n",
                "        (baseline_mape - blackbox_mape) / baseline_mape *
100\n",
                "    ]\n",
                "})\n",
                "display(improvement_df)\n",
                "plt.figure(figsize=(9, 7))\n",
                "plt.bar(improvement_df[\"Метрика\"],
improvement_df[\"Покращення, %\"])\n",
                "plt.xlabel(\"Метрика\")\n",
                "plt.ylabel(\"Покращення, %\")\n",
                "plt.title(\"Рисунок 7.3 – Покращення black-box моделі
порівняно з математичною\")\n",
                "plt.grid(True, axis=\"y\")\n",
                "plt.tight_layout()\n",
                "plt.savefig(\"figure_7_3_blackbox_improvement.png\", dpi=200,
bbox_inches=\"tight\")\n",
                "plt.show()"
              ]
            },
            {
              "cell_type": "markdown",
              "metadata": {},
              "source": [
                "## 10. Permutation Importance"
              ]
            },
```

```
            {
              "cell_type": "code",
              "metadata": {},
              "execution_count": null,
              "outputs": [],
              "source": [
                "perm_result = permutation_importance(\n",
                "    xgb_model, X, y, n_repeats=30, random_state=42,
scoring=\"neg_mean_absolute_error\"\n",
                ")\n",
                "perm_importance = pd.DataFrame({\n",
                "    \"feature\": X.columns,\n",
                "    \"importance\": perm_result.importances_mean\n",
                "}).sort_values(\"importance\", ascending=False)\n",
                "display(perm_importance)\n",
                "\n",
                "plot_perm = perm_importance.sort_values(\"importance\",
ascending=True)\n",
                "plt.figure(figsize=(12, 7))\n",
                "plt.barh(plot_perm[\"feature\"],
plot_perm[\"importance\"])\n",
                "plt.title(\"Рисунок 5.1 – Важливість ознак за методом
Permutation Importance\")\n",
                "plt.xlabel(\"Зростання похибки після перемішування
ознаки\")\n",
                "plt.ylabel(\"Ознака\")\n",
                "plt.grid(True, axis=\"x\")\n",
                "plt.tight_layout()\n",
                "plt.savefig(\"figure_5_1_permutation_importance.png\",
dpi=200, bbox_inches=\"tight\")\n",
                "plt.show()"
              ]
            },
            {
              "cell_type": "markdown",
              "metadata": {},
              "source": [
                "## 11. SHAP-аналіз"
              ]
            },
            {
              "cell_type": "code",
              "metadata": {},
              "execution_count": null,
              "outputs": [],
              "source": [
                "explainer = shap.TreeExplainer(xgb_model)\n",
                "shap_values = explainer.shap_values(X)\n",
                "shap_df = pd.DataFrame(shap_values, columns=X.columns)\n",
                "mean_abs_shap =
shap_df.abs().mean().sort_values(ascending=False)\n",
                "display(mean_abs_shap)\n",
                "\n",
                "plot_shap = mean_abs_shap.sort_values(ascending=True)\n",
                "plt.figure(figsize=(12, 7))\n",
                "plt.barh(plot_shap.index, plot_shap.values)\n",
                "plt.title(\"Рисунок 8.1 – Глобальна важливість ознак за
SHAP-аналізом\")\n",
                "plt.xlabel(\"Середнє абсолютне SHAP-значення\")\n",
                "plt.ylabel(\"Ознака\")\n",
                "plt.grid(True, axis=\"x\")\n",
                "plt.tight_layout()\n",
                "plt.savefig(\"figure_8_1_shap_global_importance.png\",
dpi=200, bbox_inches=\"tight\")\n",
```

```
                "plt.show()\n",
                "\n",
                "shap.summary_plot(shap_values, X, feature_names=X.columns,
show=True)\n",
                "\n",
                "sample_index = 0\n",
                "shap.plots.waterfall(\n",
                "    shap.Explanation(\n",
                "        values=shap_values[sample_index],\n",
                "        base_values=explainer.expected_value,\n",
                "        data=X.iloc[sample_index],\n",
                "        feature_names=X.columns\n",
                "    )\n",
                ")"
              ]
            },
            {
              "cell_type": "markdown",
              "metadata": {},
              "source": [
