

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА**

**Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»**

Кафедра математичного моделювання та статистики

Освітньо-професійна програма

Галузь знань

Спеціальність

Економічна кібернетика

05 Соціальні та поведінкові науки

051 Економіка

Форма навчання: очна (денна)

КВАЛІФІКАЦІЙНА БАКАЛАВРСЬКА РОБОТА

на тему **«Моделювання динаміки обсягів продажів за допомогою інструментів
машинного навчання»**

(назва теми)

здобувача Струка Олега Олександровича

(ПІБ, підпис)

Науковий керівник: кандидат економічних наук,
доцент Катуніна О. С.

(науковий ступінь, учене звання, ПІБ)

(підпис)

**Робота допущена до захисту перед екзаменаційною
комісією з атестації здобувачів вищої освіти (ЕК)**

Завідувач кафедри кандидат фізико-математичних наук,
професор Великоіваненко Г.І.

Київ 2023

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА**

**Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»
Кафедра математичного моделювання та статистики**

Освітньо-професійна програма
Галузь знань
Спеціальність

Економічна кібернетика
05 Соціальні та поведінкові науки
051 Економіка

ПОГОДЖЕНО

Керівник проектної групи (гарант)
освітньо-професійної програми
О.М.Притоманова

(підпис)

20__ р.

ЗАТВЕРДЖУЮ

Завідувач кафедри математичного
моделювання та статистики
Г.І.Великоіваненко

(підпис)

20__ р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

**здобувачу вищої освіти Струку Олегу Олександровичу
очної (денної) форми навчання**

на підготовку кваліфікаційної бакалаврської роботи

**на тему «Моделювання динаміки обсягів продажів за допомогою інструментів
машинного навчання»**

Тему затверджено наказом ректора Університету від " " _____ 2023 р. №

Кваліфікаційна бакалаврська робота виконується на матеріалах

ТОВ “АГРОС ЮА”

План кваліфікаційної бакалаврської роботи

Розділ 1	Теоретичні засади моделювання тенденцій розвитку ринку предметів споживання в умовах нестабільної економічної ситуації
(назва розділу)	
Розділ 2	Математичні моделі аналізу та прогнозування обсягів продажів товарів
(назва розділу)	
Розділ 3	Практичне застосування методів машинного навчання для аналізу та прогнозування обсягів продажів товарів
(назва розділу)	
Об'єкт дослідження:	Об'єктом дослідження є процеси реалізації та продажів товарів у торговельних мережах.

Предмет дослідження:	Предметом дослідження є система економіко-математичних методів, моделей аналізу та прогнозування динамічних тенденцій змінення обсягів продажів товарів у торгівельних мережах на базі інструментів одно- та багатофакторного моделювання часових рядів та методів машинного навчання.
Мета кваліфікаційної бакалаврської роботи:	Метою виконання даної роботи є побудова та експериментальне використання моделей аналізу динаміки продажів предметів споживання у торгівельній мережі з урахуванням сукупності факторів за допомогою методів машинного навчання.

Конкретні завдання, які здобувач повинен виконати для досягнення поставленої мети:

У розділі 1	Дослідити предметну область та функціонування ринку предметів споживання як об'єкту моделювання. Проаналізувати наукові джерела з наряду тенденцій розвитку ринку в умовах нестабільної економічної ситуації.
У розділі 2	Розглянути методи та інструменти моделювання. Розробити комплекс моделей для аналізу та прогнозування обсягів продажів товарів.
У розділі 3	Виконати експериментальні дослідження на основі запропонованих методів.

**Завдання підготував
науковий керівник**

(підпис)

О.С. Катуніна

(ініціали, прізвище)

« ____ » _____ 20__ р.

**Завдання одержав
здобувач**

(підпис)

О.О. Струк

(ініціали, прізвище)

« ____ » _____ 20__ р.

Реферат

Кваліфікаційна бакалаврська робота містить 68 сторінок, 15 рисунків, 4 формули, список використаних джерел з 30 найменувань, додатки.

«Моделювання динаміки обсягів продажів за допомогою інструментів машинного навчання»

Об'єктом дослідження кваліфікаційної бакалаврської роботи є процеси реалізації та продажів товарів у торгівельних мережах.

Предметом дослідження є система економіко-математичних методів, моделей аналізу та прогнозування динамічних тенденцій змінення обсягів продажів товарів у торгівельних мережах на базі інструментів одно- та багатофакторного моделювання часових рядів та методів машинного навчання.

Мета і завдання дослідження. Метою виконання даної роботи є побудова та експериментальне використання моделей аналізу динаміки продажів предметів споживання у торгівельній мережі з урахуванням сукупності факторів за допомогою методів машинного навчання.

Досягнення поставленої мети обумовлює необхідність вирішення таких завдань:

1. Дослідити принципи, предметну область та функціонування ринку предметів споживання як об'єкту моделювання.
2. Проаналізувати наукові джерела з напрямку тенденцій розвитку ринку в умовах нестабільної економічної ситуації.
3. Розглянути методи та інструменти моделювання.
4. Розробити комплекс моделей для аналізу та прогнозування обсягів продажів товарів.
5. Виконати експериментальні дослідження на основі запропонованих методів.

Теоретична, методична та практична значущість отриманих результатів. Кваліфікаційна робота має теоретичну значущість, оскільки вона

сприяє розширенню знань про використання машинного навчання для моделювання динамічних тенденцій в сфері торгівлі та їхнього впливу на бізнес-процеси. Методична значущість полягатиме в розробці практичних рекомендацій щодо використання машинного навчання в управлінні торговельними підприємствами.

Практична значущість полягатиме в можливості покращення стратегій управління, збільшення ефективності продажів та підвищення конкурентоспроможності торговельних підприємств. Цей досвід та результати дослідження можуть бути корисними для українських торговельних компаній та допоможуть їм оптимізувати свою справу для збільшення прибутку та успішного розвитку бізнесу.

Рік виконання кваліфікаційної бакалаврської роботи – 2023.

Рік захисту роботи – 2023.

Ключові слова: моделювання, моделювання динаміки, машинне навчання, модель ARIMA, продажі, торговельні мережі.

В і д г у к

про кваліфікаційну бакалаврську роботу
здобувача навчально-наукового інституту "Інститут інформаційних технологій в
економіці Київського національного економічного університету імені Вадима

Гетьмана

освітньо-професійної програми "Економічна кібернетика"

Струка Олега Олександровича

на тему «Моделювання динаміки обсягів продажів за допомогою інструментів
машинного навчання»

1. Актуальність теми. В умовах цифрової трансформації суспільства зростає актуальність стрімкого впровадження сучасних інтелектуальних технологій моделювання торгівельних мереж, у тому числі прогнозування обсягів продажів широкої сукупності товарно-асортиментних груп продукції з урахуванням системи передумов та низки факторів економічного середовища. Обсяг реалізації товарів як категорія, інтегрує ресурсні можливості торгівельних компаній, насиченість ринку та щільність конкурентних процесів на ньому, необхідність задовільного рівня ефективності діяльності підприємств тощо. Збільшення обсягів продажів є одним з ключових факторів успіху компаній та потребує комплексних досліджень на базі використання технологій математичного моделювання, науки про дані та машинного навчання. Роботу присвячено розробленню системи моделей та застосуванню модельних технологій визначення динамічних тенденцій змінення обсягів реалізації товарів у торгівельних мережах.

2. Позитивні риси кваліфікаційної бакалаврської роботи. У роботі систематизовано сучасні підходи до моделювання динаміки розвитку великих торгівельних мереж; здійснено огляд ефективних інструментів математичного моделювання для аналізу функціонування підприємств торгівлі; розроблено та реалізовано систему моделей прогнозування обсягових показників продажів товарів.

3. Наявність самостійних розробок автора. У кваліфікаційній бакалаврській роботі автором розв'язано задачу моделювання та прогнозування обсягів продажів товарів у торгівельній мережі, здійснено комп'ютерну реалізацію комплексу моделей, застосовано інструменти машинного навчання для вибору факторів, що впливають на реалізацію продукції, проведене експериментальне порівняння ефективності побудованих моделей, обґрунтовано вибір найбільш ефективних інструментів моделювання для економічної системи, яка досліджувалась.

4. Цінність теоретичних висновків та практичних рекомендацій. Запропонована система моделей є цілісним комплексом інструментів дослідження та прогнозування обсягів продажів товарів у торгівельній мережі, а тому є затребуваною учасниками відповідних ринків та може використовуватись підприємствами та компаніями. Розглянуті моделі можуть бути ефективно впроваджені в практику дослідницької, управлінської та консалтингової діяльності.

5. Наявність недоліків. В системі розглянутих в роботі моделей прогнозування обсягів продажів товарів у торгівельній мережі було б доцільним більшу увагу приділити використанню методів аналізу одномірних часових рядів, зокрема, методів декомпозиції часових рядів, а також застосуванню рекурентних нейромережових моделей, проте їхня відсутність не є суттєвим недоліком і не впливає на загальну якість розв'язання поставленої задачі.

