ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ТЕЛЕКОМУНІКАЦІЙ
КАФЕДРА СИСТЕМНОГО АНАЛІЗУ

КВАЛІФІКАЦІЙНА РОБОТА

на тему: « МЕТОДИ ТА ЗАСОБИ ПІДВИЩЕННЯ ТОЧНОСТІ ПРОГНОЗУВАННЯ СЕЗОННИХ ПРОДАЖІВ У СФЕРІ FMCG»

на здобуття освітнього ступеня магістра
зі спеціальності: 124 Системний аналіз
освітньо-професійної програми Інтелектуальні системи управління

Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ *Марина Цибенко*

Виконала: здобувачка вищої освіти гр. САДМ-61
Марина Цибенко
Керівник: Сергій Козаченко
к.е.н., доцент кафедри СА
Рецензент:
науковий

Київ 2023

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
Навчально-науковий інститут телекомунікацій

Кафедра Системного аналізу

Ступінь вищої освіти Магістр

Спеціальність 124 Системний аналіз

Освітньо-професійна програма Інтелектуальні системи управління

ЗАТВЕРДЖУЮ
Завідувач кафедрою Равіль Нафаєєв
«____» _____20___р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Цибенко Марина Юріївна

1. Тема кваліфікаційної роботи: Методи та засоби підвищення точності прогнозування сезонних продажів у сфері FMCG

керівник кваліфікаційної роботи: Сергій Козаченко, к.е.н., доцент кафедри СА

затверджені наказом Державного університету інформаційно-комунікаційних технологій від «___» _____20___р. № ____

2. Строк подання кваліфікаційної роботи «___» _____20___р.

3. Вихідні дані до кваліфікаційної роботи: _____
_____
_____

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)
1. _____
2. _____
3._____

5. Перелік ілюстративного матеріалу: *презентація*

6. Дата видачі завдання «___» _____20___р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1			
2			
3			
4			
5			
6			
7			
8			

Здобувач(ка) вищої освіти _____ Марина Цибенко
(підпис)

Керівник
кваліфікаційної роботи _____ Сергій Козаченко
(підпис)

РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття освітнього ступеня бакалавра (магістра): 67 стор., 16 рис., 14 табл., 22 джерел.

Мета роботи – метою дипломного проекту є покращення точності прогнозування сезонних продажів у сфері FMCG (товари широкого вжитку). Це буде досягнуто шляхом огляду існуючих методів прогнозування, вибору найбільш популярного та доступного методу, і подальшого вдосконалення його точності. Ця робота має на меті забезпечити більш ефективне управління запасами та планування продажів для компаній у сфері FMCG, що, в свою чергу, може призвести до зменшення витрат і збільшення доходів.

Об'єкт дослідження – об'єктом дослідження є процес прогнозування сезонних продажів у сфері FMCG. Це включає аналіз існуючих підходів та систем, використовуваних компаніями для прогнозування попиту на свої товари.

Предмет дослідження – предметом дослідження є методи та засоби прогнозування, які можуть бути використані для підвищення точності визначення майбутніх продажів FMCG продуктів. Особлива увага буде приділена алгоритмам машинного навчання та статистичним методам, що є найбільш доступними і популярними у сучасному бізнес-середовищі.

Короткий зміст роботи: Дослідження історичних даних для розуміння тенденцій та сезонності у FMCG. Вивчення впливів зовнішніх факторів: як економічні, так і соціальні фактори можуть впливати на продажі. Аналіз різних методів прогнозування, включно з традиційними статистичними методами та сучасними підходами, такими як машинне навчання.

КЛЮЧОВІ СЛОВА: FMCG, експотенційне згладжування, ковзаюче середнє, ARIMA, сезонна декомпозиція, «випадковий ліс», прогноз, метод Холта-Вінтерса.

ABSTRACT

Text part of the master's qualification work: 67 pages, 16 pictures, 14 tables, 22 sources.

The purpose of the work is to improve the accuracy of forecasting seasonal sales in the FMCG (Fast-Moving Consumer Goods) sector. This will be achieved by reviewing existing forecasting methods, selecting the most popular and accessible method, and enhancing its accuracy. The aim is to provide more efficient inventory management and sales planning, potentially reducing costs and increasing revenue for FMCG companies.

The object of the research is the process of forecasting seasonal sales in the FMCG sector, including the analysis of existing approaches and systems used by companies for demand forecasting.

The subject of the research is forecasting methods and tools that can be used to increase the accuracy of predicting future sales of FMCG products, with a focus on machine learning and statistical methods commonly used in modern business environments.

Summary of the work: Researching historical data to understand trends and seasonality in FMCG, studying the impacts of external factors – both economic and social – on sales, and analyzing various forecasting methods, including traditional statistical methods and modern approaches like machine learning.

KEYWORDS: FMCG, exponential smoothing, moving average, ARIMA, seasonal decomposition, random forest, forecast, Holt-Winters method.

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Навчально-науковий інститут телекомунікацій

ПОДАННЯ
ГОЛОВІ ЕКЗАМЕНАЦІЙНОЇ КОМІСІЇ
ЩОДО ЗАХИСТУ КВАЛІФІКАЦІЙНОЇ РОБОТИ
на здобуття освітнього ступеня бакалавра (магістра)

Направляється здобувачка Цибенко М.Ю. до захисту кваліфікаційної роботи

за спеціальністю 124 Системний аналіз

освітньо-професійної програми Інтелектуальні системи управління

на тему: «Методи та засоби підвищення точності прогнозування сезонних продажів у сфері FMCG».

Кваліфікаційна робота і рецензія додаються.

Директор ННІ _____ _____
(підпис) *(Ім'я, ПРІЗВИЩЕ)*

Висновок керівника кваліфікаційної роботи

Здобувачка _____
_____
_____
_____

Все це дозволяє оцінити виконану кваліфікаційну роботу здобувача(ки) _____на оцінку «_____» та присвоїти йому(їй) кваліфікацію _____.

Керівник кваліфікаційної роботи _____ Сергій Козаченко
(підпис) \

«____» _____20___року

Висновок кафедри про кваліфікаційну роботу

Кваліфікаційна робота розглянута. Здобувачка Цибенко М.Ю. допускається до захисту даної роботи в Екзаменаційній комісії.

Завідувач кафедрою Системного аналізу _____ Равіль Нафаєєв
(підпис)

ВІДГУК РЕЦЕНЗЕНТА
на кваліфікаційну роботу

на

здобувачки вищої освіти Цибенко Марини Юріївни

на тему « Методи та засоби підвищення точності прогнозування сезонних продажів у сфері FMCG»

Актуальність.

_____

_____

_____

Позитивні сторони.

1.
2.
3.

Недоліки.

1.
2.

Відзначені зауваження не впливають на загальну позитивну оцінку кваліфікаційної роботи магістерської.

Висновок: *кваліфікаційна робота на здобуття ступеня магістра заслуговує оцінку "_____", а здобувач(ка) _____ заслуговує присвоєння кваліфікації: ……..*

здобуття освітнього ступеня бакалавра (магістра)

Рецензент:

науковий ступінь, вчене звання	_____ *підпис*	_____ *Ім'я, ПРІЗВИЩЕ*

ЗМІСТ

ВСТУП

Прогнозування у сфері FMCG (товари широкого вжитку) є важливим для оптимізації ланцюгів поставок, управління запасами, планування виробництва та фінансового планування. Точність прогнозування сезонних продажів дозволяє компаніям уникати витрат на зайві запаси, запобігати дефіциту товарів та адаптуватися до змін у споживчих перевагах. Вдосконалення точності прогнозів також сприяє кращому розумінню ринкових тенденцій та ефективнішому використанню ресурсів, що в кінцевому підсумку підвищує конкурентоспроможність та прибутковість компаній.

Згідно з дослідженням The Aberdeen Group, компанії з передовими процесами прогнозування продажів досягають своїх квартальних цілей у 97% випадків, в порівнянні з 55% компаній, що не використовують такі передові підходи [1].

Прогнозування дозволяє командам з доходів виявляти ранні сигнали у своїх продажах та вносити корективи, поки не стало занадто пізно. Компанії, які володіють наукою прогнозування продажів, на 10% частіше збільшують свої річні доходи та удвічі частіше обганяють своїх конкурентів у галузі [2].

У сфері FMCG кон'юнктура ринку може швидко змінюватися під впливом трендів, сезонності та інших факторів. Забезпечення продукції в потрібний момент стає важливим завданням для компаній, і точні прогнози є ключовим елементом для управління цими змінами. Точні прогнози попиту дозволяють компаніям уникати надмірного або недостатнього запасу товарів. Це сприяє ефективному управлінню ланцюгом постачання, зменшує витрати на зберігання та мінімізує ризик втрат від непроданих товарів або простою виробництва.

FMCG включає в себе продукти, які споживаються щоденно або регулярно, такі як продукти харчування, напої, товари особистої гігієни тощо. Забезпечення людей

цими товарами вчасно є ключовим для їхнього комфорту та задоволення базових потреб.

Сучасні технології, такі як штучний інтелект, аналітика даних та машинне навчання, відкривають нові можливості для підвищення точності прогнозів. Дослідження та впровадження цих методів може принести значні вигоди компаніям у сфері FMCG та інших сфер. Розробка та впровадження методів прогнозування вимагає відповідної інфраструктури, яка забезпечить збір та аналіз необхідних даних, взаємодію зі стейкхолдерами та ефективне впровадження отриманих результатів.

Оцінка поточної точності прогнозів є ключовим елементом управління прогнозуванням продажів, особливо у сфері FMCG. Цей процес включає кілька кроків:

- Збір історичних даних про продажі та порівняння їх з прогнозами, які були зроблені на відповідний період.
- Виявлення та аналіз розбіжностей між прогнозованими та фактичними продажами.
- Використання таких метрик, як середня абсолютна відсоткова помилка (MAPE), коренева середньоквадратична помилка (RMSE) [3] для оцінки точності прогнозів.
- Врахування зовнішніх факторів, таких як сезонність, ринкові тенденції, та економічні умови, які можуть впливати на точність прогнозів.
- Аналіз різних методів прогнозування для визначення найефективніших.
- Аналіз даних для виявлення трендів і відхилень, які можуть вказувати на потребу у коригуванні методів прогнозування.
- Модифікація та налаштування моделей прогнозування з метою збільшення їх точності.
- Підвищення кваліфікації працівників, які відповідають за прогнозування, для забезпечення кращого розуміння та використання даних.

– Регулярний перегляд та оновлення процесів прогнозування для забезпечення їх актуальності та ефективності.

Аналіз впливу неточних прогнозів на бізнес-операції в сфері FMCG є важливим для забезпечення ефективного управління ланцюгами поставок та фінансами [4]. Неточні прогнози можуть призвести до ряду проблем та наслідків:

Таблиця 1

Проблеми та наслідки неточних прогнозів на бізнес-операції в сфері FMCG

Проблема	Наслідки
Надлишки запасів	Призводить до зайвих витрат на зберігання та потенційно до знецінення товарів.
Дефіцит запасів	Ризик втрати продажів та незадоволених клієнтів через нестачу товарів.
Нестабільність у виробництві	Викликає проблеми з плануванням та оптимізацією виробничих потужностей.
Фінанси	Фінансові втрати через неправильне планування та невикористання ресурсів.
Втрата ринкової частки	Конкуренти можуть скористатися ситуацією нестачі або перенасичення ринку.
Погіршення відносин з партнерами	Нестабільні замовлення можуть ускладнити співпрацю з постачальниками.
Підвищений ризик зіпсування продукції	Особливо актуально для товарів з обмеженим терміном придатності.
Зниження довіри споживачів	Часті зміни асортименту або цін можуть відштовхувати клієнтів.
Втрата операційної гнучкості	Компанія стає менш здатною швидко реагувати на зміни ринкових умов.
Підвищений оперативний стрес	Робоче навантаження та стрес зростають через необхідність вирішення проблем, пов'язаних з неправильним прогнозуванням.

Основна задача полягає в виборі методів, які дозволять компаніям визначити основні фактори, що впливають на точність прогнозування у сфері FMCG. Це

передбачає розробку підходів для покращення, шляхом виявлення впливових змінних та вдосконалення методів прогнозування з метою отримання кращих результатів. Виявлення основних факторів, які впливають на точність прогнозування, включаючи сезонні зміни, споживчу поведінку, економічні та ринкові умови.

РОЗДІЛ 1. ОГЛЯД ІСНУЮЧИХ МЕТОДІВ ПРОГНОЗУВАННЯ

Цей розділ присвячений найпопулярнішим методам прогнозування в сфері FMCG. Ми розглянемо як традиційні методи, такі як методи ковзаючої середньої та експоненційного згладжування, так і сучасні підходи, включаючи використання машинного навчання та аналізу часових рядів. Традиційні методи надають нам основні інструменти для прогнозування на основі історичних даних, тоді як сучасні методи використовують складніші алгоритми та технології для покращення точності прогнозування в умовах змінюючогося ринку FMCG. Наприклад, машинне навчання може аналізувати багатофакторні взаємодії та надавати більш точні прогнози, що допомагає компаніям ефективніше управляти запасами та задовольняти попит споживачів.

1.1. Традиційні методи: Опис та аналіз традиційних статистичних методів прогнозування

Традиційні статистичні методи прогнозування охоплюють кілька ключових підходів:

1. Часові ряди: Аналізують історичні дані для виявлення тенденцій та сезонності. Приклади включають методи ковзаючої середньої та експоненційного згладжування.
2. Причинно-наслідкові моделі: Використовують змінні, які вважаються причинами для прогнозування вихідних показників. Регресійний аналіз є класичним прикладом.
3. Економетричні моделі: Комбінують статистичні методи з економічною теорією для прогнозування. Вони враховують більше факторів, ніж прості часові ряди.

4. Кваліативні методи: Включають експертні оцінки та судження для ситуацій, де даних недостатньо або вони неякісні.

Кожен з наведених методів мають свої переваги та обмеження, залежно від доступності даних, потреби у точності та контексту застосування. Розглянемо детальніше найпопулярніші з них.

Метод ковзаючої середньої (Moving Average) – це техніка прогнозування, в якій для кожного прогнозуваного періоду обчислюється середнє значення попередніх спостережень на фіксованому проміжку часу (наприклад, 3 місяці). Це допомагає згладжувати коливання та виділяти тренди в часовому ряді, що робить його корисним для прогнозування сезонних чи краткострокових змін. Перевагами даного методу є його простота в розумінні та використанні; ефективний для згладжування короткотермінових коливань, але він не враховує тренди та сезонні коливання, запізнюється з реакцією на зміни у даних [5].