                "## 12. Абляційне дослідження ознак"
              ]
            },
            {
              "cell_type": "code",
              "metadata": {},
              "execution_count": null,
              "outputs": [],
              "source": [
                "def evaluate_model(feature_list):\n",
                "    model = XGBRegressor(n_estimators=100, max_depth=2,
learning_rate=0.05, random_state=42)\n",
                "    model.fit(X[feature_list], y)\n",
                "    preds = model.predict(X[feature_list])\n",
                "    mae = mean_absolute_error(y, preds)\n",
                "    rmse = np.sqrt(mean_squared_error(y, preds))\n",
                "    mape = np.mean(np.abs((y - preds) / y)) * 100\n",
                "    return mae, rmse, mape\n",
                "\n",
                "all_features = list(X.columns)\n",
                "lag_features = [\"lag1_actual\", \"lag2_actual\",
\"lag3_actual\"]\n",
                "rolling_features = [\"rolling_3q_actual\"]\n",
                "diff_features = [\"diff_lag1_lag2\", \"diff_lag2_lag3\"]\n",
                "budget_features = [\"budgeted_target\"]\n",
                "records_features = [\"records_target\"]\n",
                "department_features = [col for col in X.columns if
col.startswith(\"Department_\")]\n",
                "\n",
                "experiments = {\n",
                "    \"Повна black-box модель\": all_features,\n",
                "    \"Без лагових ознак\": [f for f in all_features if f not
in lag_features],\n",
                "    \"Без ковзного середнього\": [f for f in all_features if
f not in rolling_features],\n",
                "    \"Без різниць між лагами\": [f for f in all_features if f
not in diff_features],\n",
                "    \"Без бюджетного плану\": [f for f in all_features if f
not in budget_features],\n",
                "    \"Без департаменту\": [f for f in all_features if f not
in department_features],\n",
                "    \"Без records_target\": [f for f in all_features if f not
in records_features],\n",
                "    \"Тільки лагові ознаки\": lag_features,\n",
```

```
                "    \"Тільки бюджетні й категоріальні ознаки\": 
budget_features + records_features + department_features\n",
                "}\n",
                "\n",
                "ablation_results = []\n",
                "for name, features_list in experiments.items():\n",
                "    mae, rmse, mape = evaluate_model(features_list)\n",
                "    ablation_results.append({\"Варіант моделі\": name, 
\"MAE\": mae, \"RMSE\": rmse, \"MAPE\": mape})\n",
                "ablation_df = pd.DataFrame(ablation_results)\n",
                "display(ablation_df)"
              ]
            },
            {
              "cell_type": "markdown",
              "metadata": {},
              "source": [
                "## 13. Prophet — підготовка, навчання та прогноз"
              ]
            },
            {
              "cell_type": "code",
              "metadata": {},
              "execution_count": null,
              "outputs": [],
              "source": [
                "prophet_df = quarterly[[\"Quarter\", \"actual\"]].copy()\n",
                "prophet_df[\"ds\"] = pd.to_datetime([\"2020-01-01\", 
\"2020-04-01\", \"2020-07-01\", \"2020-10-01\"])\n",
                "prophet_df[\"y\"] = prophet_df[\"actual\"]\n",
                "prophet_df = prophet_df[[\"ds\", \"y\"]]\n",
                "display(prophet_df)\n",
                "\n",
                "prophet_model = Prophet(yearly_seasonality=False, 
weekly_seasonality=False, daily_seasonality=False)\n",
                "prophet_model.fit(prophet_df)\n",
                "future = prophet_model.make_future_dataframe(periods=1, 
freq=\"Q\")\n",
                "forecast = prophet_model.predict(future)\n",
                "display(forecast[[\"ds\", \"yhat\", \"yhat_lower\", 
\"yhat_upper\"]])\n",
                "\n",
                "fig = prophet_model.plot(forecast)\n",
                "plt.title(\"Рисунок 10.1 – Прогноз бюджетних витрат методом 
Prophet\")\n",
                "plt.xlabel(\"Період\")\n",
                "plt.ylabel(\"Сума фактичних витрат\")\n",
                "plt.grid(True)\n",
                "plt.show()\n",
                "\n",
                "fig = prophet_model.plot_components(forecast)\n",
                "plt.show()"