6. Загальна оцінка кваліфікаційної магістерської роботи та її допущення до захисту перед ЕК: Зважаючи на те, що кваліфікаційна бакалаврська робота студента Струка Олега Олександровича на тему «Моделювання динаміки обсягів продажів за допомогою інструментів машинного навчання» є актуальною, дослідження є самостійним і ґрунтовним, вважаю, що дана кваліфікаційна робота відповідає вимогам та може бути допущена до захисту перед ЕК і заслуговує позитивної оцінки.

Науковий керівник

Катуніна О.С., к.е.н, доцент, доцент кафедри математичного моделювання та статистики Київського національного економічного університету імені Вадима Гетьмана

«___» _____ 2023 р.

Катуніна О.С.

ЗМІСТ

ВСТУП.....	4
 РОЗДІЛ 1. ТЕОРЕТИЧНІ ЗАСАДИ МОДЕЛЮВАННЯ ТЕНДЕНЦІЙ РОЗВИТКУ РИНКУ ПРЕДМЕТІВ СПОЖИВАННЯ В УМОВАХ НЕСТАБІЛЬНОЇ ЕКОНОМІЧНОЇ СИТУАЦІЇ.....	 7
1.1 Сутність, структура та особливості ринку предметів споживання як об'єкту моделювання.....	7
1.2 Аналіз тенденцій розвитку ринку предметів споживання в Україні.....	12
1.3 Огляд підходів до математичного моделювання обсягів продажів предметів споживання	18
 РОЗДІЛ 2. МАТЕМАТИЧНІ МОДЕЛІ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ОБСЯГІВ ПРОДАЖІВ ТОВАРІВ.....	 21
2.1 Методи розвідувального аналізу даних часових рядів	21
2.2 Моделювання динаміки обсягів реалізації на базі методів вирівнювання та згладжування часових рядів	27
2.3 Багатофакторне моделювання динаміки обсягів реалізації товарів.....	35
 РОЗДІЛ 3. ПРАКТИЧНЕ ЗАСТОСУВАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ОБСЯГІВ ПРОДАЖІВ ТОВАРІВ.....	 40
3.1 Розвідувальний аналіз та декомпозиція системи часових рядів, що визначають динаміку продажів товарів у торгівельній мережі	40
3.2 Аналіз результатів згладжування та ARIMA - моделювання динаміки продажів товарів.....	49
3.3 Аналіз результатів багатофакторного моделювання обсягів продажів товарів.....	54

ВИСНОВКИ	63
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	66
ДОДАТКИ	70

ВСТУП

Актуальність теми. Зв'язок машинного навчання та бізнесу набуває все більшого значення в сучасному світі. Із зростанням обсягів даних та швидкості їх обробки, компанії мають можливість аналізувати великі об'єми інформації та здійснювати швидкі та точні прогнози. Таким чином, вони знаходять нові можливості для зменшення витрат, покращення продуктивності та визначення стратегії розвитку. Однією з галузей, де машинне навчання може знайти своє застосування, є торгівля. Компанії можуть використовувати машинне навчання для прогнозування попиту, розуміння поведінки споживачів, оптимізації цін та багато іншого. Особливо ця тема актуальна для України, оскільки компанії мають адаптуватися під сучасні, нові реалії.

Аналіз останніх досліджень і публікацій. Останні дослідження та публікації підтверджують потенціал машинного навчання в сфері торгівлі. Одна зі статей, яка привернула мою увагу, це "Великі дані: нові трюки для економетрії" від Гала Р. Варіана, опублікована в журналі "Економічні перспективи" [1]. У цій статті автор розглядає різні аспекти використання алгоритмів машинного навчання в аналізі споживачів, прогнозуванні попиту, управлінні запасами, персоналізації рекомендацій та оптимізації ціноутворення.

Продовжуючи аналізувати останні дослідження та публікації, одна з інших важливих статей є "Прогнозування попиту за допомогою моделей машинного навчання; Сучасний стан і систематичний огляд досягнень" авторів Амїра Мосавї, Сїна Ардабіла та Шахабодїна Шамширбанда [2]. У цій статті автори провели систематичний огляд сучасних досягнень у використанні моделей машинного навчання для прогнозування попиту. Висвітлюючи значення цієї роботи, вона дозволяє прослідкувати найновіші тенденції у використанні машинного навчання для прогнозування обсягів продажів. Вона надає комплексний огляд різних моделей машинного навчання та їх ефективності в прогнозуванні попиту, що є важливим фактором для бізнесу в сфері торгівлі.

Крім того, певний ряд провідних дослідників приймали вирішальну роль у започаткуванні машинного навчання та його впровадженні у галузь торгівлі. Такі вчені, як Том М. Мітчелл, Джером Х. Фрідман, Стюарт Дж. Рассел, Пітер Норвіг, Альтман, Наомі та інші, внесли великий внесок. Їхні дослідження та внески визнані та широко використовуються у практичному застосуванні машинного навчання в бізнесі, зокрема у сфері торгівлі. Вони встановили основи та розробили ключові концепції, які сьогодні використовуються для розв'язання різних завдань у галузі бізнесу. Однак, хоча ці дослідження надають важливі вказівки, деякі аспекти проблеми ще залишаються невирішеними. Наприклад, важливо розробити ефективні методи моделювання сезонності, аналізувати вплив зовнішніх факторів на обсяги продажів та розробляти нові стратегії прогнозування з урахуванням динамічного середовища бізнесу.

Мета і завдання дослідження. Метою виконання даної роботи є побудова та експериментальне використання моделей аналізу динаміки продажів предметів споживання у торгівельній мережі з урахуванням сукупності факторів за допомогою методів машинного навчання.

Досягнення поставленої мети обумовлює необхідність вирішення таких завдань:

1. Дослідити принципи, предметну область та функціонування ринку предметів споживання як об'єкту моделювання.
2. Проаналізувати наукові джерела з напряду тенденцій розвитку ринку в умовах нестабільної економічної ситуації.
3. Розглянути методи та інструменти моделювання.
4. Розробити комплекс моделей для аналізу та прогнозування обсягів продажів товарів.
5. Виконати експериментальні дослідження на основі запропонованих методів.

Об'єкт дослідження. Об'єктом дослідження є процеси реалізації та продажів товарів у торгівельних мережах.

Предмет дослідження. Предметом дослідження є система економіко-математичних методів, моделей аналізу та прогнозування динамічних тенденцій

змінення обсягів продажів товарів у торгівельних мережах на базі інструментів одно- та багатofакторного моделювання часових рядів та методів машинного навчання.

Методи дослідження. У кваліфікаційній роботі для вирішення поставлених завдань використано загальнонаукові методи лінійної регресії, ARIMA моделі, VAR, та Random Forest, зокрема розглянемо й інші методи машинного навчання.

Теоретична, методична та практична значущість отриманих результатів. Кваліфікаційна робота має теоретичну значущість, оскільки вона сприяє розширенню знань про використання машинного навчання для моделювання динамічних тенденцій в сфері торгівлі та їхнього впливу на бізнес-процеси. Методична значущість полягатиме в розробці практичних рекомендацій щодо використання машинного навчання в управлінні торговельними підприємствами. Практична значущість полягатиме в можливості покращення стратегій управління, збільшення ефективності продажів та підвищення конкурентоспроможності торговельних підприємств. Цей досвід та результати дослідження можуть бути корисними для українських торговельних компаній та допоможуть їм оптимізувати свою справу для збільшення прибутку та успішного розвитку бізнесу.

Інформаційна база дослідження. Для проведення дослідження використані зарубіжні наукові статті, книги та публікації в журналах, статистичні дані офіційних сайтів та спеціалізованих інтернет-джерел. В рамках практичного дослідження використовуються історичні дані про обсяги продажу торговельної мережі Walmart з 05.02.2010 до 01.11.2012.

Структура роботи. Робота складається зі вступу, трьох розділів, списку використаних джерел та додатків.

РОЗДІЛ 1

ТЕОРЕТИЧНІ ЗАСАДИ МОДЕЛЮВАННЯ ТЕНДЕНЦІЙ РОЗВИТКУ РИНКУ ПРЕДМЕТІВ СПОЖИВАННЯ В УМОВАХ НЕСТАБІЛЬНОЇ ЕКОНОМІЧНОЇ СИТУАЦІЇ

1.1 Сутність, структура та особливості ринку предметів споживання як об'єкту моделювання

В сучасній економіці ринок предметів споживання відіграє ключову роль і включає широкий спектр товарів та послуг, які задовольняють потреби споживачів. Його сутність полягає в обміні товарами та послугами між продавцями та покупцями на підставі попиту та пропозиції. Для повного розуміння сутності ринку необхідно розглянути його основні функції [3]:

1. Остаточне визначення вартості товарів та послуг, а також їх реалізація і перетворення продукту праці на товар.
2. Забезпечення безперервності процесу суспільного відтворення (зокрема зв'язку між виробництвом та споживанням), сприяння формуванню цілісної національної економічної системи та її зв'язку з іншими національними економіками в масштабі світового ринку.
3. Стимулювання виробників товарів та послуг до зниження індивідуальних витрат порівняно з суспільно необхідними. Це сприяє покращенню економічної ефективності, й підвищенню загальної корисності для суспільства.
4. Підтримка контролю споживачів над виробництвом та вирівнювання цін.
5. Посилення конкуренції між виробниками товарів та послуг, як на внутрішньому ринку країни, так і на світовому.