Експоненційне згладжування (Exponential Smoothing) – це метод прогнозування, в якому ваги надаються попереднім значенням часового ряду з експоненційною зміною ваги. Це дозволяє виділяти тренди та сезонні компоненти в даному ряді даних. Перевагами даного методу є краще реагування на недавні зміни в даних та є простим у застосуванні. Обмеженнями може бути неточність при сильних трендах чи сезонності [6].

ARIMA (Autoregressive Integrated Moving Average) є статистичним методом прогнозування часових рядів. Він включає авторегресійну (AR) компоненту для моделювання залежності від попередніх значень, інтегровану (I) частину для стабілізації ряду, і компоненту ковзаючої середньої (MA) для моделювання шуму. ARIMA може прогнозувати тренди та сезонні зміни в часових рядах. Цей метод застосовується в економетриці, фінансах та інших галузях для прогнозування часових рядів. Метод нучкий, може моделювати різноманітні часові ряди та враховує тренди

та сезонність, але складний у розумінні та налаштуванні, вимагає значної кількості історичних даних [7].

Сезонна декомпозиція часових рядів – це процес розділення часового ряду на кілька складових: тренд, сезонність, і випадкові відхилення. Це дозволяє аналізувати окремі компоненти серії та краще розуміти її поведінку. Наприклад, використовуючи метод сезонної декомпозиції, можна виявити типові шаблони повторення в даних, які пов'язані з певними сезонами року або іншими циклічними факторами. Метод дозволяє якісно розділити тренд, сезонність та випадковість, корисний для планування на основі сезонних факторів, проте менш ефективний для нетипових або нерегулярних часових рядів [8].

1.2. Сучасні підходи: Вивчення методів, заснованих на машинному навчанні та інших інноваційних технологіях

Сучасні підходи до прогнозування включають використання машинного навчання, глибокого навчання та штучного інтелекту для аналізу та прогнозування складних даних. Вони дозволяють автоматизувати процес прогнозування, враховуючи багатофакторні взаємодії та нестаціонарність даних. Сучасні підходи також використовують розширені методи оцінки точності, такі як крос-валідація та метрики ефективності, для покращення результатів прогнозування. Нейронні мережі здатні виявляти складні шаблони та залежності у даних. Методи ансамблю можуть комбінувати прогнози з декількох моделей для покращення точності. Машинне навчання використовує історичні дані для тренування моделі.

Ці методи вимагають значних обсягів даних та обчислювальних ресурсів, але здатні давати більш точні та адаптивні прогнози.

Машинне навчання з використанням методу «Випадковий ліс» (Random Forest) – це потужний алгоритм, який використовується для класифікації та регресії. Він

працює шляхом створення великої кількості дерев рішень, кожне з яких навчається на випадковій підмножині даних. Потім результати цих дерев комбінуються для покращення точності та стабільності прогнозу. Особливо ефективний у випадках, коли є великі набори даних з багатьма функціями. Він здатний виявляти важливі функції та має високу здатність до узагальнення, знижуючи ризик перенавчання. Однак, він може бути відносно важким для інтерпретації та вимагає великої обчислювальної потужності для тренування та виконання. Ефективний для складних наборів даних із високою точністю. Виявляє важливі ознаки, але вимагає значних обчислювальних ресурсів. Складний для інтерпретації результатів. Має хорошу здатність до узагальнення, але менш прозорий у порівнянні з лінійними моделями [9].

Методи машинного навчання вимагають значних зусиль та інфраструктури, тому іноді ці методи не є доступними з фінансової точки зору, тому широкого розповсюдження набирають no-code та low-code платформи для прогнозування, такі як Azure AI від компанії Microsoft.

Azure AI пропонує великий набір інструментів та послуг для прогнозування та аналізу часових рядів. Сервіс Azure Machine Learning дозволяє користувачам створювати та розгортати власні моделі машинного навчання для прогнозування. Azure Time Series Insights надає можливість дослідження та аналізу даних в режимі реального часу. Крім того, Azure пропонує готові AI-моделі для прогнозування, виявлення аномалій та прогнозування попиту завдяки Azure Cognitive Services. Ці послуги дозволяють підприємствам використовувати штучний інтелект і машинне навчання для точних прогнозів, оптимізації операцій та покращення прийняття рішень на основі історичних і реальних даних. Azure AI завдяки масштабованості і інтеграції з іншими послугами Azure стає потужною платформою для прогнозування та аналізу часових рядів [10].

Azure AI має численні переваги для прогнозування, включаючи:

1. Масштабованість: Azure AI може обробляти великі обсяги даних та

працювати з великою кількістю моделей, що дозволяє розглядати складні завдання прогнозування.

2. Гнучкість: Платформа надає можливість використовувати готові AI-моделі або створювати власні, в залежності від потреб бізнесу.
3. Інтеграція: Azure AI легко інтегрується з іншими послугами Azure, що дозволяє побудувати повний стек рішень для аналізу та прогнозування даних.
4. Підтримка реального часу: Azure Time Series Insights дозволяє виконувати аналіз даних в режимі реального часу для оперативного прийняття рішень.
5. Готові рішення: Azure пропонує готові моделі для різних видів прогнозування, що спрощує розгортання та використання AI в бізнесі.

Ці переваги роблять Azure AI потужним інструментом для прогнозування та аналізу часових рядів у сучасних бізнес-сценаріях, але він має також ряд і недоліків, які включають високі витрати для великих обсягів даних, необхідність експертної підготовки для використання деяких функцій, а також можливу складність налаштування та оптимізації моделей. Підприємствам також слід бути обережними зі зберіганням та захистом даних, коли вони використовують хмарні рішення. До інших можливих недоліків включають обмежені можливості безпосереднього доступу до алгоритмів машинного навчання та необхідність стежити за оновленнями та змінами в платформі.

1.3. Критичний аналіз: Порівняння ефективності різних методів.

Порівнюючи різні методи прогнозування в сфері FMCG, можна визначити їхні переваги та обмеження. Традиційні методи, як ковзаюча середня та експоненційне згладжування, мають низькі витрати на обчислення та простоту в реалізації, але можуть бути менш точними при складних динаміках даних. ARIMA дозволяє

моделювати більш складні залежності, але вимагає більше фаховості та обчислювальних ресурсів. Сезонна декомпозиція надає можливість точно прогнозувати сезонні ефекти, але потребує додаткового аналізу. Машинне навчання, таке як «Випадковий ліс», може бути дуже точним, але вимагає великої кількості даних та обчислювальних ресурсів. Вибір методу повинен залежати від конкретних вимог бізнесу, доступних ресурсів та бажаного рівня точності.

Експоненційне згладжування краще використовувати, коли в часовому ряді присутня сезонність або коли тренд в даних є нестабільним. Цей метод дозволяє ефективно моделювати як тренд, так і сезонні варіації у даних, що робить його більш точним у таких випадках. Крім того, експоненційне згладжування може бути корисним, коли дуже важливо враховувати недавні значення у прогнозі, оскільки воно надає більшу вагу останнім спостереженням.

Метод ковзаючої середньої може бути кращим в тих випадках, коли треба здійснити швидкий та простий прогноз, і коли сезонність або тренд не є критичними факторами. Він використовується для стабільних часових рядів без значних варіацій. Метод ковзаючої середньої підходить для швидких оцінок, але менш точний, оскільки не враховує складніші залежності в даних.

Авторегресійний інтегрований метод ковзаючої середньої (ARIMA) краще використовувати в тих випадках, коли у часовому ряді присутні складні тренди, сезонні коливання та автокореляція між спостереженнями. ARIMA дозволяє моделювати та прогнозувати такі складні залежності у даних і зазвичай є точнішим за прості методи, такі як ковзаюча середня. Він підходить для складних часових рядів, де важливо враховувати історичні зміни та зв'язки між даними.

Сезонну декомпозицію часових рядів найкраще використовувати, коли в часовому ряді існує виражений сезонний компонент, який повторюється на регулярній основі. Цей метод дозволяє відокремити тренд, сезонність та випадковий шум у даних, що полегшує аналіз та прогнозування. Він корисний для прогнозування явищ, які

підпорядковані сезонним коливанням, таким як сезонні продажі у сфері FMCG або метеорологічні дані.

Машинне навчання з використанням методу «Випадковий ліс» (Random Forest) є кращим в випадках, коли у часовому ряді є складні та нелінійні залежності, а також багато факторів, які впливають на прогноз. Випадковий ліс ефективно враховує велику кількість параметрів і може адаптуватися до різних структур даних. Він підходить для завдань, де точність прогнозування є критично важливою, і коли є велика кількість даних для навчання моделі.

РОЗДІЛ 2. ВИБІР МЕТОДУ ДЛЯ ПІДВИЩЕННЯ ТОЧНОСТІ

Для вибору методу підвищення точності прогнозування, необхідно провести ретельний аналіз та розрахунки для кожного методу, враховуючи критерії важкості імплементації, часу на навчання та прогнозування, рівня кваліфікації персоналу і наявної інфраструктури. Це допоможе визначити, який метод найбільше відповідає конкретним потребам компанії та бюджету, забезпечуючи найкращий баланс між точністю та витратами ресурсів, для цього буде обрано критерії та застосований метод експертних оцінок для визначення основного методу.

2.1. Критерії вибору: Визначення критеріїв для вибору найбільш підходящого методу

Методи, що були обрані для аналізу будуть використані для прогнозу на реальних даних отриманих з відкритого ресурсу Kaggle для Sales in Norway в локальному рітейлі. Дані для машинного навчання будуть розбиті на дві частини – тренувальні (90%) та порівняльні (10%), для визначення якості прогнозу.

Датасет було перевірено за допомогою теста Стьюдента. Тест Стьюдента (t-тест) використовується для визначення статистичної значущості різниці між середніми значеннями двох груп даних. Він допомагає визначити, чи є ця різниця обумовленою статистично, чи вона може бути випадковою.

Для зрозуміння, коли різниця є статистично значущою, використовують значення p (p-value), якщо p-value менше ніж 0,05 (або інший попередньо встановлений рівень значущості), то можна вважати, що різниця обумовлена статистично. В нашому випадку p-value менше ніж 0,05 (3.16E-164).

Основні характеристики для порівняння: точність прогнозу, час для імплементації, вартість інфраструктури, вартість ресурсів для імплементації, підтримка та редагування, складність інтерпритації.

2.1.1. Метод ковзаючої середньої

Зважаючи на те, що прогнозування за допомогою методу ковзаючої середньої є досить простим, для цього ми будемо використовувати Excel, так як наші дані по місяцям, то за основу ми братимемо річну ковзаючу середню.

Основа ковзаючої середньої – це визначення середньої на виділеному відрізку часу:

$$SMA = \frac{\sum_{i=1}^{n} Pi}{n},$$

де *Pi* – величина, що прогнозується (продажі) відвідно до обраного періоду часу, n – довжина згладжування або період SMA.

При розрахунку даним методом, деталі якого наведені в додатку ми отримуємо таблицю похибок (|Error| – сума різниць прогнозованого та фактичного значень по модулю, Error ^2 – сума різниць прогнозованого та фактичного значень в квадраті, |% Error| – сума відхилення прогнозованого та фактичного значень в відсотках):

Таблиця 2.1

Похибки методу ковзаючої середньої

|Error|	Error ^2	|% Error|
101,953.00	159,194,006.33	998.98%

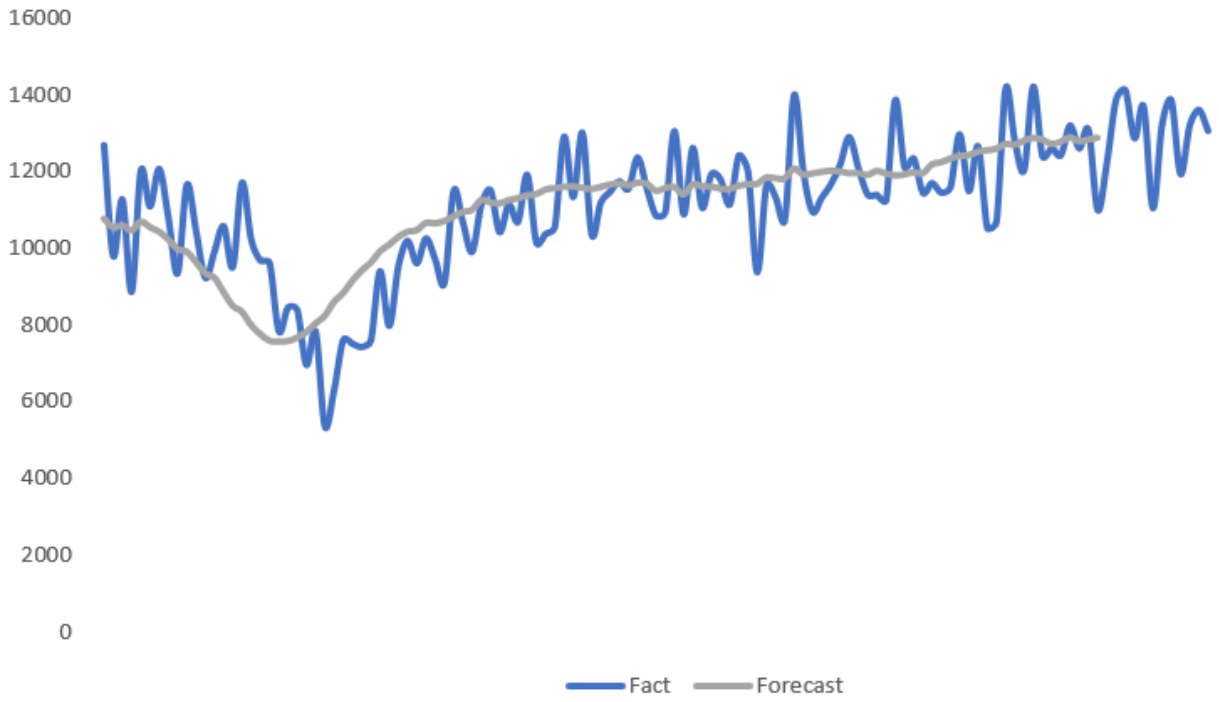


Рис. 2.1 Фактичні та прогнозовані дані кількості продажів в Норвегії за період з 2007 по 2017 роки методом ковзаючої середньої

Для отримання оцінки якості прогнозу, розрахуємо основні похибки:

1. Середня абсолютна похибка (MAE) – це метрика, яка використовується для оцінки точності прогнозів у задачах прогнозування. Вона вимірює середню величину абсолютних похибок між прогнозованими значеннями і фактичними спостереженнями. Щоб розрахувати MAE, потрібно взяти модуль різниці між кожним прогнозом і відповідним фактичним значенням, потім обчислити середнє арифметичне цих модулів. MAE вимірюється в тих же одиницях, що і вихідні дані і дозволяє оцінити середню величину похибок у прогнозах. Чим менше MAE, тим точніше прогнози [11]

$$MAE = \frac{1}{n}\sum |e_i|,$$

де e_i – сума модулів різниці між кожним прогнозом і відповідним фактичним значенням, n – кількість прогнозів.