              ]
            },
            {
              "cell_type": "markdown",
              "metadata": {},
              "source": [
                "## 14. Ручний графік Prophet у квартальному вигляді"
              ]
            },
            {
              "cell_type": "code",
              "metadata": {},
```

```
            "execution_count": null,
            "outputs": [],
            "source": [
              "plt.figure(figsize=(12, 7))\n",
              "plt.plot([\"Q1\", \"Q2\", \"Q3\", \"Q4\"], prophet_df[\"y\"],
marker=\"o\", linewidth=2, label=\"Фактичні витрати\")\n",
              "plt.plot([\"Q1\", \"Q2\", \"Q3\", \"Q4\", \"Наступний
квартал\"], forecast[\"yhat\"], marker=\"o\", linestyle=\"--\", linewidth=2,
label=\"Прогноз Prophet\")\n",
              "plt.fill_between([\"Q1\", \"Q2\", \"Q3\", \"Q4\", \"Наступний
квартал\"], forecast[\"yhat_lower\"], forecast[\"yhat_upper\"], alpha=0.2,
label=\"Довірчий інтервал\")\n",
              "plt.title(\"Рисунок 10.1 – Прогноз бюджетних витрат методом
Prophet\")\n",
              "plt.xlabel(\"Період\")\n",
              "plt.ylabel(\"Сума фактичних витрат\")\n",
              "plt.grid(True)\n",
              "plt.legend()\n",
              "plt.tight_layout()\n",
              "plt.savefig(\"figure_10_1_prophet_forecast.png\", dpi=200,
bbox_inches=\"tight\")\n",
              "plt.show()"
            ]
          },
          {
            "cell_type": "markdown",
            "metadata": {},
            "source": [
              "## 15. Порівняння прогнозу з фактичними даними"
            ]
          },
          {
            "cell_type": "code",
            "metadata": {},
            "execution_count": null,
            "outputs": [],
            "source": [
              "actual_vs_pred = model_data[[\"Department\",
\"target_actual\", \"blackbox_forecast\", \"blackbox_abs_error\",
\"blackbox_ape\"]].copy()\n",
              "actual_vs_pred.columns = [\"Департамент\", \"Фактичні витрати
Q4\", \"Прогноз black-box моделі\", \"Абсолютна помилка\", \"Відносна
помилка, %\"]\n",
              "display(actual_vs_pred)\n",
              "\n",
              "departments = model_data[\"Department\"]\n",
              "actual_q4 = model_data[\"target_actual\"]\n",
              "predicted_q4 = model_data[\"blackbox_forecast\"]\n",
              "x_pos = np.arange(len(departments))\n",
              "width = 0.35\n",
              "plt.figure(figsize=(12, 7))\n",
              "plt.bar(x_pos - width/2, actual_q4, width, label=\"Фактичні
витрати Q4\")\n",
              "plt.bar(x_pos + width/2, predicted_q4, width, label=\"Прогноз
black-box моделі\")\n",
              "plt.xticks(x_pos, departments)\n",
              "plt.xlabel(\"Департамент\")\n",
              "plt.ylabel(\"Сума витрат\")\n",
              "plt.title(\"Рисунок 11.1 – Порівняння фактичних і
прогнозованих значень black-box моделі за департаментами\")\n",
              "plt.legend()\n",
              "plt.grid(True, axis=\"y\")\n",
              "plt.tight_layout()\n",
```

```
                "plt.savefig(\"figure_11_1_actual_vs_blackbox.png\", dpi=200,
bbox_inches=\"tight\")\n",
                "plt.show()"
              ]
            },
            {
              "cell_type": "markdown",
              "metadata": {},
              "source": [
                "## 16. Аналіз тренду та залишків"
              ]
            },
            {
              "cell_type": "code",
              "metadata": {},
              "execution_count": null,
              "outputs": [],
              "source": [
                "quarters = quarterly[\"Quarter\"].tolist()\n",
                "actual_spending = quarterly[\"actual\"].tolist()\n",
                "x = np.arange(len(quarters))\n",
                "trend_coeffs = np.polyfit(x, actual_spending, 1)\n",
                "trend_line = np.polyval(trend_coeffs, x)\n",
                "plt.figure(figsize=(12, 7))\n",
                "plt.plot(quarters, actual_spending, marker=\"o\",
linewidth=2, label=\"Фактичні витрати\")\n",
                "plt.plot(quarters, trend_line, linestyle=\"--\", linewidth=2,
label=\"Лінійний тренд\")\n",
                "plt.title(\"Рисунок 12.1 – Динаміка фактичних витрат і тренд
за кварталами\")\n",
                "plt.xlabel(\"Квартал\")\n",
                "plt.ylabel(\"Сума фактичних витрат\")\n",
                "plt.grid(True)\n",