Виходячи з цього, можна сформулювати загальне визначення сутності ринку. Ринок - це складна економічна система відносин між різними суб'єктами

господарювання, де відбувається купівля-продаж товарів та послуг, визнання їх вартості та реалізація. Це сприяє посиленню конкуренції між виробниками товарів, що спонукає їх знижувати витрати виробництва та підвищувати суспільну корисність товарів. Ринок також сприяє пропорційному розвитку національної економічної системи, забезпеченню безперервності суспільного відтворення та контролю споживачів над виробництвом.

Умови, необхідні для функціонування ринку, включають:

1. Здатність різних суб'єктів підприємницької діяльності впливати на рівень цін.
2. Наявність ефективного антимонопольного законодавства.
3. Розвинута система економічного, правового та адміністративного регулювання економіки державою.
4. Конкурентна боротьба між різними суб'єктами підприємницької діяльності.
5. Розгалужений комплекс об'єктів власності, які можуть бути предметом купівлі-продажу.

Структура ринку предметів споживання є досить різноманітною. Існують ринки праці, природних ресурсів, промислової продукції, землі, робочої сили, нерухомості, послуг, інтелектуальної власності та багато інших. Більшість з них можна поділити на різні сегменти або категорії, наприклад, ринок інтелектуальної власності поділяється на ринок патентів, ліцензій програмного забезпечення тощо. Товарний ринок поділяється на оптовий та роздрібний ринок.

Також, залежно від ступеня монополізації існують монополістичні та олігополістичні ринки, на монополізованому ринку один продавець в своїх руках концентрує всю масу виробництва, всю кількість товарів певного виду та диктує правила ринку. На немонополізованому ринку присутні багато продавців, кожен з яких не може впливати на процес формування цін. Олігополістичний ринок складається із декількох продавців, що мають певний товар чи послугу, та які поділили між собою ринок та вплив на рівень ціни. Незважаючи на таку різноманітність та залежить від галузі або категорії, у загальному сенсі структуру можна поділити на [4]:

1. Постачальників, тих хто виробляє товари чи надає послуги.

2. Споживачів – особи або організації, чиї потреби задовольняються.
3. Конкурентів – інші суб'єкти, які працюють у тій же галузі та конкурують за споживачів. Конкуренція може бути як локальною, так і глобальною, в залежності від розмірів ринку та місця діяльності компаній.
4. Регуляторів (держава, органи регулювання) – це органи управління, що встановлюють правила та норми для діяльності на ринку, забезпечуючи рівень безпеки, якість продуктів, права споживачів та інші аспекти.

Основними особливостями ринку є велика конкуренція, зміни у попиті, інноваційність та сезонність.

Ринок товарів та послуг зазвичай є конкурентним середовищем, де багато підприємств пропонують подібні або альтернативні товари та послуги. Конкуренція спонукає підприємства покращувати якість своїх пропозицій, знижувати ціни або шукати інші способи виокремитися на ринку.

Ринок товарів та послуг залежить від співвідношення між попитом та пропозицією. Попит визначається потребами та бажаннями споживачів, тоді як пропозиція відображає кількість товарів або послуг, які підприємства готові надати на ринку. Рівновага між попитом та пропозицією встановлює ціни та кількість товарів, які будуть продані.

Ринок товарів та послуг постійно змінюється через появу нових технологій, інновацій та розвиток нових продуктів. Підприємства, які успішно адаптуються до змін, можуть мати перевагу на ринку, пропонуючи нові або вдосконалені товари та послуги.

Ці особливості роблять ринок цікавим об'єктом для моделювання і дослідження. Завдяки моделюванню підприємства можуть аналізувати та прогнозувати попит, розуміти поведінку споживачів, визначати конкурентну стратегію та розробляти ефективні маркетингові рішення. Маркетинг, як одна із особливостей ринку, відіграє важливу роль у визначенні споживацьких тенденцій та формуванні попиту. Він включає в себе стратегії, такі як реклама, просування товарів, ціноутворення, позиціонування бренду та інші інструменти, які мають вплив на споживачів і їх рішення про покупку. Моделювання ринку допомагає

маркетологам краще розуміти та передбачати поведінку споживачів, а маркетингові рішення, з своєї сторони, надають цінну інформацію для побудови моделей. Наприклад, в процесі моделювання ринку можна враховувати різні фактори, такі як ціноутворення, реклама, позиціонування бренду, які впливають на споживачів і їх рішення про покупку. Ці фактори можуть бути враховані в математичних моделях, що дозволяє аналізувати їх вплив на попит та прибутковість продукту чи послуги.

Розглянемо останні тенденції на ринку. Стрімкий розвиток економіки та сфери бізнесу в сучасному світі вимагає постійного вдосконалення методологій, інструментів управління та планування. Щоб досягти успіху на конкурентному ринку, необхідно мати ефективну стратегію продажів, що заснована на комплексному аналізі ринкових тенденцій, поведінки споживачів та конкурентного середовища.

Сучасний світ переживає безліч економічних та бізнес-тенденцій, головною з яких є стрімке зростання онлайн-продажів та супутнє падіння продажів у звичайних магазинах. За тенденцією останніх років, що виникла в результаті світових потрясінь, кількість онлайн-покупців зросла у порівнянні з попередніми періодами. Це вказує на нагальну потребу компаній у запровадженні онлайн-продажів та оптимізації роботи своїх традиційних магазинів, аби вони не залишилися у минулому.

Ще одна помітна тенденція, яка з'явилася, — зростаюче значення машинного навчання та штучного інтелекту в сфері бізнесу. Оптимізація та автоматизація процесів за допомогою технологій машинного навчання дозволяє компаніям скорочувати витрати та підвищувати ефективність своєї діяльності [4].

Іншою важливою тенденцією є зростаюча популярність цифрової аналітики та великих даних (Big Data) у бізнесі, що привернула увагу компаній у всьому світі. Заглиблюючись у аналіз, великих масивів даних, компанії можуть отримати безцінне уявлення про поведінку клієнтів, ринкові тенденції та стратегії конкурентів, що допоможе їм приймати обґрунтовані рішення. Важливо також відзначити, що загострення конкуренції на ринку спонукає компанії постійно

шукати нові методи та інструменти для підвищення ефективності своєї діяльності. Таким чином, у цьому контексті використання машинного навчання та інструментів цифрової аналітики може надати компаніям значну конкурентну перевагу [5].

Наряду з цими тенденціями, також виникають й інші напрямки, що мають вплив на розвиток ринку. Наприклад, зростання свідомості споживачів щодо екологічної відповідальності. За останні роки, споживачі стали все більше усвідомлювати вплив своїх покупок на довкілля. Це ставить перед компаніями завдання розвитку бізнесу та розробки екологічно орієнтованих стратегій маркетингу.

Ще однією важливою тенденцією є зростання ролі соціальних медіа. Соціальні мережі надають компаніям можливість ефективно комунікувати зі споживачами, рекламувати свої товари та послуги. Соціальних медіа також стали важливим джерелом інформації про споживачів, їхні вподобання, що дозволяє компаніям налаштовувати свою стратегію на більш персоналізований підхід.

Не можна також не згадати про вплив глобалізації на ринок предметів споживання. Через зростання міжнародної торгівлі та технологічного прогресу, компанії мають можливість проникати на нові ринки та конкурувати зі світовими лідерами. Також, через глобалізацію, у споживачів відкривається доступ до широкого асортименту товарів та послуг з усього світу.

Ринок предметів споживання постійно розвивається та змінюється під впливом різних чинників. Нові технології, аналіз ринкових тенденцій, екологічна відповідальність та інноваційні підходи є важливими факторами успіху компаній, що діють на ринку. Адаптація до цих тенденцій дозволяє підприємствам бути конкурентоспроможними та успішними в сучасному економічному середовищі.

1.2 Аналіз тенденцій розвитку ринку предметів споживання в Україні

За останні роки український ринок споживання набрав обертів та розвивався, але на даний момент він все ще знаходиться у незадовільному стані. Значна проблема, що сповільнює розвиток ринку предметів споживання в Україні, полягає в поширеності бідності (рис. 1.1) та низької покупної спроможності населення.