У випадку прогнозування на відкритих даних за допомогою ковзаючого середнього величина MAE = 935.35.

2. Середньоквадратична помилка (MSE) – це метрика для вимірювання точності прогнозів у задачах прогнозування. Вона обчислюється як середнє арифметичне квадратів різниць між прогнозованими і фактичними значеннями. MSE дозволяє виявити величину помилок та визначити, наскільки вони відрізняються від фактичних даних. Чим менше значення MSE, тим точніше прогнози. Метрика особливо корисна в задачах, де більші помилки пов'язані з більшими штрафами, оскільки вона враховує квадратичну природу помилок [11]

$$MSE = \frac{1}{n}\sum \left|e_i^2\right|,$$

де e_i – сума модулів різниці між кожним прогнозом і відповідним фактичним значенням, n – кількість прогнозів.

У випадку прогнозування на відкритих даних за допомогою ковзаючого середнього величина MSE = 159,194,006.

3. Середньоквадратична помилка (RMSE) – це метрика для вимірювання точності прогнозів. Вона обчислюється як квадратний корінь з середньоквадратичної відстані між прогнозованими і фактичними значеннями. RMSE показує величину середньої помилки у тих самих одиницях вимірювання, що і вихідні дані. Чим менше значення RMSE, тим точніше прогнози.

$$RMSE = \sqrt{MSE},$$

де MSE – середнє арифметичне квадратів різниць між прогнозованими і фактичними значеннями.

У випадку прогнозування на відкритих даних за допомогою ковзаючого середнього величина RMSE = 1,208.51.

4. Середня абсолютна відсоткова помилка (MAPE) – це метрика для вимірювання точності прогнозів. Вона обчислюється як середнє арифметичне відсоткових різниць між прогнозованими і фактичними значеннями. MAPE виражає точність прогнозів у відсотках і дозволяє оцінити, наскільки великі відсоткові помилки в прогнозах. Чим менше значення MAPE, тим точніше прогнози. Метрика широко використовується в задачах прогнозування, особливо в економічних і фінансових аналітичних дослідженнях [11].

$$\text{MAPE} = \frac{1}{n}\sum \left|\frac{e_i}{y_i}\right|,$$

де e_i – прогнозоване значення, yi – фактичне значення, n – кількість прогнозів.

У випадку прогнозування на відкритих даних за допомогою ковзаючого середнього величина MAPE = 9.16%.

Метод ковзаючого середнього не потребує конкретних навичок чи додаткової інфраструктури, він також достатньо безпечний та не потребує багато часу на імплементацію, але він не є достатньо точним.

2.1.2. Експоненційне згладжування

Експоненційне згладжування є методом прогнозування, що використовується для аналізу часових рядів. Він вважається удосконаленням простої ковзаючої середньої, оскільки зосереджується на останніх даних, надаючи їм більшу вагу.

У методі експоненційного згладжування використовується коефіцієнт згладжування (α), що варіюється від 0 до 1. Вищий коефіцієнт згладжування означає, що більше ваги надається останнім спостереженням. Розглянемо два варіанти при α = 0.2 та α = 0.5.

Формула однократного експоненційного згладжування:

$$S_t = \alpha Y_t + (1 - \alpha) S_{t-1},$$

де α – коефіцієнт згладжування, Y_t – спостереження в момент часу t, S_{t-1} – згладжене значення у попередній момент часу.

Початкове згладжене значення (S_0) може бути встановлене як перше спостереження або як середнє значення серії [12].

При розрахунку методом експоненційного згладжування, деталі якого наведені в додатку ми отримуєму таблицю похибок (|Error| – сума різниць прогнозованого та фактичного значень по модулю, Error ^2 – сума різниць прогнозованого та фактичного значень в квадраті, |% Error| – сума відхилення прогнозованого та фактичного значень в відсотках) для кожного варіанту коефіцієнт згладжування:

Таблиця 2.2

Похибки методу експоненційного згладжування при α = 0.2 та α = 0.5

	|Error|	Error ^2	|% Error|
α = 0.2	113,555.02	177,844,455.36	1102.86%
α = 0.5	105,573.16	151,319,560.22	980.56%

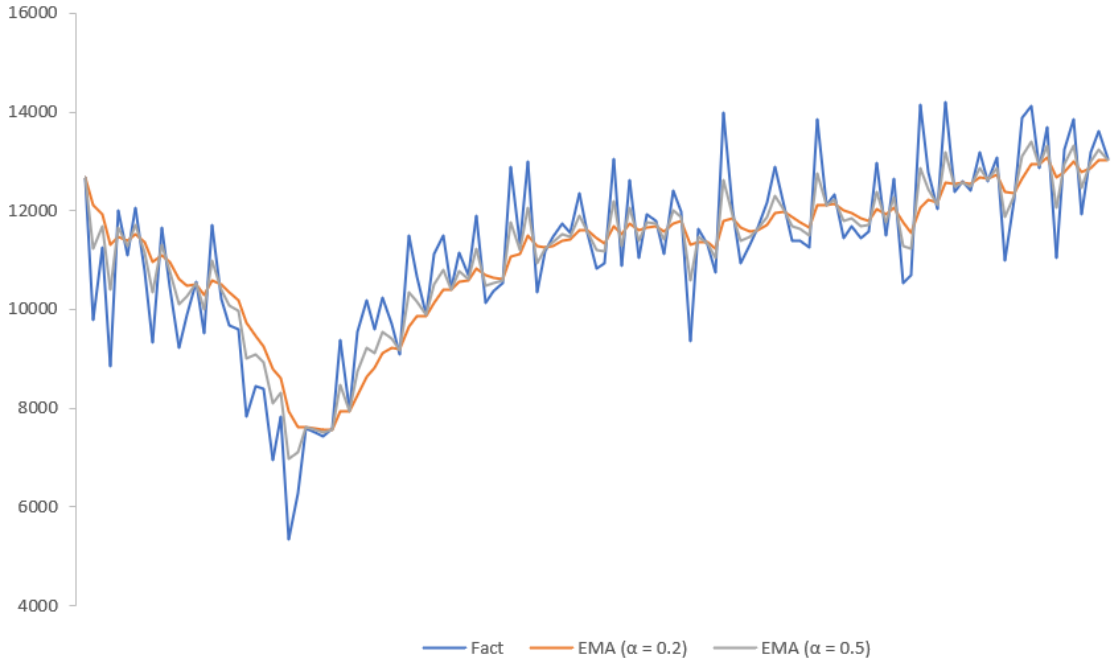


Рис. 2.2 Фактичні та прогнозовані дані кількості продажів в Норвегії за період з 2007 по 2017 роки методом експоненційного згладжування при α = 0.2 та α = 0.5

Для отримання оцінки якості прогнозу, розрахуємо основні похибки:

Таблиця 2.3

Оцінка якості прогнозу експоненційного згладжування при α = 0.2 та α = 0.5

	α = 0.2	α = 0.5
MAE	938.47	872.51
MSE	1,469,788.89	1,250,574.88
RMSE	1,212.35	1,118.29
MAPE	9.11%	8.10%

Метод експоненційного згладжування допомагає ефективно реагувати на зміни у даних, але може бути менш точним при виражених трендах або сильній сезонності. Це також однокроковий метод згладжування, який не враховує тренди чи сезонні компоненти, що може бути обмеженням у деяких сценаріях. Даний метод, як і метод

ковзаючого середнього не потребує конкретних навичок чи додаткової інфраструктури, він також достатньо безпечний та не потребує багато часу на імплементацію, точність краща за ковзаюче середнє, але все ще недостатня.

2.1.3. ARIMA (Авторегресійний інтегрований метод ковзаючої середньої)

ARIMA (Авторегресійний інтегрований метод ковзної середньої) – це складний статистичний метод, використовуваний для аналізу та прогнозування часових рядів. ARIMA об'єднує три основні компоненти: авторегресію (AR), інтеграцію (I), та ковзаючу середню (MA). ARIMA моделі дуже гнучкі і можуть адаптуватися до різноманітних структур даних, але вони вимагають ретельного вибору параметрів та великої кількості історичних даних для точного прогнозування [13].

Даний метод прогнозування може бути реалізований шляхом автоматичних функцій в Excel, а також за допомогою мови програмування Python або R.

Основною метою дипломної роботи є вивчення та покращення методів прогнозування для широкого загалу компаній незалежно від розмірів та з мінімальними витратами та ресурсами, тому реалізацію методу ARIMA буде імплементовано за допомогою Excel.

Першим кроком для імплементації моделі необхідно вивчити тренди наших даних. На рисунку 2.3 ми бачимо тренди продажів з 2007 по 2017 роки в розрізі місяців.

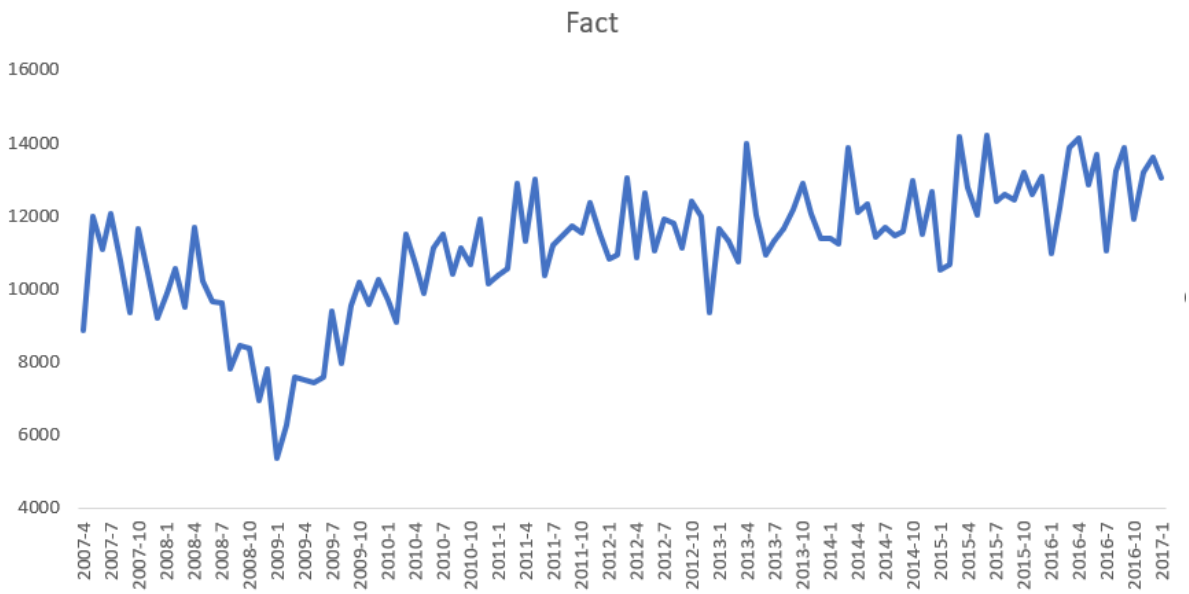


Рис. 2.3 Фактичні дані кількості продажів в Норвегії за період з 2007 по 2017 роки

Якщо ми подивимося на графік продажів, то побачимо, що він має тренд, тобто він формує вищі максимуми та вищі мінімуми. Давайте зафіксуємо тенденцію в нашій моделі (*dY*). Для цього необхідно відняти продажі в момент часу *t* від продажів в момент часу *t-1*.

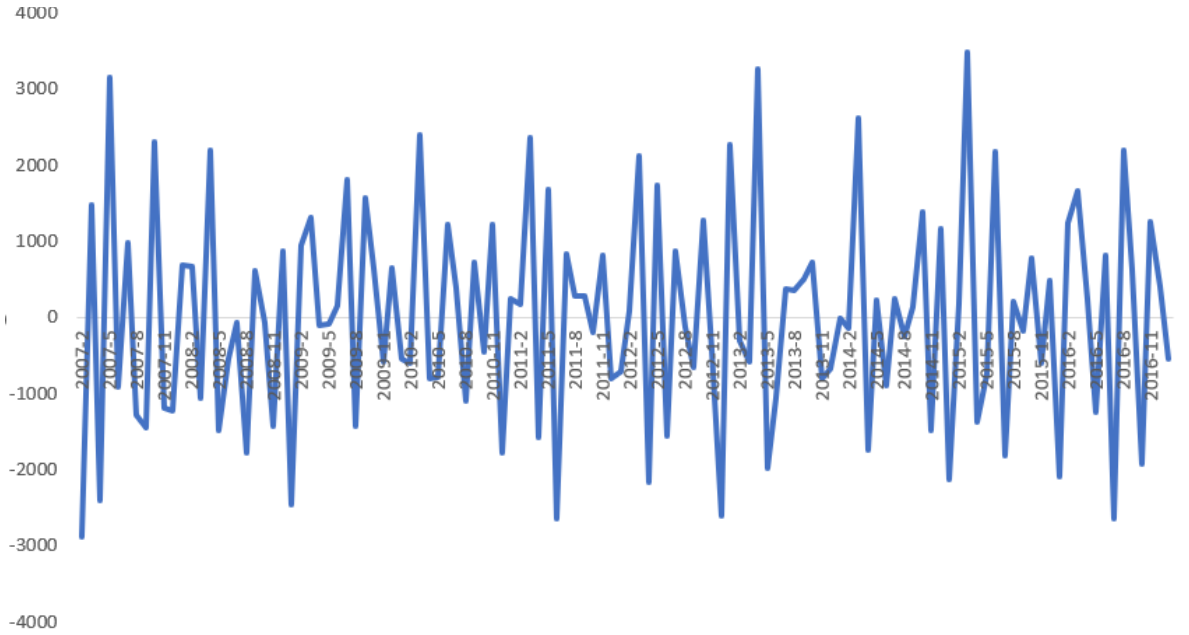


Рис 2.4 Тренд різниці першого порядку

Наступним кроком є Auto Regressive Model (AR-модель). Це метод моделювання часових рядів, що регресує змінну (dY) на її власні минулі значення, щоб виявити сезонність і циклічність в цінах. Цей метод включає в себе використання запізнених (лагованих) значень dY. Наприклад, лаговане значення порядку 1 означає використання значення dY вчорашнім днем разом із сьогоднішнім, а лаговане значення порядку 2 означає використання значення від позавчора разом із сьогоднішнім. Це допомагає в розумінні та моделюванні тимчасових залежностей в даних (рис. 2.5).

Year	Month	Quantity	dY	Lag 1
2007	1	12685		
2007	2	9793	-2892	
2007	3	11264	1471	-2892
2007	4	8854	-2410	1471
2007	5	12007	3153	-2410
2007	6	11083	-924	3153
2007	7	12062	979	-924
2007	8	10786	-1276	979
2007	9	9340	-1446	-1276
2007	10	11646	2306	-1446

Рис 2.5 Приклад лагованого значення порядку 1

На цьому етапі ми проводимо лінійну регресію dY зі значенням запізненого dY. Для спрощення припускається, що порядок запізнення дорівнює 1.