                "plt.legend()\n",
                "plt.tight_layout()\n",
                "plt.savefig(\"figure_12_1_actual_spending_trend.png\",
dpi=200, bbox_inches=\"tight\")\n",
                "plt.show()\n",
                "\n",
                "model_data[\"residual\"] = model_data[\"target_actual\"] -
model_data[\"blackbox_forecast\"]\n",
                "display(model_data[[\"Department\", \"target_actual\",
\"blackbox_forecast\", \"residual\"]])\n",
                "plt.figure(figsize=(12, 7))\n",
                "plt.bar(model_data[\"Department\"],
model_data[\"residual\"])\n",
                "plt.axhline(0, linestyle=\"--\", linewidth=1)\n",
                "plt.title(\"Рисунок 12.2 – Залишки black-box моделі за
департаментами\")\n",
                "plt.xlabel(\"Департамент\")\n",
                "plt.ylabel(\"Залишок: фактичне значення мінус прогноз\")\n",
                "plt.grid(True, axis=\"y\")\n",
                "plt.tight_layout()\n",
                "plt.savefig(\"figure_12_2_blackbox_residuals.png\", dpi=200,
bbox_inches=\"tight\")\n",
                "plt.show()"
              ]
            },
            {
              "cell_type": "markdown",
              "metadata": {},
              "source": [
                "## 17. Аналіз аномалій і залишків Prophet"
              ]
```

```
            },
            {
              "cell_type": "code",
              "metadata": {},
              "execution_count": null,
              "outputs": [],
              "source": [
                "actual = quarterly[\"actual\"].values\n",
                "quarters = quarterly[\"Quarter\"].values\n",
                "mean_value = actual.mean()\n",
                "std_value = actual.std(ddof=0)\n",
                "z_scores = (actual - mean_value) / std_value\n",
                "anomaly_df = pd.DataFrame({\"Квартал\": quarters, \"Фактичні
витрати\": actual, \"Відхилення від середнього\": actual - mean_value,
\"z-score\": z_scores})\n",
                "display(anomaly_df)\n",
                "\n",
                "plt.figure(figsize=(12, 7))\n",
                "plt.plot(quarters, actual, marker=\"o\", linewidth=2,
label=\"Фактичні витрати\")\n",
                "plt.axhline(mean_value, linestyle=\"--\", linewidth=2,
label=f\"Середній рівень: {mean_value:.2f}\")\n",
                "max_idx = int(np.argmax(actual))\n",
                "min_idx = int(np.argmin(actual))\n",
                "plt.scatter([quarters[max_idx]], [actual[max_idx]], s=140,
marker=\"D\", label=\"Локальний максимум\")\n",
                "plt.scatter([quarters[min_idx]], [actual[min_idx]], s=140,
marker=\"s\", label=\"Локальний мінімум\")\n",
                "plt.title(\"Рисунок 13.1 – Виявлення потенційних аномалій у
квартальних витратах\")\n",
                "plt.xlabel(\"Квартал\")\n",
                "plt.ylabel(\"Сума фактичних витрат\")\n",
                "plt.grid(True)\n",
                "plt.legend()\n",
                "plt.tight_layout()\n",
                "plt.savefig(\"figure_13_1_quarterly_anomalies.png\", dpi=200,
bbox_inches=\"tight\")\n",
                "plt.show()\n",
                "\n",
                "prophet_fitted = forecast.iloc[:4].copy()\n",
                "prophet_fitted[\"actual\"] = prophet_df[\"y\"].values\n",
                "prophet_fitted[\"residual\"] = prophet_fitted[\"actual\"] -
prophet_fitted[\"yhat\"]\n",
                "display(prophet_fitted[[\"ds\", \"actual\", \"yhat\",
\"residual\"]])\n",
                "plt.figure(figsize=(12, 7))\n",
                "plt.bar([\"Q1\", \"Q2\", \"Q3\", \"Q4\"],
prophet_fitted[\"residual\"])\n",
                "plt.axhline(0, linestyle=\"--\", linewidth=1)\n",
                "plt.title(\"Рисунок 13.2 – Залишки моделі Prophet за
кварталами\")\n",
                "plt.xlabel(\"Квартал\")\n",
                "plt.ylabel(\"Залишок: фактичне значення мінус прогноз\")\n",
                "plt.grid(True, axis=\"y\")\n",
                "plt.tight_layout()\n",
                "plt.savefig(\"figure_13_2_prophet_residuals.png\", dpi=200,
bbox_inches=\"tight\")\n",
                "plt.show()"
              ]
            },
            {
              "cell_type": "markdown",
              "metadata": {},
              "source": [
```

```
            "## 18. Порівняльний аналіз моделей"
          ]
        },
        {
          "cell_type": "code",