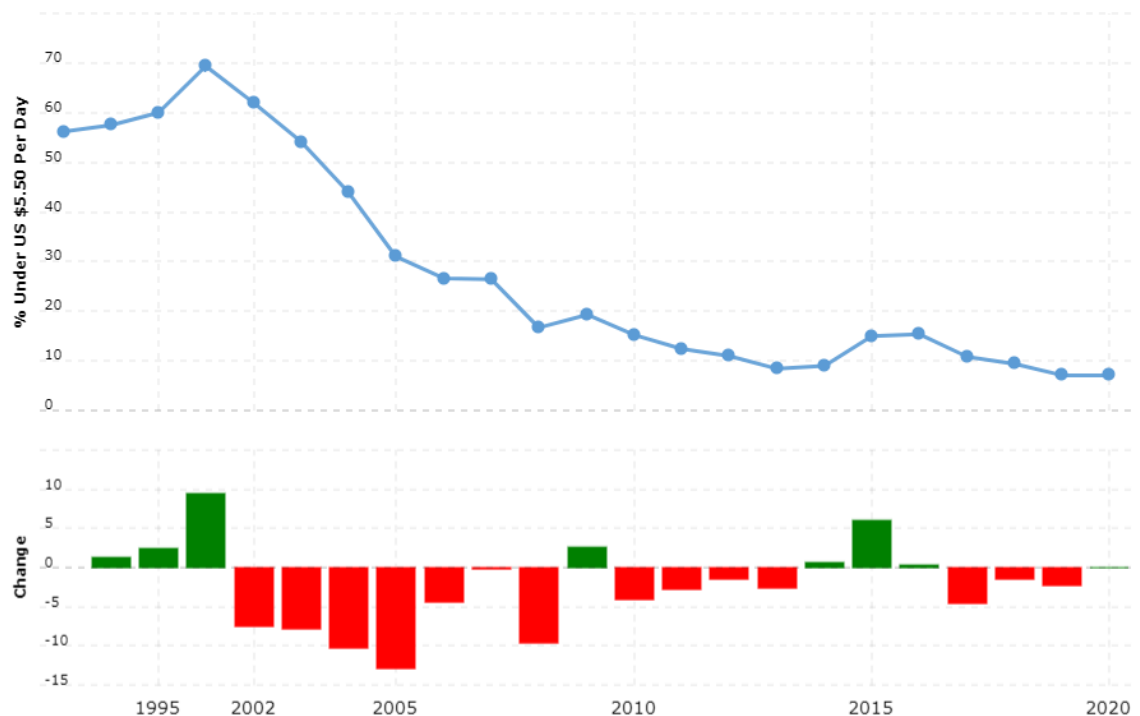


Рисунок 1.1 – Рівень бідності в Україні 1992-2020 рр.

Джерело: [6]

Високий рівень бідності впливає на споживчу активність та здатність населення купувати товари та послуги. За даними, які наведені на графіку, можна спостерігати, що відсоток бідності в Україні знижувався протягом усього періоду, що свідчить про певні позитивні зміни в економічному та соціальному розвитку країни.

Можна спостерігати, як у період з 2014 року, після політичної нестабільності, рівень бідності дещо виріс, однак з 2017 року почав показувати стійкий спад. У період наступної кризи, яка була спричинена пандемією у світі, цей показник вдалося утримати сталим, за різними джерелами, він не критично варіюється. Якщо говорити про сьогоднішній день, по неофіційним даним, цей показник росте швидкими темпами.

Однією з причин, що уповільнюють розвиток ринку предметів споживання в Україні, є значний рівень державного втручання в економіку, недостатня ефективність податкової політики, бюрократичні перешкоди та непрозорість в управлінні, що створює негативну атмосферу для розвитку ринку. Непередбачуваність управлінських рішень та зміни в законодавстві можуть створювати несприятливі умови для бізнесу, що обмежує інвестиції та розвиток нових підприємств.

Також, негативними аспектами, що сповільнюють розвиток ринку, є низький рівень довіри до якості товарів, корупція, недостатня захищеність прав інтелектуальної власності, недостатня розвиненість логістичної інфраструктури та високі витрати на транспортування.

Нові реалії, спричинені глобальною економічною нестабільністю, примусили ринок України зазнати суттєвих змін. Криза, пандемія та особливо війна суттєво вплинули на ринок України, насамперед, внаслідок економічних труднощів та невизначеності, споживачі змінили свої пріоритети та спосіб витрати коштів. Більшість споживачів стали більш обережними при витрачанні грошей та більш уважними до цін, шукаючи більш доступні товари та послуги.

Обмеження на рух та фізичну взаємодію призвели до значного зростання популярності онлайн-покупок. Для роздрібних магазинів, це обмеження вимагає впровадження додаткових заходів безпеки та перегляду своєї логістики. Багато магазинів переключилися на онлайн-продажі та доставку товарів [7].

Під час тривалого карантину та обмежень безпеки, деякі галузі економіки постраждали найбільше, а саме: готельний та ресторанний сектор, туристичний, розважальний, транспортний сектори та виробничий бізнес (рис.1.2).

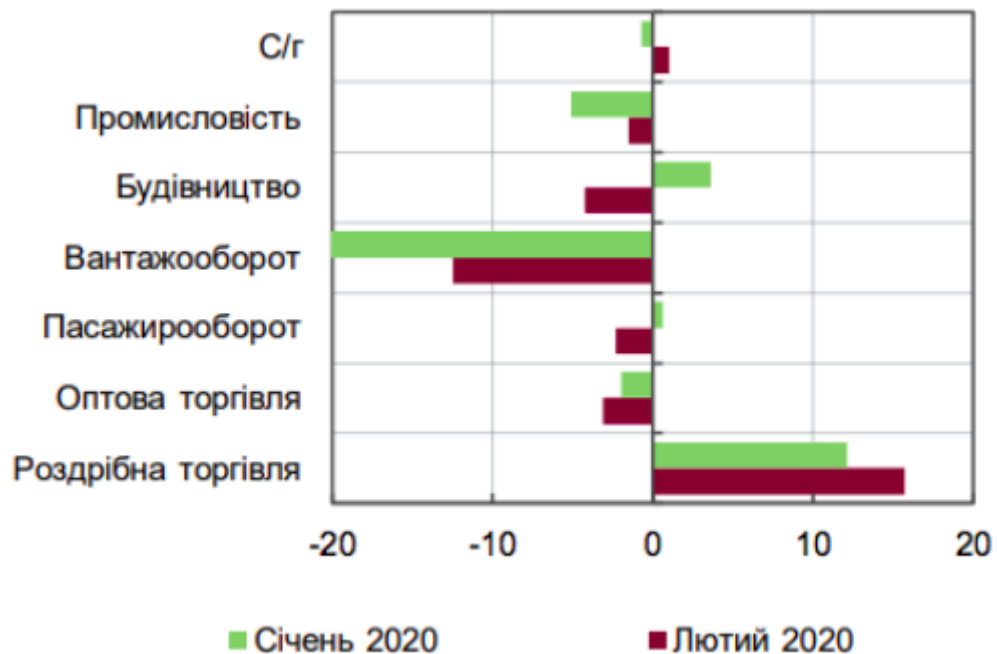


Рисунок 1.2 – Випуск в окремих видах діяльності, приріст (або зменшення) порівняно з попереднім роком, %.

Джерело: [7]

Також, пандемія викликала зміни в споживчому попиті, зростання попиту на медичні товари, засоби індивідуального захисту, продукти харчування, доставку їжі та інші схожі послуги.

Всі ці тенденції показують адаптацію ринку України до складних умов, підприємства змушені змінювати свої стратегії та швидко адаптуватися до нових реалій.

Незважаючи на низку проблем, ринок роздрібної торгівлі в Україні є досить конкурентним та намагається постійно розвиватися, що зумовлює появу нових технологій. В наш час, для торговельних компаній важливо мати ефективні моделі прогнозування продажів, щоб забезпечити своєчасне поповнення товарного запасу, високу якість обслуговування клієнтів та ефективне планування бізнес-процесів. На сьогоднішній день, українські торговельні компанії використовують різні моделі прогнозування продажів для вирішення своїх завдань. Розглянемо існуючі моделі прогнозування продажів, їх переваги та недоліки, що дозволить краще зрозуміти, яким чином можна ефективно працювати в умовах українського ринку.

Найбільші торгівельні мережі в Україні, які володіють мережами супермаркетів та гіпермаркетів по всій країні, у своїй діяльності використовують різноманітні методи та моделі.

Одним із поширених таких методів є використання нейронних мереж для прогнозування продажів. Для цього компанії збирають дані про історію продажів, сезонність, рекламні акції, погоду, календарні свята та інші фактори, що можуть вплинути на продажі. Ці дані використовуються для навчання нейронної мережі, яка прогнозує продажі на наступний період. За допомогою цієї моделі, компанія може визначити оптимальний запас товарів, який задовольнятиме попит, а також зменшити кількість витрачених коштів на зберігання запасів.

Крім того, така модель дозволяє компанії планувати рекламні кампанії та промо-акції з урахуванням очікуваних продажів та максимізувати прибуток. Методу нейронних мереж властиві висока точність прогнозування, нейронні мережі можуть прогнозувати продажі з високою точністю, що дозволяє компаніям ефективніше управляти своїми запасами товарів та забезпечувати їх своєчасним поповненням.

Використання нейронних мереж дозволяє зменшити кількість ручної роботи, оскільки модель може автоматично аналізувати великі обсяги даних та робити прогнози. Крім того, нейронні мережі можуть працювати в режимі реального часу, що дозволяє компанії оперативно реагувати на зміни в умовах ринку та адаптувати свої стратегії продажів [8].

Іншим методом машинного навчання є градієнтний бустинг (gradient boosting), який дозволяє створювати комплексні моделі на основі декількох базових моделей. Кожна наступна модель намагається покращити помилки попередньої моделі, що дозволяє підвищувати точність прогнозування. Щоб застосувати цей метод, компанії використовують дані про історію продажів, дані про попит, ціну, конкурентну ситуацію та інші фактори. Далі дані використовуються для створення базових моделей, які пізніше комбінуються в єдину модель градієнтного бустингу.