У програмі Excel, для проведення регресії використовуємо пакет Data Analysis з функцією регресії. Результатом цієї команди є набір коєфіцієнтів (рис. 2.6)

На основі коефіцієнтів, які ми отримали в підсумку вище, остаточне рівняння, яке ми отримуємо:

$$dY(t) = 31.8324 - 0.5159\, dY(n-1) \quad (1)$$

Після виконання моделі інтеграції та автоматичної регресії, ми зафіксували тенденцію та сезонність, але в помилці залишаться деякі закономірності. Це означає, що ми побудували модель для певної міри передбачення, але залишаться деякі

закономірності, які ми все ще можемо охопити та які ми охопимо, використовуючи підхід ковзаючого середнього. Щоб змоделювати цей крок, необхідно обчислити прогнозований *dY.*

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.524607718
R Square	0.275213258
Adjusted R Square	0.269018499
Standard Error	1175.275963
Observations	119

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	61365556.88	61365556.88	44.42679376	9.12527E-10
Residual	117	161609009.9	1381273.589		
Total	118	222974566.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	31.83244262	107.7390179	0.291282056	0.771351375	-181.9890366	244.7539218	-181.9890366	244.7539218
X Variable 1	-0.515943813	0.077307174	-6.665342734	9.12527E-10	-0.668381621	-0.362176005	-0.668381621	-0.362176005

Рис. 2.6 Узагальнення регресії авторегресійної моделі (AR)

Щоб обчислити прогнозований *dY*, ми використаємо рівняння 1, яке ми розрахували на попередньому кроці. Обчислити термін помилки в момент часу *t*, тобто *e(t).* Щоб обчислити *e(t),* ми віднімемо прогнозований *dY* від фактичного *dY*.

$$e(t) = \text{Actual } dY(t) - \text{Predicted } dY(t) \quad (2)$$

Коли розрахуємо похибку та виконаємо кроки, які ми зробили для моделі AR. Єдина відмінність полягає в тому, що для моделі AR ми працювали з *dY*, а для цього кроку ми робимо це для стовпця *e(t).* Як обговорювалося вище, для спрощення ми використовуємо лагове значення порядку 1 (рис. 2.7)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.473452914
R Square	0.224157662
Adjusted R Square	0.217526531
Standard Error	1215.966111
Observations	119

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	49981457.59	49981457.59	33.8038351	5.38958E-08
Residual	117	172993109.2	1478573.583		
Total	118	222974566.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.443094102	111.4679856	0.227569492	0.820378135	-195.3897916	246.1232172	-195.3897916	246.1232172
X Variable 1	-0.16629743	0.082752864	5.81410656	5.38958E-08	0.31724626	0.645021682	0.31724626	0.645021682

Рис. 2.7 Узагальнення регресії ковзаючого середнього (MA)

Після регресії e(t) на e(t-1) ми знаходимо, що значення коефіцієнтів. Отже, остаточне рівняння таке:

$$e(t) = 0.4431 - 0.1663 * e(t-1) \quad (3)$$

Результатом нашої моделі буде:

$$Predict\ dY(t) = 31.8324 - 0.5159\, dY(t-n2) + 0.4431 - 0.1663 * e(t-n3)$$

Таблиця 2.4

Похибки авторегресійного інтегрованого методу ковзаючої середньої

|Error|	Error ^2	|% Error|
91,015.73	113,806,268.09	839.03%

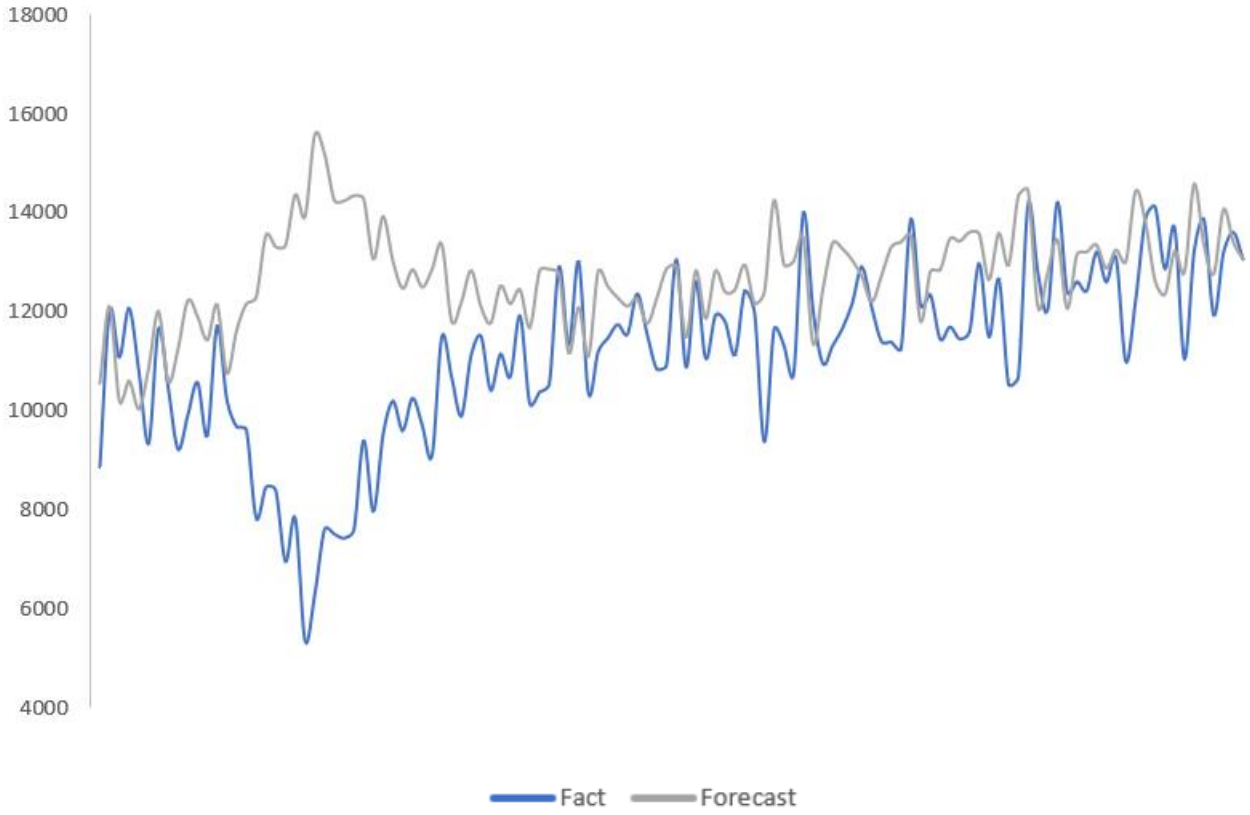


Рис. 2.8 Фактичні та прогнозовані дані кількості продажів в Норвегії за період з 2007 по 2017 роки авторегресійного інтегрованого методу ковзаючої середньої

Для отримання оцінки якості прогнозу, розрахуємо основні похибки:

Таблиця 2.5

Оцінка якості прогнозу авторегресійного інтегрованого методу ковзаючої середньої

MAE	764.84
MSE	956,355.19
RMSE	977.93
MAPE	7.05%

Авторегресійний інтегрований метод ковзаючої середньої більш складний і вимагає більше експертизи, результати досить схожі з попередніми методами.

2.1.4. Сезонна декомпозиція часових рядів

Метод сезонної декомпозиції виділяє сезонність, тренд та ірегулярні компоненти. Корисна для детального аналізу сезонних даних. Не є ефективна для нерегулярних або нетипових часових рядів. Вимагає чітко вираженої сезонності та досить великої кількості даних.

Найпростішим методом сезонної декомпозиції часових рядів є лінійна регресія. Excel надає автоматичні функції лінійної регресії та надає рівняння.

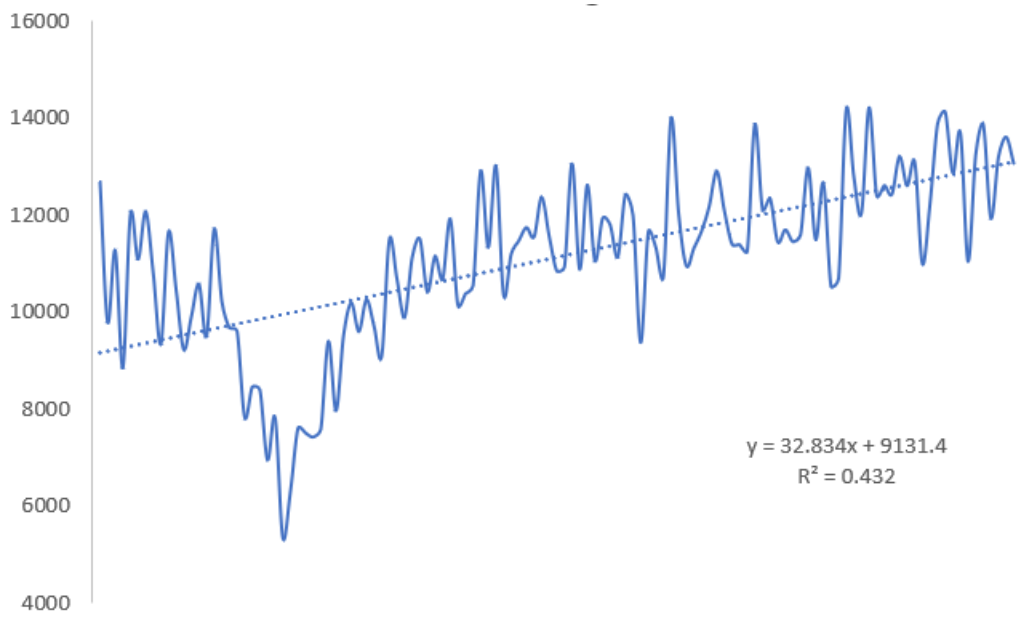


Рис. 2.9 Фактичні дані кількості продажів в Норвегії за період з 2007 по 2017 роки та тренд лінійної регресії

Таблиця 2.6

Похибки методу лінійної регресії

|Error|	Error ^2	|% Error|
116,908.44	209,217,895.24	1175.99%

Для отримання оцінки якості прогнозу, розрахуємо основні похибки:

Таблиця 2.7

Оцінка якості прогнозу методом лінійної регресії

MAE	966.19
MSE	1,729,073.51
RMSE	1,314.94
MAPE	9.72%

Перевагами даного методу є простота та доступність, проте є суттєві недоліки, такі як:

- Передбачення обмежено лінійними зв'язками, що не дає вловити складні нелінійні зв'язки між змінними.
- Вона може бути чутливою до викидів, що впливають на точність прогнозів.
- Лінійна регресія передбачає, що помилки мають нормальний розподіл, що може бути неправдивим для деяких даних.
- Робота з категоріальними змінними вимагає додаткових перетворень.
- Вибір правильних змінних важливий для успішного застосування лінійної регресії.

2.1.5. Декомпозиція часових рядів (TSD)

Щоб отримати реалістичний прогноз, нам потрібно брати до уваги всі компоненти історичних даних, а не лише тенденцію, яку враховує лінійна регресія. Найважливішим, оскільки його можна певною мірою передбачити, є сезонність.

Декомпозиція часових рядів бере за основу тренд лінійної регресії, враховує середній сезонний компонент і накладає їх один на одного [14].

Вихідними даними є тренд, який лінійної регресії, обчислюється за допомогою значень нахилу та перетину. Додаткові дані, які необхідні – ідентифікатор, який визначає сезонність. Оскільки ми робимо місячну сезонність, використовуємо ідентифікатор місяця. Сезонний компонент історичних даних обчислюється простим відніманням лінійного тренду від фактичних даних. Середня сезонність, визначається як середнє значення для кожного ідентифікатору. Прогнозовані дані – це сума середньої сезонності та лінійного тренду. На рис. 2.10 зображені фактичні та прогнозовані дані, що показують якісну тенденцію.

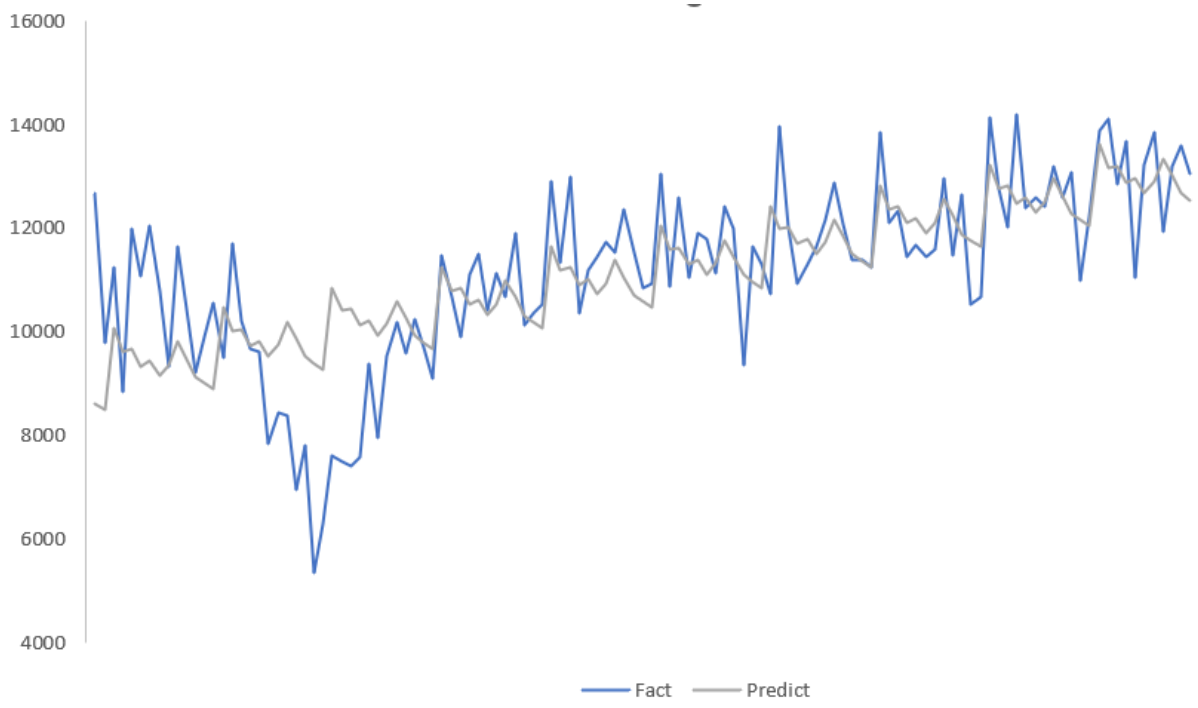


Рис. 2.10 Фактичні та прогнозовані дані кількості продажів в Норвегії за період з 2007 по 2017 роки декомпозиції часових рядів

Таблиця 2.8

Похибки методу декомпозиції часових рядів

|Error|	Error ^2	|% Error|
136,938.43	275,558,677.85	1355.48%

Для отримання оцінки якості прогнозу, розрахуємо основні похибки:

Таблиця 2.8

Оцінка якості прогнозу методом декомпозиції часових рядів

MAE	1,131.72
MSE	2,277,344.45
RMSE	1,509.09
MAPE	11.20%

Декомпозиція часових рядів має переваги і недоліки. Переваги включають можливість виділення тренду та сезонності для кращого розуміння даних, спрощення прогнозування та виявлення аномалій. Проте, цей підхід має обмеження, такі як припущення про стаціонарність та складність визначення періодичності сезонності. Крім того, він може бути менш ефективним для рядів зі складними паттернами.