          "metadata": {},
          "execution_count": null,
          "outputs": [],
          "source": [
            "prophet_mae = mean_absolute_error(prophet_fitted[\"actual\"],
prophet_fitted[\"yhat\"])\n",
            "prophet_rmse =
np.sqrt(mean_squared_error(prophet_fitted[\"actual\"],
prophet_fitted[\"yhat\"]))\n",
            "prophet_mape = np.mean(np.abs((prophet_fitted[\"actual\"] -
prophet_fitted[\"yhat\"]) / prophet_fitted[\"actual\"])) * 100\n",
            "\n",
            "final_comparison = pd.DataFrame({\n",
            "    \"Модель\": [\"Математична модель\", \"Black-box
модель\", \"Prophet\"],\n",
            "    \"Рівень оцінювання\": [\"департамент — квартал\",
\"департамент — квартал\", \"агрегований квартальний ряд\"],\n",
            "    \"MAE\": [baseline_mae, blackbox_mae, prophet_mae],\n",
            "    \"RMSE\": [baseline_rmse, blackbox_rmse,
prophet_rmse],\n",
            "    \"MAPE\": [baseline_mape, blackbox_mape,
prophet_mape]\n",
            "})\n",
            "display(final_comparison)\n",
            "\n",
            "plt.figure(figsize=(10, 7))\n",
            "plt.bar(final_comparison[\"Модель\"],
final_comparison[\"MAPE\"])\n",
            "plt.title(\"Рисунок 14.1 – Порівняння MAPE використаних
моделей\")\n",
            "plt.xlabel(\"Модель\")\n",
            "plt.ylabel(\"MAPE, %\")\n",
            "plt.grid(True, axis=\"y\")\n",
            "plt.tight_layout()\n",
            "plt.savefig(\"figure_14_1_mape_models_comparison.png\",
dpi=200, bbox_inches=\"tight\")\n",
            "plt.show()\n",
            "\n",
            "criteria = [\"Точність\", \"Інтерпретованість\", \"Простота
реалізації\", \"Департаментальний аналіз\", \"Аналіз тренду\"]\n",
            "math_model_scores = [3, 5, 5, 3, 3]\n",
            "black_box_scores = [4, 3, 3, 5, 3]\n",
            "prophet_scores = [4, 4, 3, 1, 5]\n",
            "x = np.arange(len(criteria))\n",
            "width = 0.25\n",
            "plt.figure(figsize=(13, 7))\n",
            "plt.bar(x - width, math_model_scores, width,
label=\"Математична модель\")\n",
            "plt.bar(x, black_box_scores, width, label=\"Black-box
модель\")\n",
            "plt.bar(x + width, prophet_scores, width,
label=\"Prophet\")\n",
            "plt.xticks(x, criteria, rotation=25, ha=\"right\")\n",
            "plt.ylim(0, 5.5)\n",
            "plt.xlabel(\"Критерій\")\n",
            "plt.ylabel(\"Оцінка за умовною шкалою 1–5\")\n",
            "plt.title(\"Рисунок 14.2 – Узагальнене порівняння підходів за
критеріями\")\n",
            "plt.legend()\n",
```

```
                "plt.grid(True, axis=\"y\")\n",
                "plt.tight_layout()\n",

"plt.savefig(\"figure_14_2_approaches_criteria_comparison.png\", dpi=200,
bbox_inches=\"tight\")\n",
                "plt.show()"
              ]
            },
            {
              "cell_type": "markdown",
              "metadata": {},
              "source": [
                "## 19. Збереження основних таблиць у Excel"
              ]
            },
            {
              "cell_type": "code",
              "metadata": {},
              "execution_count": null,
              "outputs": [],
              "source": [
                "with pd.ExcelWriter(\"budget_modeling_results.xlsx\") as
writer:\n",
                "    df.to_excel(writer, sheet_name=\"Original data\",
index=False)\n",
                "    quarterly.to_excel(writer, sheet_name=\"Quarterly
aggregation\", index=False)\n",
                "    dept_quarter.to_excel(writer, sheet_name=\"Department
quarter\", index=False)\n",
                "    model_data.to_excel(writer, sheet_name=\"Model data\",
index=False)\n",
                "    comparison_metrics.to_excel(writer, sheet_name=\"Model
comparison\", index=False)\n",
                "    perm_importance.to_excel(writer, sheet_name=\"Permutation
importance\", index=False)\n",
                "    ablation_df.to_excel(writer, sheet_name=\"Ablation
study\", index=False)\n",
                "    anomaly_df.to_excel(writer, sheet_name=\"Anomaly
analysis\", index=False)\n",
                "    final_comparison.to_excel(writer, sheet_name=\"Final
comparison\", index=False)\n",
                "\n",
                "files.download(\"budget_modeling_results.xlsx\")"
              ]
            }
          ]
        }
```

ДОДАТОК Б ПРЕЗЕНТАЦІЯ

1.