Основні властивості градієнтного бустингу полягають у тому що, цей метод може працювати з великою кількістю різних факторів, що може значно підвищити

точність прогнозування. Градієнтний бустинг може підкреслити значущість окремих факторів, що дозволяє компанії зосередитися на найважливіших аспектах своєї діяльності. Градієнтний бустинг є досить гнучким методом, який може працювати з різними типами даних та моделями. Крім того, цей метод може допомогти зменшити помилки прогнозування та підвищити ефективність діяльності компанії [9].

Також, в деяких торгівельних компаніях використовують аналіз геоданих (geospatial analysis) для прогнозування продажів. Аналіз геоданих дозволяє враховувати просторові фактори, такі як місце розташування магазину, популяційні дані, дані про конкурентів та інші фактори, що можуть вплинути на продажі [10]. Аналіз геоданих дозволяє компанії створити точні прогнози продажів для кожного магазину.

Загалом, існує багато різних моделей продажів, які можуть бути використані в торгівельних компаніях. Кожна модель має свої переваги та недоліки, тому вибір моделі залежить від конкретної ситуації та обставин. Наприклад, модель лінійної регресії може бути використана, якщо існує пряма лінійна залежність між факторами та продажами, а метод ARIMA може бути корисним для прогнозування продажів на основі історичних даних про продажі.

При виборі моделі важливо також враховувати доступні дані та ресурси компанії. Деякі моделі можуть вимагати великих обсягів даних та складних обчислень, тому не кожна компанія може їх використовувати.

Крім того, важливим є процес підготовки даних для моделювання. Для досягнення точних прогнозів необхідно мати достовірні та релевантні дані. Дані повинні бути зібрані та організовані у відповідний спосіб, а також проведена попередня аналітика даних для виявлення взаємозв'язків між факторами та продажами.

Якщо говорити про світовий ринок, то різні торгівельні компанії використовують різні моделі прогнозування продажів, залежно від їхніх потреб та можливостей. Компанії збирають велику кількість даних про продажі, ціни, запаси, попит, погоду та інші фактори, які можуть вплинути на продажі. З цих даних

створюють моделі прогнозування продажів, які використовуються для планування запасів, управління ланцюгом поставок та інших аспектів діяльності компанії.

Іншим, схожим варіантом для прогнозування продажів є спосіб, коли компанії збирають дані про продажі, пошукові запити, відгуки клієнтів та інші дані, щоб створювати прогнози попиту на товари та послуги. Використовуючи ці прогнози, компанія може забезпечити свої склади та партнерів-постачальників необхідною кількістю товарів в потрібний момент.

Якщо порівняти використання моделей прогнозування продажів світових компаній та українських, то ці методи засновані на одному принципі та відрізняються тільки глибиною аналізу, масштабами можливостей автоматизації та використання технологій у роботі. Однак, компанії мають спільну мету - ефективно управління запасами та підвищення рівня продажів за рахунок точного прогнозування попиту.

Один з основних факторів, який впливає на продажі, - це попит на товари. Торгівельні компанії можуть використовувати моделі прогнозування попиту, які враховують такі фактори, як історію продажів, ціну товару, маркетингові акції та конкурентну ситуацію. Це дозволяє компанії визначати оптимальну кількість товарів, які потрібно замовити, щоб задовольнити попит на них.

Торгівельні компанії можуть використовувати прогнозування продажів, щоб оцінити ефективність рекламних кампаній. Наприклад, компанія може порівняти продажі в період, коли була проведена рекламна кампанія, з продажами в попередній та наступні періоди, щоб визначити, наскільки успішною була кампанія. Багато товарів мають сезонний характер, тобто їх попит змінюється залежно від пори року.

Торгівельні компанії можуть використовувати метод прогнозування сезонності, щоб визначити очікуваний попит на даний товар в конкретний період часу. Це дозволяє компанії підготувати запаси товарів на пік сезонного попиту та забезпечити належний рівень обслуговування клієнтів. Ціна товару є одним з основних факторів, який впливає на продажі. Оптимальна ціна дозволяє збільшити обсяги продажів та маржинальну прибутковість.

1.3 Огляд підходів до математичного моделювання обсягів продажів предметів споживання

У сучасному мінливому бізнес-середовищі точне прогнозування продажів стає все більш важливим завданням для торговельних компаній. Складність ринкової динаміки вимагає від компаній точних прогнозів продажів, які можуть забезпечити їх необхідними обсягами товарів та послуг для задоволення попиту клієнтів.

Традиційні методи прогнозування, такі як експертні оцінки та статистичні методи, можуть бути недостатньо ефективними, оскільки вони не враховують складні залежності та нелінійні зв'язки між різними факторами, що впливають на продажі. Саме тут у гру вступає машинне навчання - воно може допомогти підвищити точність прогнозування, аналізуючи великі обсяги даних і виявляючи складні взаємозв'язки між факторами та продажами [11].

Машинне навчання може бути використане для аналізу поведінки покупців, що дозволяє торговельним компаніям зрозуміти, які товари покупці купують, коли вони їх купують і як часто вони купують. Ця інформація може бути використана для планування стоків і акцій, які збільшать продажі.

Зокрема, одним з конкретних підходів, який може бути використаний, є застосування нейронних мереж, що можуть виявити неявні взаємозв'язки між різними факторами та прогнозуванням продажів, які може бути важко виявити людині.

Точне прогнозування продажів є вирішальним елементом для успіху будь-якої торгової компанії. Це дозволяє компанії підтримувати достатні запаси товарів на складах, запобігати надлишковості та забезпечувати належний рівень обслуговування клієнтів. Використовуючи інструменти машинного навчання, компанії можуть значно підвищити точність прогнозування продажів, тим самим розширюючи свої можливості планування виробничих процесів і управління

запасами, знижуючи складські витрати та в кінцевому підсумку, збільшуючи прибуток компанії.

Щоб підвищити точність прогнозування продажів торговельної компанії за допомогою інструментів машинного навчання - необхідно виконати декілька завдань [12]:

1. Необхідно зібрати та підготувати відповідні дані про продажі компанії: обсяг продажів, економічні показники, дані про споживачів, ринки збуту та інші фактори, які можуть впливати на продажі.
2. Необхідно провести попередню обробку та аналіз даних для відбору важливих ознак та підготовки даних для побудови моделі.
3. Розробка та навчання моделі прогнозування обсягів продажів вимагатиме вибору найбільш підходящих методів машинного навчання, які будуть застосовані.
4. Тестування та оцінка точності моделі на реальних даних торговельної компанії також будуть необхідні для оцінки її ефективності.
5. Провести аналіз результатів тестування - це дасть змогу зробити висновки про ефективність методів машинного навчання для прогнозування обсягів продажів компанії.

Машинне навчання досліджує глибини обчислювальних моделей, здатних динамічно адаптуватися та покращувати свої прогностичні здібності шляхом аналізу та розпізнавання величезних обсягів даних. Розглянемо деякі популярні алгоритми машинного навчання, які є предметом цього дослідження [13]:

1. Лінійна регресія.

Алгоритм лінійної регресії застосовується для моделювання залежності між змінними. Він використовує лінійну комбінацію вхідних даних для передбачення вихідного значення. Уявімо, що ми маємо дані про обсяги продажів (залежна змінна) та рекламні витрати (незалежна змінна), за допомогою лінійної регресії ми можемо побудувати модель, яка прогнозує обсяги продажів на основі рекламних витрат. Наприклад, якщо ми спостерігаємо, що кожні \$1000 витрат на рекламу призводять до збільшення обсягів продажів на 50 одиниць, то ми можемо

використати цю залежність для передбачення обсягів продажів за новими значеннями витрат.

2. Експоненційне згладжування.

Даний метод прогнозування часових рядів використовує попередні значення ряду для прогнозу майбутніх значень. Зазвичай, цей метод добре підходить для даних з трендами та сезонністю.

3. ARIMA (Авторегресійне інтегроване ковзне середнє).

Це статистичний метод, що використовується для прогнозування часових рядів. Наприклад, якщо у нас є щомісячні дані про обсяги продажів протягом останніх кількох років, ми можемо застосувати ARIMA для прогнозування майбутніх обсягів продажів на основі попередніх значень.

4. VAR (Векторна авторегресія).

Це модель, що використовується для прогнозування часових рядів, в якій кожна змінна залежить від своїх попередніх значень та значень інших змінних. Наприклад, якщо ми спостерігаємо залежність між обсягами продажів та цінами на ринку, VAR може допомогти в прогнозуванні майбутніх обсягів продажів на основі попередніх значень обсягів продажів та цін.

5. Random Forest.

Це алгоритм машинного навчання, що використовується для класифікації та регресії. В контексті прогнозування продажів, Random Forest може використовуватись для створення моделі, яка аналізує різні ознаки, такі як рекламні витрати, ціни на ринку, часові фактори тощо, і прогнозує обсяги продажів на основі цих ознак.

6. Нейронні мережі.