2.1.6. Машинне навчання з використанням методу «Випадковий ліс»

Метод «Випадковий ліс» (Random Forest) є потужним алгоритмом машинного навчання, який використовується для класифікації та регресії. Він базується на ідеї ансамблю дерев рішень. Метод створює велику кількість дерев із випадковими підвибірками даних і використовує їх для прийняття прогнозів. Важливою перевагою «Випадкового лісу» є його здатність уникнути перенавчання і покращувати точність прогнозів. Цей метод використовується в багатьох галузях, включаючи фінанси, медицину і аналітику даних [15].

Основними бібліотеками для аналізу даних та машинного навчання є pandas, numpy, scikit-learn, joblib, seaborn, matplotlib.

Розглянемо детальніше кожну з них та її функції.

Pandas – це потужна бібліотека для обробки та аналізу даних у Python. Вона надає структури даних, такі як DataFrame і Series, що дозволяють легко завантажувати, зберігати, обробляти і аналізувати дані. Pandas дозволяє виконувати операції з вибором, фільтрацією, групуванням і об'єднанням даних, а також використовувати функції для обчислення статистичних показників. Вона є важливою для аналізу даних, машинного навчання та візуалізації. Pandas робить роботу з даними більш ефективною та зручною в Python.

NumPy (Numerical Python) – це бібліотека для Python, яка надає підтримку для роботи з масивами та матрицями чисел високої якості. Вона є важливою для наукових обчислень, обробки даних та машинного навчання. NumPy забезпечує векторизовані операції, що роблять обчислення ефективнішими, і надає багато математичних функцій для роботи з числами. Ця бібліотека дозволяє створювати, індексувати та змінювати масиви, а також виконувати операції лінійної алгебри, статистики та інші обчислення. NumPy є однією з основних бібліотек у екосистемі Python для обчислювальних завдань [16].

Scikit-learn, також відомий як sklearn, є відкритою бібліотекою машинного навчання для Python. Вона надає широкий спектр інструментів та алгоритмів для задач класифікації, регресії, кластеризації, вимірювання важкості, а також для виконання операцій навчання з учителем і без учителя. Scikit-learn включає в себе підтримку валідації моделей, вибору гіперпараметрів, підготовки даних та візуалізації результатів. Ця бібліотека стала популярною серед дослідників та інженерів машинного навчання завдяки своїй простоті використання та потужним можливостям для розробки моделей на основі даних [17].

Joblib – це бібліотека для Python, яка використовується для ефективного збереження та завантаження об'єктів Python, таких як моделі машинного навчання, безпосередньо на диск. Вона є корисною для збереження навчених моделей і об'єктів, щоб потім легко відновлювати їх для використання без повторного навчання. Joblib

працює швидко і ефективно завдяки використанню механізму серіалізації і кешування, що дозволяє прискорити операції збереження та завантаження великих об'єктів. Ця бібліотека часто використовується в машинному навчанні та аналізі даних для зручності роботи з моделями та даними.

Seaborn – це бібліотека для створення красивих та інформативних статистичних графіків в мові програмування Python. Вона побудована на основі бібліотеки Matplotlib і надає вищий рівень абстракції для створення графіків. Seaborn пропонує широкий спектр графіків, призначених для аналізу даних, включаючи гістограми, ящики з вусами, теплові карти, графіки розподілу, лінійні графіки і багато інших. Вона надає можливості для налаштування вигляду та стилю графіків і допомагає спростити процес візуалізації даних для аналітиків та дослідників [18].

Matplotlib – це бібліотека для створення графіків та візуалізації даних в мові програмування Python. Вона надає різноманітні можливості для створення різних типів графіків, таких як лінійні графіки, гістограми, діаграми розсіювання та інші. Matplotlib дає повний контроль над виглядом та стилем графіків і дозволяє налаштовувати їх до потреб користувача. Вона є однією з основних бібліотек для візуалізації даних в науковому та аналітичному програмуванні з використанням Python [19].

Було використано код з відкритих джерел [20] та проведено аналіз та прогноз даних продажів.

Першим кроком за допомогою бібліотеки Matplotlib було проаналізовано історичні дані (рис. 2.11)

```
plt.figure(figsize=(10, 3))
years=list(data_eda['Year'].unique())
for i in years:
plt.plot(data_eda['Month'][data_eda['Year']==i],
data_eda['Quantity'][data_eda['Year']==i], label=i, linewidth=1)
```

```
fmt = '{x:,.0f}'
ytick = mtick.StrMethodFormatter(fmt)
plt.gca().yaxis.set_major_formatter(ytick)
plt.xticks(range(1,13,1),['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec'])
plt.legend(ncol=1, loc='upper right',facecolor='white',bbox_to_anchor=(1.2, 1))
plt.gca().set_facecolor("white")
plt.title("Quantity per month over years")
plt.xlabel("Month")
plt.ylabel("Quantity")
plt.grid(axis = 'x',color='lightgrey', linestyle='--', linewidth=0.5)
plt.grid(axis = 'y',color='lightgrey', linestyle='--', linewidth=0.5)
```

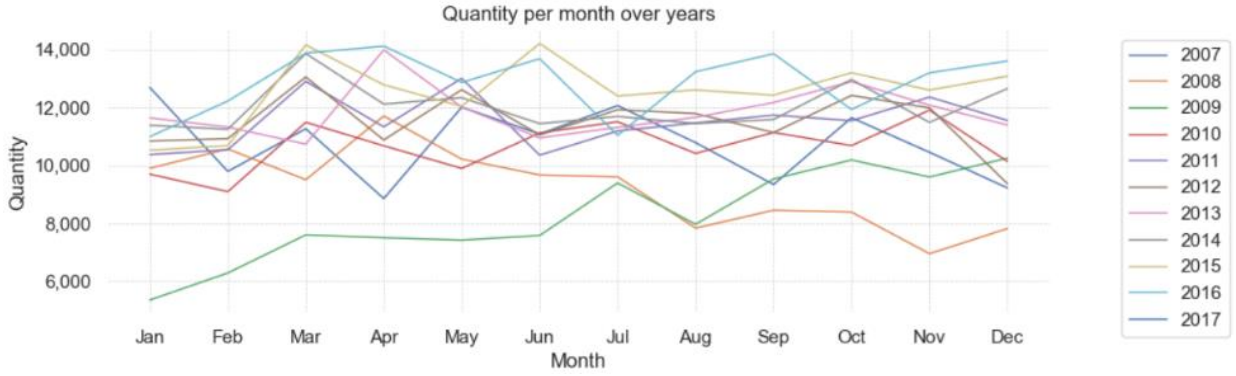


Рис. 2.11 Фактичні дані кількості продажів в Норвегії за період з 2007 по 2017

В нашому випадку випадковий ліс використаний для розв'язання задач регресії. У випадку регресії, завдання полягає в прогнозуванні числового значення вихідної змінної на основі інших ознак. Випадковий ліс будує декілька дерев рішень, які розглядаються як ансамбль. Кожне дерево робить прогноз регресії, і кінцевий результат обчислюється як середнє або медіанне значення прогнозів всіх дерев. Це дозволяє використовувати випадковий ліс для прогнозування числових значень.

Для чистоти експерименту ми використовували лише дані продажів. RMSE використовувався як метрика для вимірювання точності моделі регресії.За

найкращими параметрами моделі RandomForestRegressor ми отримали RMSE = 2342.85 (рис. 2.12) на тестовому наборі даних з наступними показниками:

- середня абсолютна похибка на тестових даних – 2141
- середня відносна похибка – 16%.

Після видалення всіх даних до 2010 року з навчального набору даних (через значний спад продажів в 2008-2009 роках) ми досягли трохи вищої точності:

- середній абсолютний похибка на тестових даних – 1529
- середня відносна похибка – 11.5%.

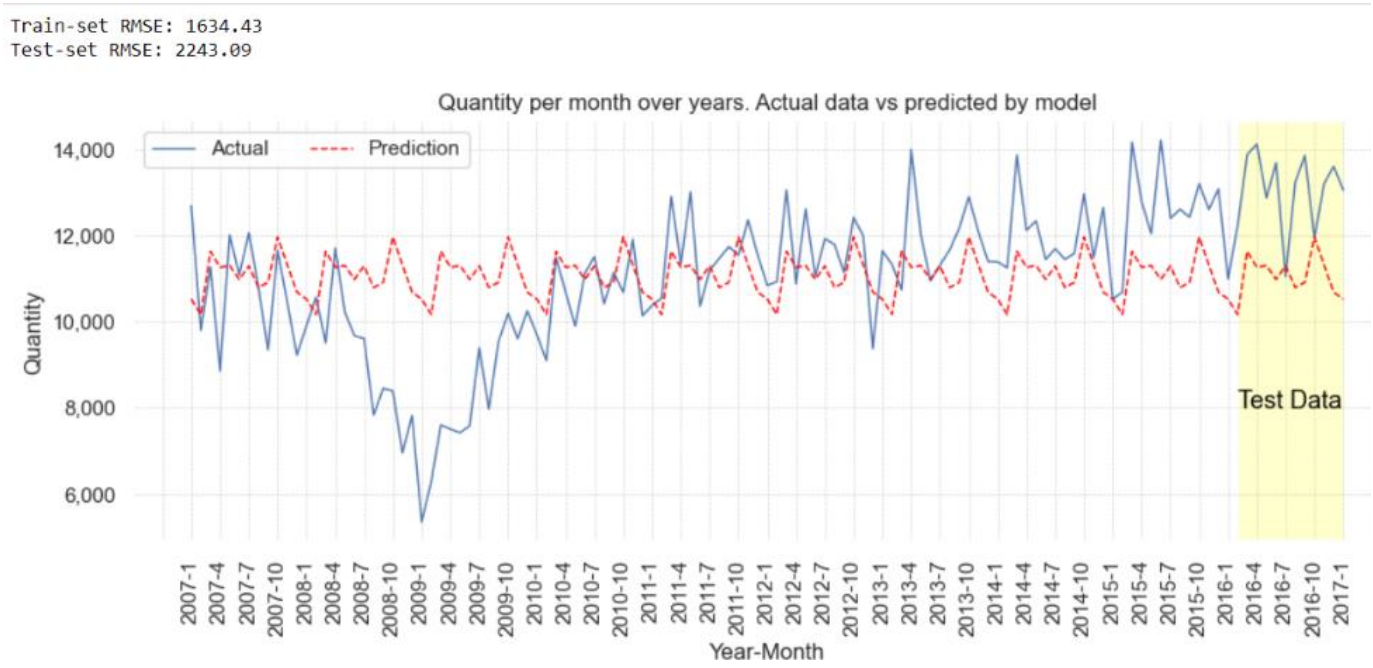


Рис. 2.12 Фактичні та прогнозовані дані кількості продажів в Норвегії за період з 2007 по 2017 роки за допомогою «випадкого лісу».

Незважаючи на багато переваг, у випадкового лісу є деякі недоліки. По-перше, він може стати перенавчанням на великих наборах даних, що призводить до низької узагальнюючої здатності. По-друге, він не дуже ефективний в умовах великої кількості рідкісних класів у задачах класифікації. Крім того, випадковий ліс не надає

інтерпретації результатів, що робить його менш відповідним для деяких додатків, де необхідно пояснити, як саме модель зробила прогнози.

2.2. Вибір та детальний опис обраного методу та його характеристик

Для вибору найперспективнішого методу ми скористаємося методом експертних оцінок, на основі зібраних даних та результатів прогнозів.

Метод експертних оцінок – це підхід до прогнозування, де експерти в галузі надають свої прогнози та оцінки на основі свого досвіду та знань. Цей метод застосовується в ситуаціях, коли немає достатньої кількості інформації для використання інших статистичних методів. Експерти надають свої оцінки, і їхні думки агрегуються, наприклад, за допомогою середньої або медіанної оцінки, щоб отримати прогноз. Цей метод може бути корисним в ситуаціях, де інші методи не працюють ефективно, але він також може бути суб'єктивним і піддатливим до впливу більшості або недостатньо обгрунтованих оцінок експертів [21].

Основними критеріями аналізу будуть:

1. Використання таких метрик, як середня абсолютна відсоткова помилка (MAPE), коренева середньоквадратична помилка (RMSE) для оцінки точності прогнозів.
2. Необхідність специфічних знань та ресурсів.
3. Необхідність додаткової інфраструктури.
4. Обмеження в даних (неможливість використання категоріальних даних, необхідність великих масивів даних в перерахунку від 0 до 5, де 5 – немає взагалі обмежень та 0 – суцільні обмеження).
5. Складність підтримки.

Аналіз різноманітних методів та детальний структурований опис щодо оцінки кожного методу на підставі конкретних критеріїв, які наведено в таблиці 2.9. Інформація та критерії, що стосується кожного методу дозволяють провести ретельну оцінку їх ефективності за численними критеріями.

Таблиця 2.9

Зведені дані результатів аналізу кожного описаного методу прогнозування

Метод / Критерії	MAPE	RMSE	Специфічні знання	Додаткова інфраструктура	Категоріальні дані	Обмеження кількістю даних	Складність підримки
Метод ковзаючої середньої	9%	1,209	Не потребує	Не потребує	Не підтримує	Лише обмеження інструменту реалізації	Низький рівень
Експоненційне згладжування	8%	1,118	Потребує (низький рівень)	Не потребує	Не підтримує	Лише обмеження інструменту реалізації	Низький рівень
ARIMA (Авторегресійний інтегрований метод ковзаючої середньої)	7%	977	Потребує (середній рівень)	Не потребує (можливе використвння програмування)	Не підтримує	Лише обмеження інструменту реалізації	Середній рівень
Лінійна регресія	10%	1,315	Не потребує	Не потребує	Не підтримує	Лише обмеження інструменту реалізації	Низький рівень
Декомпозиція часових рядів (TSD)	11%	1,509	Потребує (низький рівень)	Не потребує	Не підтримує	Лише обмеження інструменту реалізації	Низький рівень
«Випадковий ліс»	16%	2,343	Потребує (високий рівень)	Потребує	Підтримує	Лише обмеження інструменту реалізації (перенавчання)	Високий рівень

Згідно з таблицею 2.9, ми можемо перетворити ці категоріальні дані в метод експертних оцінок, використовуючи оцінки від 0 до 5, щоб визначити найбільш перспективний метод (табл. 2.10).