1

Аналіз структури федеральних видатків США на основі статистичних даних

Дипломна робота бакалавра
Ващенко Віктор Костянтинович
Керівник: Сидорський Володимир Сергійович
КПІ ім. Ігоря Сікорського, 2026

2.

2

Актуальність і мета роботи

Аналіз бюджетних видатків за основними напрямами державного фінансування є важливим для фінансового планування, оцінювання відхилень між плановими та фактичними показниками й підтримки управлінських рішень.

Мета роботи:
побудувати й порівняти моделі прогнозування бюджетних видатків за функціональними напрямами та визначити ознаки, що найбільше впливають на результат.

Об’єкт:
процес аналізу та прогнозування бюджетних видатків за напрямами державної діяльності.

Предмет:
методи прогнозування фінансово-бюджетних часових рядів та інтерпретації моделей машинного навчання.

3.

3

Завдання і методи роботи

Основні завдання:

- провести первинний аналіз даних про бюджетні видатки за напрямами фінансування;
- підготувати дані до моделювання;
- побудувати базову математичну модель;
- побудувати модель **XGBoost Regressor**;
- застосувати **Prophet** для аналізу часового ряду;
- порівняти моделі за **MAE, RMSE, MAPE**;
- пояснити результати через **SHAP** та **Permutation Importance**.

Дані	→	**Підготовка**	→	**Моделі**	→	**Метрики**	→	**Інтерпретація**

4.

4

Датасет і підготовка даних

У роботі використано відкритий датасет **Budget Spending Dataset**, що містить дані про бюджетні видатки за шістьма напрямами фінансування організаційної діяльності: :

Operations, Sales, IT, HR, Marketing, Finance.

Змінна	Зміст
Month	місяць
Department	департамент
Budgeted Amount	заплановані витрати
Actual Spending	фактичні витрати
Variance	відхилення факту від плану

Підготовка даних:

місяці → квартали → лагові ознаки → прогноз Q4

Дані було агреговано за кварталами, а для прогнозування сформовано ознаки за трьома попередніми кварталами.

**Датасет описує бюджетні видатки за функціональними напрямами державної діяльності, а не окремими державними установами.*

5.

5

Якість і загальна структура даних

Перевірка якості даних показала, що в датасеті відсутні пропущені значення та повністю дубльовані записи. Дані охоплюють 12 місяців спостережень за шістьма напрямами державного фінансування.

Показник	Значення
Кількість записів	5000
Кількість змінних	5
Пропущені значення	0
Дублікати рядків	0
Кількість місяців	12
Кількість напрямів фінансування	6
Помилки у розрахунку Variance	0

Було додатково перевірено правильність розрахунку змінної Variance. Для всіх 5000 записів вона дорівнює різниці між фактичними та запланованими витратами:

$$Variance = Actual\ Spending - Budgeted\ Amount$$

Дані є повними та узгодженими, а загальний бюджет майже збалансований.

6.

6

Структура бюджетних витрат

Агрегація за кварталами:

Квартал	Фактичні витрати	Відхилення
Q1	15 488 338	−48 504
Q2	16 194 633	+10 390
Q3	15 462 789	+103 308
Q4	15 163 233	+30 953

Загальні витрати між департаментами розподілені відносно рівномірно, але відхилення від плану суттєво відрізняються.

7.