Алгоритм, який називається нейронною мережею, імітує поведінку нейронів мозку для вирішення завдань машинного навчання. Нейронні мережі, завдяки ретельному процесу навчання здатні досягати вражаючої точності у вирішенні різних завдань, включаючи класифікацію, прогнозування і розпізнавання образів, але не обмежуючись ними [14].

РОЗДІЛ 2

МАТЕМАТИЧНІ МОДЕЛІ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ОБСЯГІВ ПРОДАЖІВ ТОВАРІВ

2.1 Методи розвідувального аналізу даних часових рядів

Процес дослідницького аналізу даних, що передбачає всебічне, комплексне вивчення певного набору даних з метою виявлення закономірностей, залежностей та інших важливих особливостей називається розвідувальним аналізом даних (Exploratory data analysis, EDA). Як правило, він виконується на початковому етапі дослідження та слугує для попереднього вивчення даних перед тим, як вони будуть піддані більш детальному аналізу. Процес EDA складається з низки важливих кроків, які дозволяють досліднику краще зрозуміти та проаналізувати наявний набір даних [15]:

- Завантаження даних.
- Описова статистика.

Описова статистика охоплює низку методів, які дають всебічний огляд певного набору даних. Сюди входить розрахунок основних показників, таких як середнє, медіана, мінімальне та максимальне значення, стандартне відхилення та коефіцієнти кореляції між змінними. За допомогою описової статистики дослідник ретельно вивчає атрибути кожної змінної в наборі даних. Описова статистика дає змогу узагальнювати та інтерпретувати дані. Вони використовуються для вивчення даних, виявлення закономірностей, аномалій або відхилень.

Використовуючи описову статистику, дослідники можуть визначити важливі характеристики даних, такі як центральна тенденція, варіабельність та кореляція між змінними. Показники, такі як середнє значення та медіана, дають загальне уявлення про дані та дозволяють дослідникам визначити типові значення

змінної. Ці показники корисні для розуміння розподілу даних і для порівняння між групами.

Показники варіабельності - дають уявлення про розкид або дисперсію даних. Висока варіабельність може свідчити, що дані розпорошені та не сконцентровані навколо середнього значення. Низька варіабельність може вказувати на те, що дані щільно згруповані навколо середнього значення. Такі показники можна використовувати для виявлення викидів або аномальних значень у даних, які можуть спотворити аналіз. Кореляційний аналіз вимірює зв'язок між змінними. Обчислюючи коефіцієнт кореляції, дослідники можуть визначити силу та напрямок зв'язку між змінними.

Така інформація може бути використана для того щоб визначити, які змінні мають найбільший вплив на результат, що нас цікавить, а також для розробки прогностичних моделей.

Загалом, описова статистика надає дослідникам інструмент для вивчення та розуміння даних. Узагальнюючи дані в осмислений спосіб, описова статистика дозволяє приймати обґрунтовані рішення щодо даних та розробляти гіпотези для подальших досліджень.

- Візуалізація даних.

Спосіб відображення даних у вигляді графіків, діаграм та інших графічних зображень, називається візуалізацією даних. Всебічний аналіз даних може бути здійснений за допомогою ефективного методу візуалізації даних, який матеріалізує дані, ілюструючи їх в такий спосіб, можна розпізнати взаємозв'язки та кореляції між різними змінними.

Візуалізація даних не лише покращує розуміння складних наборів даних, але й надає візуальні підказки, які часто неможливо розгледіти за допомогою інших традиційних аналітичних методів. Візуалізація даних може допомогти виявити взаємозв'язки між змінними, показати розподіл даних, виявити викиди та аномалії. Це дозволяє зробити більш об'єктивні висновки про дані та спростити їхнє розуміння. Серед основних методів візуалізації даних можна виділити чимало підходів, зокрема такі як: гістограми, діаграми розсіювання, box plots та heat maps.

Кожен із них має свої унікальні властивості, гістограма - показує розподіл даних за кількістю спостережень у певному діапазоні значень. Діаграма розсіювання - ефективно демонструє залежність між двома змінними. Box plot - представляє розподіл даних графічно, заповнений визначальними деталями, такими як мінімальне та максимальне значення, медіана та квартилі, підвищує розуміння даних у їхній цілісності. Heat map - відображає залежність між двома змінними у вигляді кольорової матриці, підкреслює взаємозв'язок між даними з великою чутливістю.

Крім того, інтерактивна візуалізація даних може виявитися дуже корисною, оскільки дозволяє ретельно вивчати дані під різними кутами та на різних рівнях деталізації, використовуючи інтерактивні функції, такі як масштабування і перетягування. Таким чином, можна вивчити дані більш глибоко та виявити візуальні закономірності, які можуть залишатися прихованими на традиційних графіках.

- **Перевірка гіпотез.**

Після проведення розвідувального аналізу даних можна перейти до формулювання кількох гіпотез, що стосуються складних взаємозв'язків або наслідкових відмінностей між змінними. Для перевірки цих гіпотез і підтвердження їхньої істини чи хибності можна застосувати низку ретельно розроблених і науково обґрунтованих статистичних методів і прийомів, наприклад, тест Стюдента для перевірки різниці між середніми значеннями двох груп, ANOVA для перевірки різниці між середніми значеннями більше ніж двох груп, чи кореляційний аналіз для визначення ступеня залежності між двома змінними.

Перевірений часом статистичний тест Стюдента, має високу репутацію як потужний інструмент для перевірки різниці між середніми значеннями двох груп. Тест Стюдента розраховує t -статистику, яка відображає різницю між середніми значеннями груп в стандартних одиницях. Значення t -статистики порівнюється з критичним значенням, яке залежить від розміру вибірок та рівня значущості. Якщо t -статистика перевищує критичне значення, то можна зробити висновок про статистичну значущість різниці між групами [16].

Метод, який дає глибоке розуміння взаємозалежності між двома змінними, вимірює ступінь зв'язку однієї змінної та іншою, називається кореляційним аналізом.

Кореляційний аналіз дозволяє дослідникам встановити величину і напрямок статистично значущого зв'язку між двома змінними. Аналізуючи дві числові змінні, які можуть бути показниками різних атрибутів або вимірами одного і того ж явища, дослідники можуть визначити наявність кореляції.

Коефіцієнт кореляції, який коливається від -1 до 1, дає чітке уявлення про зв'язок між змінними. Зокрема, коефіцієнт -1 означає повну обернену залежність, коефіцієнт 0 - відсутність кореляції, а коефіцієнт 1 - абсолютну пряму кореляцію між двома змінними. Таким чином, кореляційний аналіз є безцінним інструментом для дослідників, які прагнуть зрозуміти ступінь зв'язку між двома змінними та кількісно оцінити силу і характер їхнього взаємозв'язку [17].

Розглянемо, наприклад, залежність між віком та доходом. За допомогою кореляційного аналізу можна отримати неоціненну інформацію про ступінь взаємозалежності між цими двома змінними.

Якщо коефіцієнт позитивний і наближається до 1, дослідники можуть зробити висновок, що зі збільшенням віку зростає і дохід. І навпаки, якщо коефіцієнт від'ємний і близький до -1, можна зробити висновок, що між віком і доходом існує обернений зв'язок, тобто зі збільшенням віку дохід зменшується.

У випадку, коли коефіцієнт наближається до 0, дослідники можуть зробити висновок, що вік і дохід не залежать один від одного, що свідчить про відсутність статистичної значущості. Таким чином, кореляційний аналіз надає дослідникам потужний інструмент для вивчення і розшифровки складних зв'язків між різними змінними, що дозволяє їм робити цінні висновки і розкривати розуміння їх взаємозв'язку.

- Обробка пропущених даних.

Іноколи статистичні дані можуть бути неповними, тобто містити пропущені значення. Коли дані містять пропущені значення, це може спотворити результати аналізу даних, викликати помилки в оцінках залежності між змінними, вплинути

на результати аналізу та вивести з грядущої гіпотези. Дані могли бути пропущені в результаті відсутності інформації, помилок у зборі даних або відмови респондентів від відповіді на певні запитання.

Для обробки пропущених даних можна використовувати різні методи. Один з найпростіших методів - це використання середнього значення або медіани для заповнення пропущених значень. Цей метод є досить ефективним, якщо пропущені дані не становлять великої частини всієї вибірки та якщо пропуски є випадковими.

Інший метод - це використання методів множинної регресії або аналізу даних, щоб передбачити відсутні значення на основі інших змінних. Цей метод може бути корисним, якщо є декілька змінних, які корелюють зі змінною, яка має пропущені значення.

У будь-якому випадку, необхідно пам'ятати, що використання методів заповнення пропущених значень може вплинути на статистичну точність та достовірність результатів аналізу. Як наслідок, необхідно приділити увагу вибору методу і переконатися, що підхід відповідає конкретним вимогам дослідження.

- Обробка викидів та аномалій.

Викиди та аномалії представляють собою значення, які значно відрізняються від решти значень у наборі даних, вони можуть спричинити хаос в аналітичних результатах, вносячи заплутані ефекти та викривлення, тому їх потрібно обробляти.

Враховуючи їхній значний вплив на аналіз даних, вкрай важливо, щоб до викидів ставилися з великою обережністю. Для того, щоб впоратися з цими особливими точками даних, дослідники можуть використовувати ряд стратегій.