Таблиця 2.10

Визначення найперспективнішого методу шляхом експертних оцінок

Метод / Критерії	MAPE	RMSE	Специфічні знання	Додаткова інфраструктура	Категоріальні дані	Обмеження кількістю даних	Складність підримки	Результат
Метод ковзаючої середньої	3	3	5	5	0	5	5	26
Експоненційне згладжування	4	4	5	5	0	5	5	28
ARIMA (Авторегресійний інтегрований метод ковзаючої середньої)	5	5	4	4	0	5	4	27
Лінійна регресія	3	3	5	5	0	5	5	26
Декомпозиція часових рядів (TSD)	2	2	5	5	0	5	5	24
«Випадковий ліс»	1	1	2	3	5	4	2	18

На підставі аналізу та експертних оцінок, було вирішвибрати метод експоненційного згладжування. Цей метод виявився найбільш підходящим і належним чином оціненим у відповідності до встановлених критеріїв.

Експоненційне згладжування – це метод прогнозування часових рядів, який використовує експоненційно зменшення вагів для минулих спостережень. Для покращення цього методу можна розглянути наступні кроки:

- Налаштування параметрів згладжування (альфа, бета, гамма) для знаходження найкращого підходу до даних.
- Додавання сезонності або тенденцій для виявлення складніших залежностей.
- Обробка викидів або аномалій, які можуть вплинути на точність прогнозу.
- Дослідження більш складних варіацій, таких як модель Хольта-Вінтерса чи моделі на основі простору станів, для отримання кращих результатів.

Оптимізувати якість прогнозу за допомогою експотенційного згладжування ми спробуємо методом Хольта-Вінтерса.

Метод Хольта-Вінтерса – це метод експоненційного згладжування для прогнозування часових рядів, який враховує як тренд, так і сезонність в даних. Він включає в себе трендовий згладжувач, сезонний згладжувач та згладжувач для виправлення невипадкових коливань. Даний метод дозволяє прогнозувати дані з високою точністю, враховуючи їхню динаміку та сезонні коливання. Для налаштування параметрів цього методу можна використовувати оптимізаційні підходи та метрики оцінки точності прогнозів [22].

Основні кроки:

- Початкові значення тренду (T_0), рівня (L_0) та сезонного компонента (S_0) визначаються на підставі початкових даних.

- Оновлення тренду, рівня та сезонного компонента проводиться згідно зі спеціальними формулами, які враховують актуальні дані та параметри згладжування.
- На основі оновлених значень розраховуються прогнози на майбутні періоди.
- Модель регулярно оновлюється з кожним новим значенням часового ряду.

РОЗДІЛ 3: РОЗРОБКА ТА АНАЛІЗ МОДЕЛІ МЕТОДОМ ХОЛЬТА-ВІНТЕРСА

3.1 Розробка моделі: Створення алгоритму на основі методу Холта-Вінтерса

Згладжування Холта-Вінтерса також називають потрійним експоненційним згладжуванням, за рахунок трьох параметрів, що згладжують (альфа, гама і сезонний фактор – дельта). Наприклад, якщо є 12-місячний сезонний цикл:

$$Yt = (l_{t-3} + 3 * Tr_{t-3}) * S_{t-12},$$

де l_{t-3} – рівень за період *t-3*, Tr_{t-3} – тренд за період *t-3*, S_{t-12} – сезонність за період *t-12*.

Аналізуючи дані, необхідно з'ясувати, що в серії даних є трендом, а що сезонністю. Щоб виконати обчислення за методом Холта-Вінтерса, необхідно:

- Згладити історичні дані методом ковзаючого середнього.
- Порівняти згладжену версію часового ряду даних із оригіналом, щоб отримати приблизну оцінку сезонності.
- Отримати нові дані без сезонного компонента.
- Знайти наближення рівня та тренду на основі цих нових даних.

Метод ковзаючого середнього використовується для згладжування історичних даних. Він полягає в обчисленні середнього значення даних за певний період, який «ковзає» по часовому ряду. Це допомагає виділити загальний тренд та відсіяти шуми, полегшуючи аналіз та прогнозування. Результат згладжування наведено на рисунку 3.1.

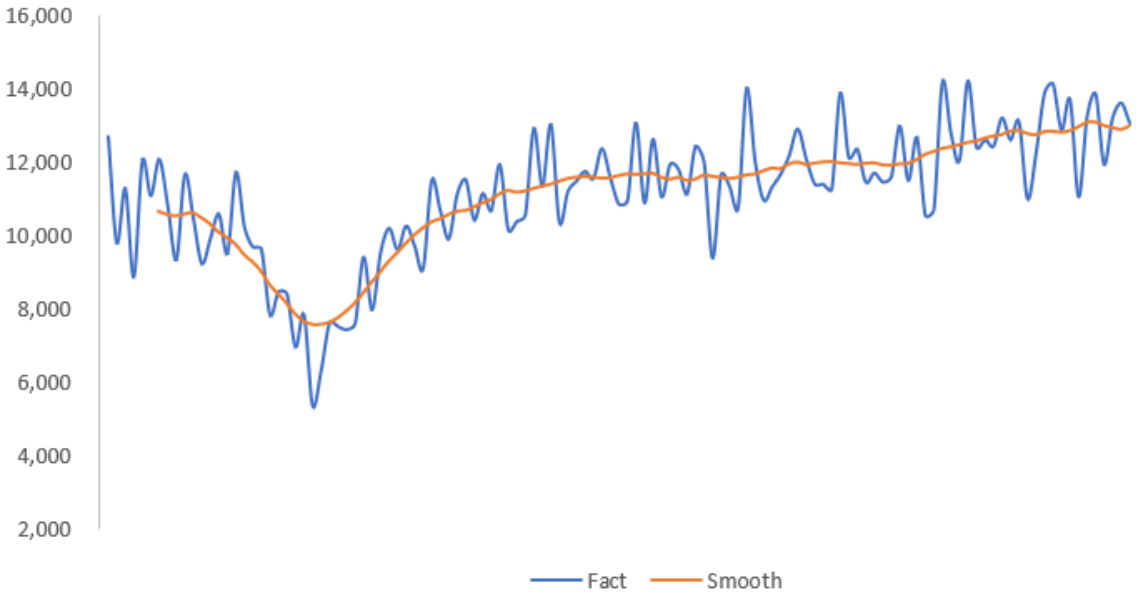


Рис. 3.1 Фактичні та згладжені дані попиту

Рівень сезонних коливань можна відобразити графічно (рис. 3.2).

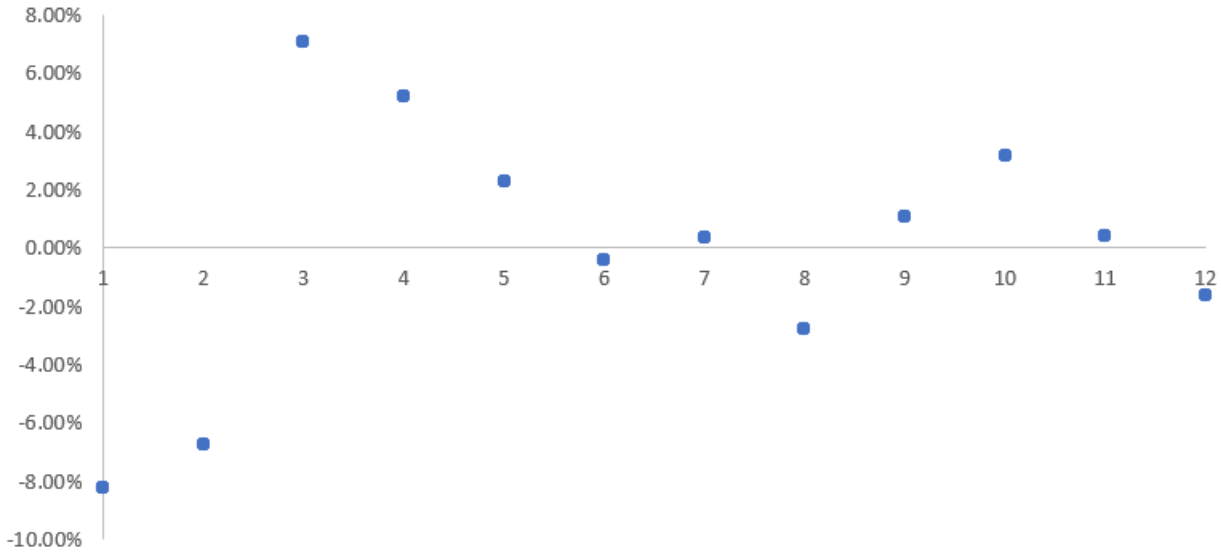


Рис. 3.2 Сезонні коливання

Тепер можна отримати дані, скориговані сезонними коливаннями. На основі отриманих даних за допомогою лінійної регресії будуємо графік та доповнюємо його лінією тренда. Коєфіцієнти рівняння лінії тренду є основою для розрахунків методу Холта-Вінтерса (рис. 3.3).

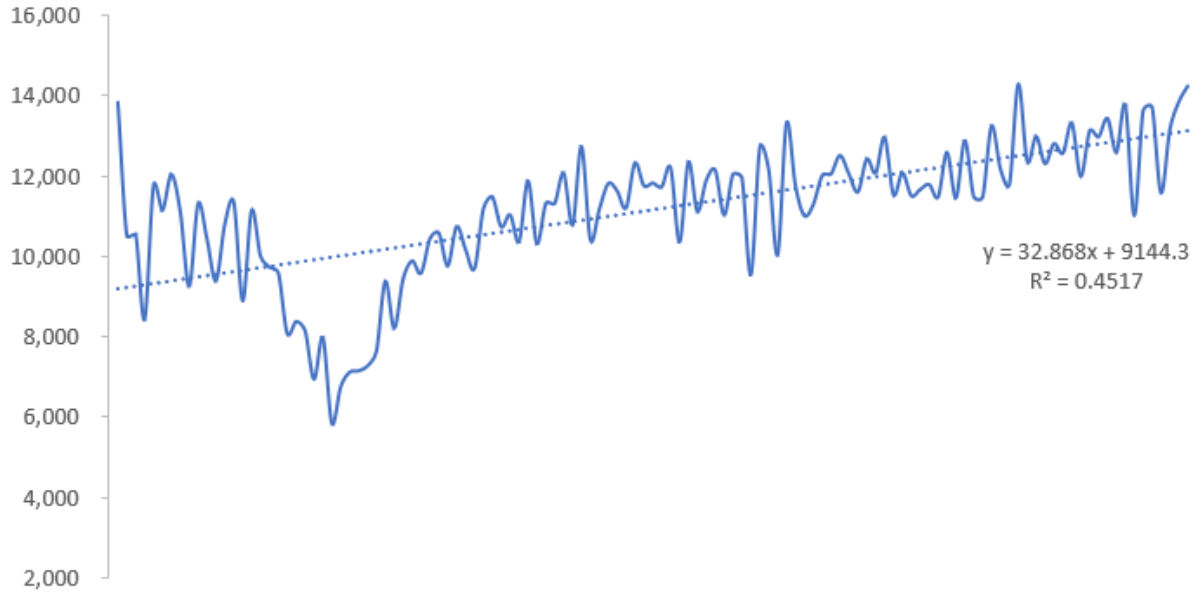


Рис. 3.3 Скориговані сезонними коливаннями дані та лінія тренду

Однокроковий прогноз методом Холта-Вінтерса базується на основі констант зладжування, тренду, сезонності та рівня, тренду та коєфіцієнту сезонності.

Константи були обрані на рівні 0.5.

Однокроковий прогноз розраховується за формулою:

$$Pr_{t1} = (L_{t0} + Tr_{t0}) * S_{AVG1},$$

де L_{t0} – рівень для нульового періоду, Tr_{t0} – тренд для нульового періоду, S_{AVG1} – середня сезонність відповідного періоду.

Рівень для нульового періоду визначається за допомогою рівняння лінії тренду (9144.3), кожний наступний розраховується за формулою:

$$L_{t1} = L_{t0} + Tr_{t0} + \alpha * \frac{E}{S_{AVG1}},$$

де L_{t0} – рівень для нульового періоду, Tr_{t0} – тренд для нульового періоду, α – константа згладжування (0.5), E – помилка прогнозу (різниця факту та прогнозу), S_{AVG1} – середня сезонність відповідного періоду.

Тренд для нульового періоду визначається за допомогою рівняння лінії тренду (32.868), кожней наступний розраховується за формулою:

$$Tr_{t1} = Tr_{t0} + \alpha * \beta * \frac{E}{S_{AVG1}},$$

де Tr_{t0} – тренд для нульового періоду, α – константа згладжування (0.5), β – константа згладжування тренду (0.5), E – помилка прогнозу (різниця факту та прогнозу), S_{AVG1} – середня сезонність відповідного періоду.

Середня сезонність для початкового періоду була розрахована на першому етапі, а наступні сезонності розраховується за формулою:

$$S_{AVG1} = S_{AVG0} * \gamma * (1 - \alpha) * \frac{E}{Tr_{t0} + L_{t0}},$$

де S_{AVG0} – середня сезонність для початкового періоду, L_{t0} – рівень для нульового періоду, Tr_{t0} – тренд для нульового періоду, α – константа згладжування (0.5), γ – константа згладжування сезонності, E – помилка прогнозу (різниця факту та прогнозу).

3.2 Аналіз моделі методом Хольта-Вінтерса

Перспективи покращення прогнозу методом експоненційного згладжування Холта-Вінтерса включають оптимізацію параметрів моделі для кращого урахування тренду та сезонності. Автоматизовані методи підбору параметрів можуть сприяти точнішим прогнозам. Також, комбінування методу з іншими алгоритмами машинного навчання допоможе отримати кращі результати, особливо в складних ситуаціях.

Модель Холта-Вінтерса підвищила точність прогнозу і дозволила ефективно прогнозувати на значні періоди в майбутньому (рис. 3.4).