7

Моделі прогнозування бюджетних витрат

Прогнозування виконувалося за схемою: **Q1, Q2, Q3 → Q4**

Модель	Принцип роботи	Використані дані
Базова математична модель	прогноз як середнє трьох попередніх кварталів	фактичні витрати Q1–Q3
XGBoost Regressor	прогноз на основі набору ознак	лаги, ковзне середнє, планові витрати, департамент
Prophet	аналіз агрегованого часового ряду	квартальна динаміка фактичних витрат

Базова модель використовується як контрольний еталон, а XGBoost — як точніша black-box модель.

8.

8

Результати прогнозування

Порівняння якості моделей

Модель	MAE	RMSE	MAPE
Базова модель	267 116,67	296 582,87	10,67%
XGBoost Regressor	189 758,17	193 813,40	7,63%

Покращення XGBoost порівняно з базовою моделлю:

Метрика	Покращення
MAE	28,96%
RMSE	34,65%
MAPE	28,49%

XGBoost показав кращу якість прогнозування за всіма метриками, особливо за RMSE, тобто краще зменшив великі помилки прогнозу.

9.

9

Аналіз важливості ознак

Для пояснення роботи XGBoost використано:

Метод	Призначення
Permutation Importance	визначає, які ознаки найбільше впливають на якість прогнозу
SHAP-аналіз	пояснює внесок ознак у конкретні прогнозні значення

Найважливіші групи ознак:

- лагові значення фактичних видатків;
- ковзне середнє за три квартали;
- планові видатки;
- департаментальна належність.

Ключовий висновок:
Модель найбільше спирається на попередню динаміку фактичних видатків, тоді як планові видатки та напрям фінансування мають менший вплив на результат прогнозування.

10.

10

Prophet: загальна квартальна динаміка

Prophet застосовано для аналізу агрегованого квартального ряду:

Q1 → Q2 → Q3 → Q4 → наступний квартал

Ключовий висновок:

Prophet показав зниження витрат після локального максимуму в Q2 та прогнозує наступний квартал приблизно на рівні **15,12 млн**.

Період	Фактичні видатки	Прогноз Prophet
Q1	15 488 338	15 540 000
Q2	16 194 633	15 880 000
Q3	15 462 789	15 520 000
Q4	15 163 233	15 230 000
Наступний квартал	—	15 120 000

11.

11

Залишки й аномалії прогнозування

Показник	Значення
Найбільший додатний залишок Prophet	Q2: +314 633
Найбільший від'ємний залишок Prophet	Q4: −66 767
Основне джерело помилки	локальний максимум Q2
Загальна поведінка Prophet	згладжує різкі коливання

Ключові спостереження:

- квартальний ряд не містить сильних статистичних аномалій;
- Q2 є локальним максимумом витрат;
- Q4 має найнижчий рівень витрат;
- базова модель завищує Q4 через високе значення Q2;
- XGBoost краще враховує департаментальні особливості.

12.

12

Роль кожного підходу в дослідженні

Підхід	Рівень аналізу	Переваги	Обмеження
Базова математична модель	квартали / департаменти	проста, прозора, легко пояснюється	нижча точність
XGBoost Regressor	департаменти	найкраща точність, враховує кілька факторів	потребує SHAP та Permutation Importance
Prophet	агрегований квартальний ряд	зручний аналіз тренду й динаміки	обмежений через короткий часовий ряд

Жодна модель не є універсально найкращою: базова модель потрібна як еталон, XGBoost – для точнішого прогнозування, Prophet – для аналізу загального тренду.

13.

13

Висновки роботи

У роботі було:

- проведено первинний аналіз датасету бюджетних витрат;
- виконано агрегацію даних за кварталами та департаментами;
- побудовано базову математичну модель прогнозування;
- побудовано XGBoost Regressor модель;
- застосовано Prophet для аналізу загальної динаміки;
- виконано інтерпретацію результатів через SHAP та Permutation Importance.

Основний результат:
XGBoost показав кращу якість прогнозування: **MAPE зменшилося з 10,67% до 7,63%**.

Практичне значення:
запропонований підхід дозволяє аналізувати бюджетні витрати, оцінювати відхилення між планом і фактом та визначати найважливіші фактори прогнозування.

14.

14

Дякую за увагу!