Одним із поширених методів обробки викидів є повне виключення їх з набору даних, таким чином мінімізуючи їхній вплив на аналіз. Крім того, інший метод - це заміна. Дослідники можуть замінити викиди іншими значеннями, такими як середнє або медіана, таким чином пом'якшуючи їхній вплив на набір даних, зберігаючи загальну структуру даних. Цей метод може бути більш ефективним, коли викиди є досить численними та вони впливають на головні властивості даних.

Ще одним з методів обробки викидів є видалення їх з набору даних. Цей метод може бути корисним, якщо кількість викидів є досить невеликою та вони не впливають на головні властивості розподілу даних. Однак, цей метод може вплинути на точність та повноту даних, тому видалення викидів має бути обґрунтовано та проведено з обережністю.

Незалежно від стратегії, яка використовується для усунення відхилень важливо оцінити їх наявність та вплив на результати аналізу. Наприклад, для цього можна використовувати такі графічні методи, такі як гістограма, ящик з вусами та діаграма розсіювання.

- Декомпозиція часового ряду.

Процес розбиття ряду на складові компоненти, для більшого розуміння її поведінки та внесення якості залежності між часом та значенням ряду – називається декомпозицією часового ряду. Зазвичай часовий ряд можна розкласти на основні компоненти, такі як, тренд, сезонність та шум.

Тренд відображає зміну значень часового ряду, він може бути висхідним, низхідним або плоским, що означає зростання, спадання, або стабільність значень відповідно.

Сезонність відображає повторювані коливання в часовому ряді, які стало повторюються протягом фіксованого періоду.

Шум відображає випадкові коливання та може бути причиною непередбачуваних аномалій або випадкових збурень в часовому ряді.

- Перевірка стаціонарності.

Стаціонарність часового ряду означає, що статистичні властивості ряду залишаються постійними з часом, щоб перевірити на стаціонарність можна використовувати статистичні тести, такі як тест Дікі-Фуллера, даний тест статистично перевіряє гіпотезу про наявність одиничного кореня в ряду, який свідчить про нестаціонарність

- Автокореляція.

Автокореляція вимірює ступінь залежності між значеннями часового ряду та його лагованими копіями. Автокореляційна функція показує, наскільки значення

ряду взаємозв'язані зі значеннями попередніх лагів. Це допомагає виявити наявність сезонності або інших взаємозв'язків в ряду. Графічне представлення автокореляції може бути зроблено за допомогою корелограми, вона відображає автокореляційні коефіцієнти на різних лагах та за її допомогою можна виявити сезонність та інші патерни в ряду.

2.2 Моделювання динаміки обсягів реалізації на базі методів вирівнювання та згладжування часових рядів

Одномірний ряд є дискретною функцією числової послідовності. Аналіз одномірного ряду передбачає вивчення рівнянь, закономірностей та поведінки даних у часовому ряді. Метою його аналізу є розкриття структури даних, прогнозуванні майбутніх значень та прийнятті рішень на основі цих даних.

Одномірний ряд є одним з основних об'єктів дослідження в аналізі часових рядів. Він може бути використаний для аналізу та прогнозування різних типів даних, таких як фінансові, економічні, соціальні, метеорологічні та інші. Аналіз одномірного ряду включає в себе різні методи та техніки, які дозволяють досліджувати взаємозв'язки між спостережуваними значеннями та знайти закономірності, які описують поведінку даних у часовому ряді.

Залежно від поставленої задачі, одномірний ряд може вирішувати завдання різного типу. Наприклад, в дослідженнях фінансових ринків метою може бути прогнозування цін на акції, у метеорології - прогнозування погоди, а в медицині - аналізування динаміки показників здоров'я пацієнтів.

Прогнозування майбутніх значень ряду є однією з основних цілей аналізу одновимірних рядів. Це може бути корисним для прийняття рішень у різних сферах, включаючи планування виробництва, управління запасами, фінансовий аналіз тощо.

Іншою важливою метою аналізу одномірного ряду є виявлення структури даних та закономірностей, що дозволяє зрозуміти, які фактори впливають на спостережувані значення ряду. Це може допомогти в розумінні причинно-наслідкових зв'язків та плануванні майбутніх дій.

Також аналіз одномірного ряду може допомогти виявити аномальні значення, що можуть бути показниками помилок в даних або незвичайних подій. Це може допомогти уникнути можливих проблем, виправити помилки та підготувати більш точні прогнози [18].

Експоненційне згладжування є методом згладжування часового ряду шляхом видалення випадкових змін. Цей метод полягає у тому, що кожне нове спостереження в ряду згладжується з попереднім значенням, використовуючи експоненційно зважене середнє. Експоненційно згладжене середнє відображає сезонність та тренди в даних, що робить його корисним для прогнозування майбутніх значень.

Експоненційне згладжування є одним з найпоширеніших методів згладжування часових рядів. Цей метод згладжує часовий ряд шляхом видалення випадкових змін, тобто видаляє шум та інші нерегулярні фактори, що впливають на ряд. Експоненційне згладжування базується на тому, що кожне нове спостереження в ряду згладжується з попереднім значенням, використовуючи експоненційно зважене середнє.

Метод експоненційного згладжування може бути застосований до рядів з різними сезонністю та трендами. При згладжуванні рядів з виразно вираженим трендом, експоненційно згладжене середнє буде близьким до тренду, що дозволяє виявити зміни тренду та спрогнозувати майбутні значення. У рядів з виразно вираженою сезонністю, експоненційно згладжене середнє може допомогти виявити та врахувати циклічні зміни.

У сфері аналізу часових рядів експоненціальне згладжування є дуже потужним і ефективним інструментом, який добре підходить для прогнозування майбутніх значень та тенденцій. Тим не менш, важливо пам'ятати, що експоненціальне згладжування не здатне вловити або врахувати зміни, що

відбуваються поза сезонністю та трендом, такі як зміна економічних факторів, кризи та інші випадкові події. Крім того, експоненціальне згладжування також дуже чутливе до початкових значень та параметрів згладжування.

Отже, вкрай важливо звернути увагу на правильний вибір цих параметрів при застосуванні методу. У більш складних випадках, коли ряд має складну структуру з різними трендами та сезонністю, може бути доцільним застосування більш складних методів згладжування, таких як ARIMA, SARIMA, або регресійні моделі [19].

Основна формула (2.1) експоненційного згладжування має вигляд:

$$S_{(t)} = \alpha * X_{(t)} + (1 - \alpha) * S_{(t-1)}, \quad (2.1)$$

де $S_{(t)}$ - експоненційно згладжене значення в момент часу t ;

$X_{(t)}$ - спостережене значення в момент часу t ;

$S_{(t-1)}$ - експоненційно згладжене значення в попередньому моменті часу $t-1$;

α - коефіцієнт згладжування, який приймає значення від 0 до 1 і визначає вагу, надану новому спостереженню.

Параметр α визначає ступінь згладжування: низьке значення α призводить до більшого згладжування та більшої ваги попереднього згладженого значення, тоді як високе значення α відображає менше згладжування та більшу вагу нового спостереження. Вибір оптимального значення α залежить від властивостей конкретного ряду і може бути здійснений за допомогою експерименту або застосування методів оптимізації.

Стаціонарність є важливим поняттям для аналізу часових рядів, оскільки більшість моделей базується на стаціонарності ряду. Статистична стаціонарність є одним з основних понять в аналізі часових рядів. Якщо ряд стаціонарний, то його статистичні властивості не змінюються з часом. Іншими словами, середнє, дисперсія та інші статистичні характеристики ряду залишаються постійними в часі [20].

Одним з підходів до визначення стаціонарності ряду є розгляд його часових характеристик, таких як середнє значення та дисперсія. Якщо середнє значення та дисперсія ряду не змінюються з часом, то можна вважати ряд стаціонарним. Інший підхід полягає в аналізі автокореляційних функцій ряду. Якщо автокореляційна функція змінюється з часом, то ряд не є стаціонарним.

Статистична стаціонарність є важливим поняттям для аналізу часових рядів, оскільки більшість моделей базується на стаціонарності ряду. Наприклад, для прогнозування майбутніх значень часового ряду, можна використовувати моделі, що передбачають стаціонарність ряду. Крім того, стаціонарний ряд є більш передбачуваним та менш вразливим до незалежних змін, що сприяє підвищенню точності його прогнозування. Перевірка на стаціонарність є важливим етапом аналізу часових рядів, існує кілька підходів та тестів для перевірки стаціонарності ряду:

- Візуальний аналіз – це спостереження за графіком часового ряду, що може дати певні вказівки на те, чи має ряд стаціонарність. Такий ряд буде мати постійні середнє значення та дисперсію без видимих трендів, сезонності або інших систематичних змін з часом.
- Статистичні тести – це тести, які перевіряють стаціонарність ряду, наприклад: тест ADF - перевіряє наявність одиничного кореня в ряду, що є ознакою нестаціонарності або тест KPSS, що перевіряє гіпотезу про наявність тренду або одиничного кореня в ряду.
- Статистичні показники – цей спосіб перевіряє стаціонарність ряду шляхом використання статистичних показників, таких як середнє значення та дисперсія для порівняння між різними підмножинами часового ряду. Якщо ці показники залишаються сталими або мають деяку регулярну залежність від часу, то це може свідчити про стаціонарність.