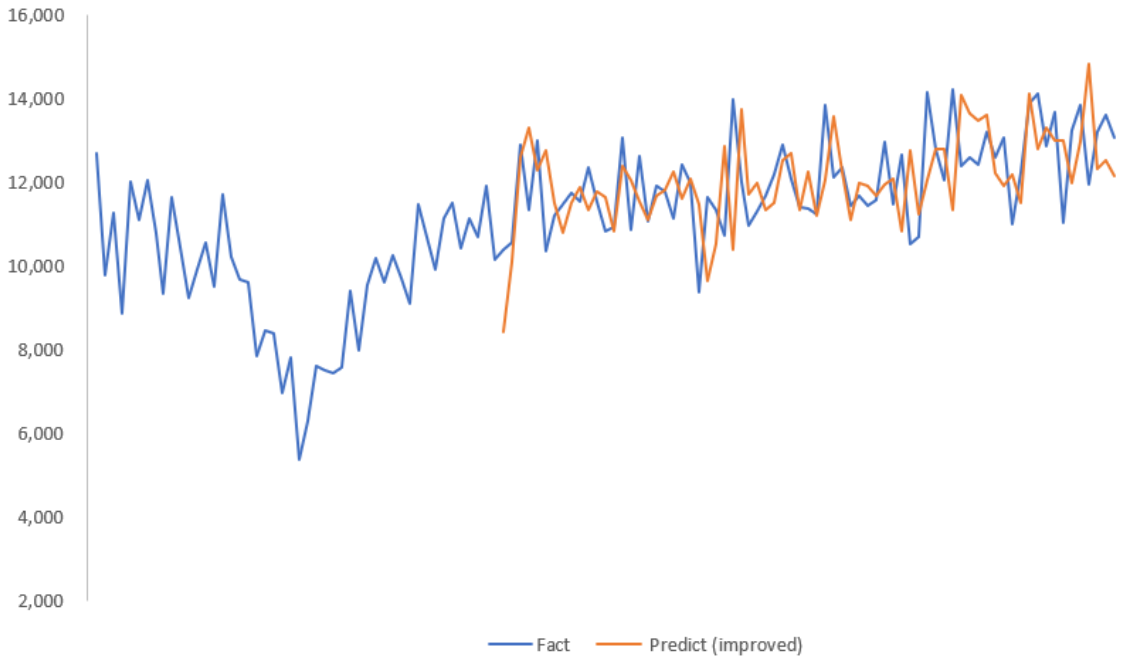


Рис. 3.4 Фактичні та прогнозовані дані кількості продажів в Норвегії за період з 2007 по 2017 роки методом Холта-Вінтерса

Таблиця 3.1

Похибки методу декомпозиції часових рядів

|Error|	Error ^2	|% Error|
63,371.13	54,926,569.48	597.32%

Для отримання оцінки якості прогнозу, розрахуємо основні похибки:

Таблиця 3.2

Оцінка якості прогнозу методом декомпозиції часових рядів

MAE	523.73
MSE	453,938.59
RMSE	673.75
MAPE	4.94%

Оцінка точності прогнозів моделі Холта-Вінтерса показала, що вона діє досить ефективно, особливо для короткострокових прогнозів. Модель дозволяє прогнозувати значення часового ряду на певний часовий горизонт і враховувати як тренд, так і сезонні коливання. Однак, для досягнення ще більшої точності можна вдосконалити параметри моделі, підбираючи їх оптимальні значення. Також, комбінування моделі Холта-Вінтерса з іншими методами, такими як алгоритми машинного навчання, може покращити точність прогнозів, особливо в умовах збільшеної складності та невпевненості.

ВИСНОВКИ

Прогнозування в сфері FMCG є важливим аспектом для планування запасів, виробництва та маркетингу. У нашому дослідженні ми ретельно проаналізували різні методи прогнозування, включаючи експоненційне згладжування, модель Холта-Вінтерса та методи машинного навчання. Кожен з цих методів має свої переваги та недоліки. Експоненційне згладжування відзначається простотою та добре підходить для стабільних часових рядів.

В свою чергою, методи машинного навчання, такі як випадковий ліс, можуть враховувати більш складні взаємозв'язки в даних, але вимагають більшого обсягу даних та обчислювальних ресурсів. Вибір методу прогнозування повинен залежати від конкретних умов і завдань компанії в сфері FMCG.

Модель Холта-Вінтерса є потужним інструментом для прогнозування в бізнесі з багатьма перевагами. Вона дозволяє враховувати тренд, сезонність та згладжувати шуми в часовому ряді, що робить її ефективною для передбачення змін у продажах і попиті. Ця модель також може адаптуватися до змін у даних, що дозволяє підтримувати актуальні прогнози в реальному часі. Її імплементація допомагає оптимізувати запаси, знизити витрати та підвищити задоволеність клієнтів завдяки точнішому управлінню запасами та постачаннями. Модель Холта-Вінтерса є незамінним інструментом для розвинених бізнес-процесів, які вимагають точних та надійних прогнозів.

В даній роботі ми досягли підвищення точності прогнозування. Від початкових значень MAPE на рівні 20% за допомогою методу ARIMA, ми досягли мінімальних MAPE на рівні 8% використовуючи метод експоненційного згладжування. Було вдосконалено результати, застосовуючи метод Холта-Вінтерса, де MAPE скоротилася

до 5%. Це свідчить про успішність застосування комбінації цих методів для отримання точних прогнозів.

Можливості для подальших розробок та досліджень у цій галузі безмежні. Для початку, можна дослідити можливість вдосконалення самої моделі Холта-Вінтерса шляхом розгляду інших функцій тренду та сезонності, що дозволить покращити точність прогнозів. Також варто дослідити вплив різних змінних на точність моделі та визначити оптимальні параметри для конкретних бізнес-сценаріїв. Крім того, можна розглянути можливість імплементації додаткових методів машинного навчання для покращення прогнозування в умовах складних даних. Такі дослідження можуть призвести до створення більш точних та ефективних інструментів для управління бізнесом.

Важливим аспектом подальшого покращення моделі – є залучення категоріальних даних. Можливість дослідити ефективність різних методів кодування категоріальних змінних та їх вплив на точність прогнозів. Також важливо вивчити можливості використання ансамблей моделей, які об'єднують прогнози з різних категоріальних даних, для отримання більш точних результатів. Крім того, дослідження з урахуванням динаміки змін в категоріальних даних може сприяти розробці адаптивних моделей, що аналізують зміни в категоріях та адаптуються до них для кращого прогнозування.

ПЕРЕЛІК ПОСИЛАНЬ

1. 7 Benefits of Accurate Sales Forecasting [Електронний ресурс] – Режим доступу до ресурсу: https://www.intangent.com/blog/7-benefits-of-accurate-sales-forecasting.
2. 7 Benefits of Accurate Sales Forecasting [Електронний ресурс] – Режим доступу до ресурсу: https://www.intangent.com/blog/7-benefits-of-accurate-sales-forecasting.
3. Forecasting: Principles and Practice – Monash University in Australia: OTexts, 2013. – 298 с. – (1).
4. Sales Forecasting Management: A Demand Management Approach – Washington DC: SAGE Publications, Inc, 2005. – 368 с. – (2).
5. J Hyndman R. Forecasting: Principles and Practice / R. J Hyndman, G. Athanasopoulos. – Monash University in Australia: OTexts, 2013. – 298 с. – (1).
6. Forecasting with Exponential Smoothing: The State Space Approach / R.Hyndman, A. Koehler, K. Ord, R. Snyder. – Berlin: Springer, 2008. – 362 с.
7. E.P. Box G. Time Series Analysis: Forecasting and Control / G. E.P. Box, G. M. Jenkins, G. C. Reinsel. – Hoboken: John Wiley and Sons Inc., 2015. – 712 с.
8. J. Brockwell P. Introduction to Time Series and Forecasting / P. J. Brockwell, R. A. Davis. – Berlin: Springer, 2016. – 450 с.
9. Breiman L. Random Forests / L. Breiman, A. Cutler. // Kluwer Academic Publishers. – 2001. – №45.
10. Azure AI [Електронний ресурс]. – 2024. – Режим доступу до ресурсу: https://azure.microsoft.com/en-us/solutions/ai.
11. C. Müller A. Introduction to Machine Learning with Python / A. C. Müller, S. Guido. – Sebastopol: O\'Reilly Media, Inc., 2016. – 398 с.

12. Samonas M. Financial Forecasting, Analysis, and Modelling: A Framework for Long-Term Forecasting / Michael Samonas. – Chichester: John Wiley & Sons, 2015. – 310 с.
13. Shumway R. Time Series Analysis and Its Applications: With R Examples / Robert H. Shumway. – Beijing: World Publishing Corporation, 2009. – 606 с.
14. S. Stoffer D. Time Series Analysis and Its Applications / David S. Stoffer. – Berlin: Springer, 2017. – 575 с.
15. Breiman L. Random Forests / L. Breiman, A. Cutler. // Kluwer Academic Publishers. – 2001. – №45.
16. Petrou T. Pandas Cookbook: Recipes for Scientific Computing, Time Series Analysis and Data Visualization using Python / Theodore Petrou. – Birmingham: Packt Publishing, 2017. – 510 с.
17. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow / Aurélien Géron. – Sebastopol: O\'Reilly Media, Inc., 2019. – 856 с.
18. McKinney W. Python for Data Analysis / Wes McKinney. – Sebastopol: O\'Reilly Media, 2022. – 550 с.
19. C. Müller A. Introduction to Machine Learning with Python / A. C. Müller, S. Guido. – Sebastopol: O\'Reilly Media, Inc, 2016. – 550 с.
20. Open source code [Електронний ресурс] – Режим доступу до ресурсу: https://www.linkedin.com/posts/lsoldatenko_crossfit-games-open-workouts-analysis-activity-90Xv?utm_source=share&utm_medium=member_ios.
21. Євсійович Б. Г. Методи експертних оцінок: теорія, методологія, напрямки використання / Борис Грабовецький Євсійович. – Вінниця: ВНТУ, 2010. – 171 с.
22. J Hyndman R. Holt-Winters' seasonal method [Електронний ресурс] / R. J Hyndman, G. Athanasopoulos – Режим доступу до ресурсу: https://otexts.com/fpp2/holt-winters.html.

ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ (Презентація)

МЕТОДИ ТА ЗАСОБИ ПІДВИЩЕННЯ ТОЧНОСТІ ПРОГНОЗУВАННЯ СЕЗОННИХ ПРОДАЖІВ У СФЕРІ FMCG

Виконала: здобувачка вищої освіти гр. САДМ-61
Марина Цибенко
Керівник: Сергій Козаченко
к.е.н., доцент кафедри СА

Товари швидкого обігу (FMCG, англ. Fast-Moving Consumer Goods)

Продукти, які продаються швидко і за порівняно низькою ціною. До них відносять продукти харчування, напої, косметику, засоби особистої гігієни, дрібні побутові товари і інші речі повсякденного вжитку. Ці товари мають короткий термін придатності, високу частоту покупки і велику оборотність.

Важливість прогнозів

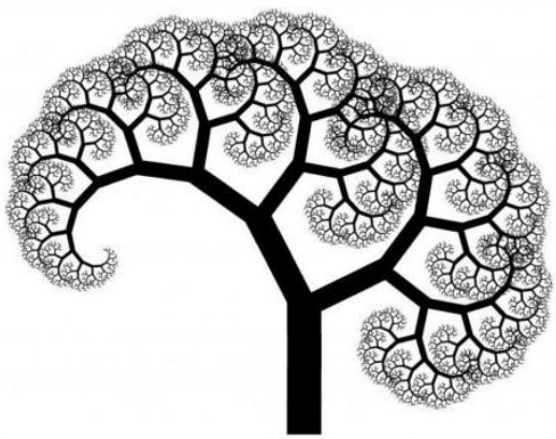
- Оптимізація запасів
- Планування виробництва
- Фінансове планування
- Маркетингова стратегія
- Відповідність попиту споживачів

Традиційні методи прогнозування

Метод ковзаючої середньої	Експоненційне згладжування	ARIMA (Autoregressive Integrated Moving Average)	Сезонна декомпозиція часових рядів
Це техніка прогнозування, в якій для кожного прогнозуваного періоду обчислюється середнє значення попередніх спостережень на фіксованому проміжку часу	Це метод прогнозування, в якому ваги надаються попереднім значенням часового ряду з експоненційною зміною ваги	Метод включає авторегресійну (AR) компоненту для моделювання залежності від попередніх значень, інтегровану (I) частину для стабілізації ряду, і компоненту ковзаючої середньої (MA) для моделювання шуму.	Процес розділення часового ряду на кілька складових: тренд, сезонність, і випадкові відхилення. Це дозволяє аналізувати окремі компоненти серії та краще розуміти її поведінку.

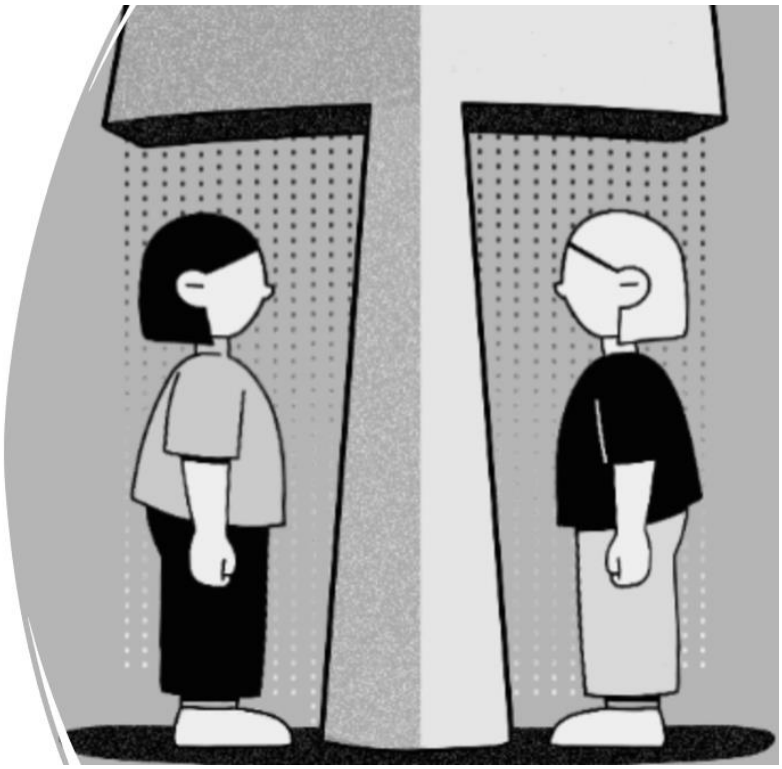
Сучасні підходи

- Сучасні підходи до прогнозування включають використання машинного навчання, глибокого навчання та штучного інтелекту для аналізу та прогнозування складних даних.
- Машинне навчання з використанням методу «Випадковий ліс» (Random Forest) – це потужний алгоритм, який використовується для класифікації та регресії. Він працює шляхом створення великої кількості дерев рішень, кожне з яких навчається на випадковій підмножині даних.



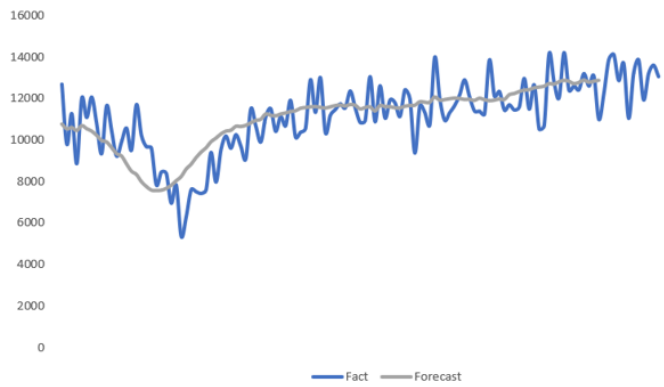
Набір даних та деталі дослідження

- Кількість продажів в Норвегії в локальному рітейлі за період з 2007 по 2017 роки
- Тест Стьюдента (t-тест): p-value менше ніж 0,05 (3.16E-164)



Метод ковзаючої середньої

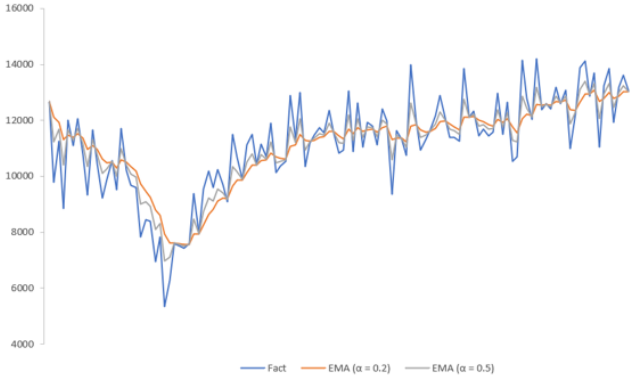
- Метод ковзаючого середнього не потребує конкретних навичок чи додаткової інфраструктури, він також достатньо безпечний та не потребує багато часу на імплементацію, але він не є достатньо точним.
- MAPE = 9.16%



Експоненційне згладжування

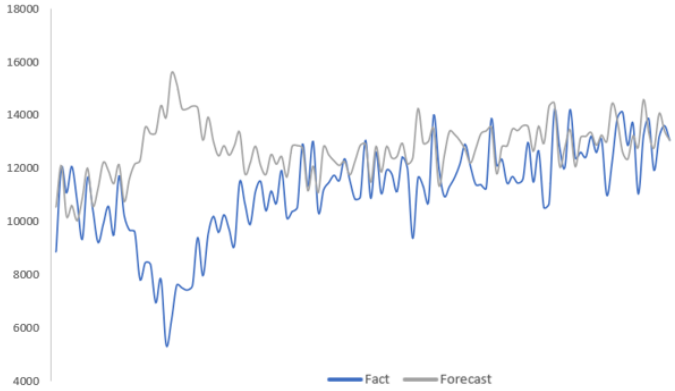
- У методі експоненційного згладжування використовується коефіцієнт згладжування (α), що варіюється від 0 до 1. Вищий коефіцієнт згладжування означає, що більше ваги надається останнім спостереженням.
- Метод експоненційного згладжування допомагає ефективно реагувати на зміни у даних, але може бути менш точним при виражених трендах або сильній сезонності.

	α = 0.2	α = 0.5
MAPE	9.11%	8.10%



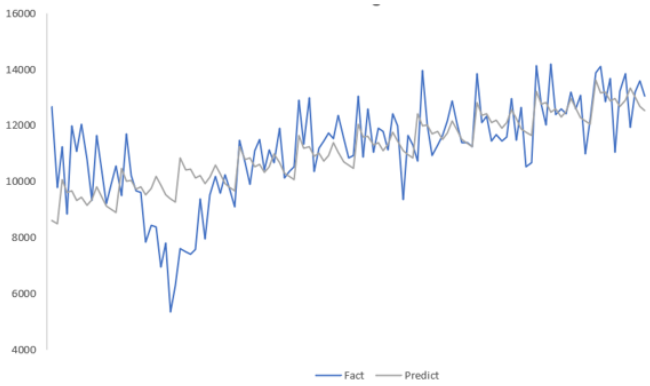
ARIMA (Авторегресійний інтегрований метод ковзаючої середньої)

- Авторегресійний інтегрований метод ковзаючої середньої більш складний і вимагає більше експертизи, результати досить схожі з попередніми методами.
- MAPE = 7.05%



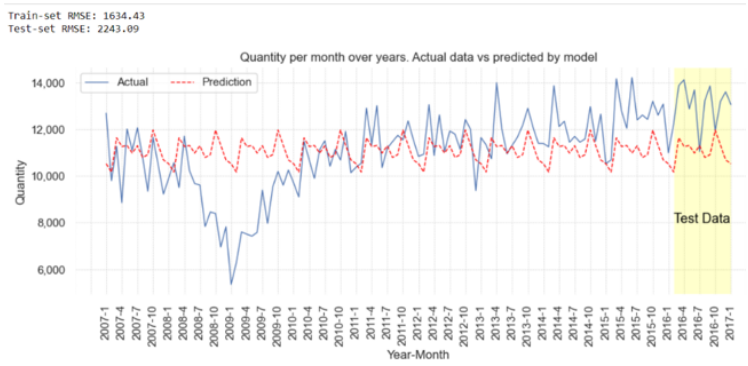
Декомпозиція часових рядів (TSD)

- Переваги включають можливість виділення тренду та сезонності для кращого розуміння даних, спрощення прогнозування та виявлення аномалій. Проте, цей підхід має обмеження, такі як припущення про стаціонарність та складність визначення періодичності сезонності. Крім того, він може бути менш ефективним для рядів зі складними паттернами
- MAPE = 11.20%



Машинне навчання з використанням методу «Випадковий ліс»

- Незважаючи на багато переваг, у випадкового лісу є деякі недоліки. По-перше, він може стати перенавчанням на великих наборах даних, що призводить до низької узагальнюючої здатності. По-друге, він не дуже ефективний в умовах великої кількості рідкісних класів у задачах класифікації. Крім того, випадковий ліс не надає інтерпретації результатів, що робить його менш відповідним для деяких додатків, де необхідно пояснити, як саме модель зробила прогнози.
- MAPE = 16.00%



Визначення найперспективнішого методу шляхом експертних оцінок

Метод / Критерії	MAPE	RMSE	Специфічні знання	Додаткова інфраструктура	Категоріальні дані	Обмеження кількістю даних	Складність підримки	Результат
Метод ковзаючої середньої	3	3	5	5	0	5	5	26
Експоненційне згладжування	4	4	5	5	0	5	5	28
ARIMA (Авторегресійний інтегрований метод ковзаючої середньої)	5	5	4	4	0	5	4	27
Лінійна регресія	3	3	5	5	0	5	5	26
Декомпозиція часових рядів (TSD)	2	2	5	5	0	5	5	24
«Випадковий ліс»	1	1	2	3	5	4	2	18

Метод Хольта-Вінтерса

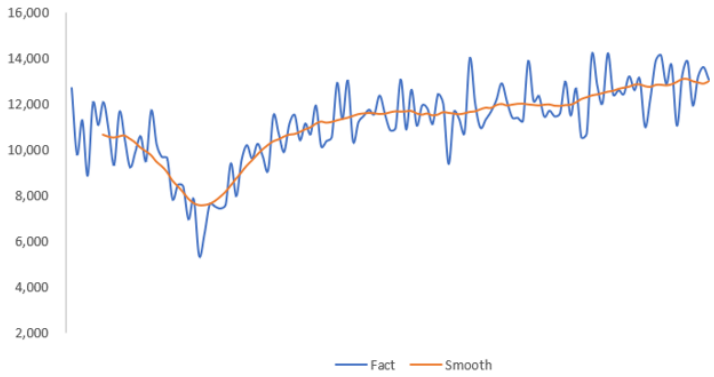
- Метод експоненційного згладжування для прогнозування часових рядів, який враховує як тренд, так і сезонність в даних. Він включає в себе трендовий згладжувач, сезонний згладжувач та згладжувач для виправлення невипадкових коливань. Даний метод дозволяє прогнозувати дані з високою точністю, враховуючи їхню динаміку та сезонні коливання. Для налаштування параметрів цього методу можна використовувати оптимізаційні підходи та метрики оцінки точності прогнозів

Основні кроки

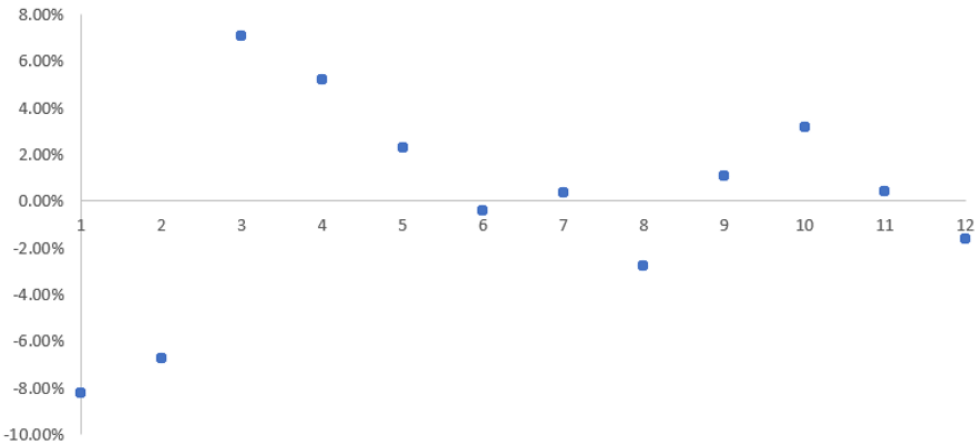
- Початкові значення тренду (T0), рівня (L0) та сезонного компонента (S0) визначаються на підставі початкових даних.
- Оновлення тренду, рівня та сезонного компонента проводиться згідно зі спеціальними формулами, які враховують актуальні дані та параметри згладжування.
- На основі оновлених значень розраховуються прогнози на майбутні періоди.
- Модель регулярно оновлюється з кожним новим значенням часового ряду.

Згладження історичних даних методом ковзаючого середнього

- Метод ковзаючого середнього використовується для згладжування історичних даних. Він полягає в обчисленні середнього значення даних за певний період, який «ковзає» по часовому ряду. Це допомагає виділити загальний тренд та відсіяти шуми, полегшуючи аналіз та прогнозування.



Рівень сезонних коливань



Однокроковий прогноз методом Холта-Вінтерса

Згладжування Холта-Вінтерса також називають потрійним експоненційним згладжуванням, за рахунок трьох параметрів, що згладжують (альфа, гама і сезонний фактор – дельта). Наприклад, якщо є 12-місячний сезонний цикл:

$$Yt = (l_{t-3} + 3 * Tr_{t-3}) * S_{t-12},$$

де l_{t-3} – рівень за період *t-3*, Tr_{t-3} – тренд за період *t-3*, S_{t-12} – сезонність за період *t-12*.

Однокроковий прогноз методом Холта-Вінтерса

Однокроковий прогноз розраховується за формулою:

$$Pr_{t1} = (L_{t0} + Tr_{t0}) * S_{AVG1},$$

де L_{t0} – рівень для нульового періоду, Tr_{t0} – тренд для нульового періоду, S_{AVG1} – середня сезонність відповідного періоду.

Визначення рівня

Рівень для нульового періоду визначається за допомогою рівняння лінії тренду (9144.3), кожний наступний розраховується за формулою:

$$L_{t1} = L_{t0} + Tr_{t0} + \alpha * \frac{E}{S_{AVG1}},$$

де L_{t0} – рівень для нульового періоду, Tr_{t0} – тренд для нульового періоду, α – константа згладжування (0.5), E – помилка прогнозу (різниця факту та прогнозу), S_{AVG1} – середня сезонність відповідного періоду.

Визначення тренду

Тренд для нульового періоду визначається за допомогою рівняння лінії тренду (32.868), кожней наступний розраховується за формулою:

$$Tr_{t1} = Tr_{t0} + \alpha * \beta * \frac{E}{S_{AVG1}},$$

де Tr_{t0} – тренд для нульового періоду, α – константа згладжування (0.5), β – константа згладжування тренду (0.5), E – помилка прогнозу (різниця факту та прогнозу), S_{AVG1} – середня сезонність відповідного періоду.

Середня сезонність

Середня сезонність для початкового періоду була розрахована на першому етапі, а наступні сезонності розраховується за формулою:

$$S_{AVG1} = S_{AVG0} * \gamma * (1 - \alpha) * \frac{E}{Tr_{t0} + L_{t0}},$$

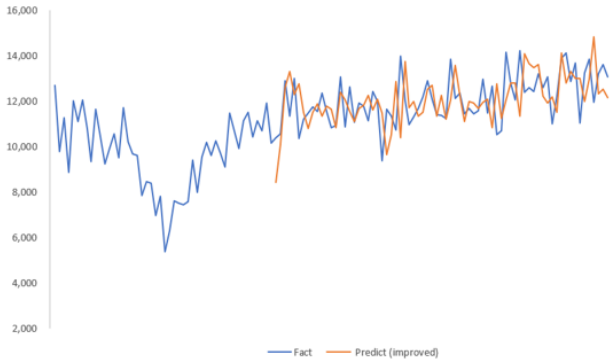
де S_{AVG0} – середня сезонність для початкового періоду, L_{t0} – рівень для нульового періоду, Tr_{t0} – тренд для нульового періоду, α – константа згладжування (0.5), γ – константа згладжування сезонності, E – помилка прогнозу (різниця факту та прогнозу).

Аналіз моделі методом Хольта-Вінтерса

Оцінка точності прогнозів моделі Холта-Вінтерса показала, що вона діє досить ефективно, особливо для короткострокових прогнозів.

MAPE = 4.94%

Важливим аспектом подальшого покращення моделі – є залучення категоріальних даних. Можливість дослідити ефективність різних методів кодування категоріальних змінних та їх вплив на точність прогнозів.



Висновки

- Прогнозування в сфері FMCG є важливим аспектом для планування запасів, виробництва та маркетингу.
- Модель Холта-Вінтерса є потужним інструментом для прогнозування в бізнесі з багатьма перевагами.
- В даній роботі ми досягли підвищення точності прогнозування. Від початкових значень MAPE на рівні 20% за допомогою методу ARIMA, ми досягли мінімальних MAPE на рівні 8% використовуючи метод експоненційного згладжування. Було вдосконалено результати, застосовуючи метод Холта-Вінтерса, де MAPE скоротилася до 5%. Це свідчить про успішність застосування комбінації цих методів для отримання точних прогнозів.
- Важливим аспектом подальшого покращення моделі – є залучення категоріальних даних.