Проте, в реальному житті більшість часових рядів не є стаціонарними, тому часто застосовують методи згладжування, щоб зробити їх більш передбачуваними та менш вразливими до змін. Наприклад, метод експоненційного згладжування

може бути використаний для згладжування рядів та вироблення їх стаціонарними або менш нестаціонарними.

Однак, стаціонарність не є єдиним важливим фактором при аналізі часових рядів, інші фактори також можуть бути важливими для прогнозування та моделювання. Наприклад, автокореляція є важливим фактором при аналізі часових рядів. Автокореляція описує ступінь залежності між значеннями ряду в різний момент часу, тобто кореляцію між значеннями ряду та його лагами. Якщо ряд має сильну автокореляцію, це може вказувати на наявність сезонності, тренду чи інших регулярних змін.

Також важливо, коли мова йдеться про прогнозування значень, потрібно враховувати можливу наявність аномальних подій або випадкових варіацій у ряді, таких як шум, тому як, вони можуть впливати на прогнозовані значення. Отже, при вивченні, аналізі та прогнозуванні часових рядів важливо враховувати ряд різних факторів, таких як: стаціонарність, автокореляцію, наявність аномальних подій та випадкових варіацій.

Моделі ARIMA (авторегресійна інтегрована середня модель) є статистичним методом для аналізу та прогнозування часових рядів. Ці моделі використовують попередні значення ряду та їх різницю для прогнозування майбутніх значень. ARIMA складається з трьох основних параметрів: p , d та q .

- Параметр p вказує на кількість попередніх значень, які використовуються в моделі для прогнозування.
- Параметр d вказує на кількість різниць, які потрібно взяти, щоб зробити часовий ряд стаціонарним.
- Параметр q вказує на кількість попередніх значень різниці, які використовуються для прогнозування.

Моделі ARIMA можуть бути використані для прогнозування майбутніх значень ряду, а також для виявлення трендів та сезонності в даних. Для покращення точності прогнозів ARIMA можна поєднувати з експоненційним згладжуванням та іншими методами аналізу часових рядів.

Метод ARIMA - це один з найбільш поширених методів аналізу та прогнозування часових рядів, який знаходить застосування в різних галузях, включаючи фінанси та економіку. Основна ідея методу полягає в тому, що майбутні значення ряду залежать від попередніх значень, різниці між ними та різниці між попередніми різницями. Це дозволяє моделі ARIMA вираховувати тренди, сезонність та інші регулярності в даних.

Параметр p (аргумент авторегресії) відображає вплив попередніх значень ряду на поточне значення, а параметр q (аргумент ковзної середньої) відображає вплив попередніх різниць між значеннями на поточне значення. Параметр d (аргумент інтегрування) відображає необхідність взяти різницю між попереднім та поточним значеннями ряду, щоб зробити ряд стаціонарним.

Для знаходження найкращої моделі ARIMA, необхідно оцінити параметри p , d та q за допомогою методу автокореляції (ACF) та часткової автокореляції (PACF) [21]. Однак, важливо пам'ятати, що ARIMA не є універсальним методом прогнозування, і деякі часові ряди можуть вимагати більш складних моделей, таких як ARIMAX (ARIMA з екзогенними змінними), SARIMA (сезонний ARIMA) та інші.

Математична модель ARIMA:

ARIMA є моделлю, що поєднує авторегресійний (AR), інтегрований (I) та ковзний середній (MA) компоненти. AR (авторегресійна) компонента враховує попередні значення ряду для прогнозування поточного значення. Її загальна формула (2.2) може бути виражена наступним чином:

$$y_t = c + \beta_1 * y_{t-1} + \beta_2 * y_{t-2} + \dots + \beta_p * y_{t-p} + \varepsilon_t, \quad (2.2)$$

де y_t - значення часового ряду на момент часу t ;

c - константа або зсув;

$\beta_1, \beta_2, \dots, \beta_p$ - коефіцієнти авторегресії, які вимірюють вплив попередніх значень на поточне значення;

$y_{t-1}, y_{t-2}, \dots, y_{t-p}$ - попередні значення часового ряду;

ε_t - помилка або шум на момент часу t .

Кожне поточне значення ряду (y_t) прогнозується шляхом комбінації попередніх значень ряду ($y_{t-1}, y_{t-2}, \dots, y_{t-p}$), помножених на відповідні параметри ($\beta_1, \beta_2, \dots, \beta_p$).

МА (ковзний середній) компонента враховує попередні різниці між значеннями ряду для прогнозування поточного значення і має вигляд (2.3) [22]:

$$y_t = \varepsilon_t + y_1 * \varepsilon_{t-1} + y_2 * \varepsilon_{t-2} + \dots + y_q * \varepsilon_{t-q} \quad (2.3)$$

де y_t - поточне значення часового ряду на момент часу t ;

ε_t - поточний шумовий член;

$y_1, y_2, \dots, y_q$ - параметри, що відображають вплив попередніх різниць на поточне значення ряду;

$\varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-q}$ - попередні шумові члени.

Кожне поточне значення ряду (y_t) прогнозується шляхом комбінації попередніх розбіжностей або залишків ($\varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-q}$), помножених на відповідні параметри ($y_1, y_2, \dots, y_q$). Значення ε_t представляє розбіжності або залишки, отримані після використання попередніх різниць.

AR-компонента моделює залежність від попередніх значень ряду, тоді як МА-компонента моделює залежність від попередніх розбіжностей або залишків.

I (інтегрований) компонента використовується для вироблення часового ряду стаціонарним шляхом виконання диференціювання. Параметр d вказує на кількість диференціювань, які необхідно застосувати.

Метрики для оцінювання ефективності моделі ARIMA:

1. Середньоквадратичне відхилення. Вимірює середньоквадратичну різницю між прогнозованими значеннями і спостережуваними значеннями. Менші значення - вказують на кращу точність моделі.
2. Середня абсолютна помилка. Вимірює середню абсолютну різницю між прогнозованими значеннями і спостережуваними значеннями. Ця метрика також оцінює точність моделі, не залежно від напрямку помилки.

3. Коефіцієнт детермінації (R^2). R^2 вказує на частку дисперсії відгуку, яку можна пояснити моделлю. Значення R^2 від 0 до 1, де 1 вказує на ідеальне узгодження між прогнозами і спостережуваними значеннями.

Підходи до налаштування моделі ARIMA:

1. Ідентифікація параметрів p , d та q . Потрібно визначити оптимальні значення параметрів.
2. Перебір параметрів. Щоб визначити оптимальні значення для параметрів, можна скористатися методом перебору. При цьому пробується різні комбінації параметрів та оцінюються метрики ефективності моделі для кожної комбінації, вибираються ті параметри, що мають найкращі результати.
3. Валідація моделі. Щоб оцінити ефективність налаштованої моделі ARIMA, важливо використовувати валідацію на незалежних даних. Для цього можна розділити наявні дані на тренувальний набір і тестовий набір. Модель налаштовується на тренувальному наборі, після чого використовується для прогнозування на тестовому наборі. Метрики ефективності оцінюються на основі порівняння прогнозованих значень зі спостережуваними значеннями на тестовому наборі.
4. Уточнення моделі. Якщо під час валідації виявляються недоліки моделі ARIMA, можна скористатися додатковими методами для поліпшення результатів.
5. Оцінка статистичної значущості. Після налаштування моделі важливо оцінити статистичну значущість коефіцієнтів моделі це допоможе зрозуміти, які коефіцієнти мають суттєвий вплив на прогнози. Для цього можна використовувати статистичні тести, такі як t-тест або статистика Вальда.
6. Перевірка залишків. Після побудови та оцінки моделі ARIMA важливо перевірити залишки (рештки) моделі на відповідність стохастичним властивостям. Це можна зробити шляхом аналізу автокореляційної функції залишків та перевірки на наявність автокореляції або гетероскедастичності. Якщо знайдені недоліки, можна застосовувати додаткові методи для моделювання та виправлення цих недоліків.

Нарешті, ARIMA може бути поєднана з іншими методами аналізу часових рядів, такими як експоненційне згладжування, щоб забезпечити більш точні прогнози. Наприклад, модель ARIMA може бути поєднана з моделлю експоненційного згладжування, щоб створити модель ARIMAX (авторегресійна інтегрована середня модель з екзогенними змінними).

ARIMAX додає додаткові змінні, які не є частиною аналізованого часового ряду, але можуть впливати на нього. Наприклад, якщо аналізується продаж продукту, то ціна та реклама можуть бути використані як екзогенні змінні, оскільки вони можуть впливати на кількість продажів.

Крім ARIMAX, існують інші моделі, які можуть бути використані для аналізу та прогнозування часових рядів. Наприклад, такі як:

- Модель звичайного експоненційного згладжування (SES);
- Модель згладжування Хольта (Holt);
- Модель згладжування Хольта-Вінтерса (Holt-Winters).

Кожна з цих моделей має свої власні переваги та недоліки, а їх вибір залежить від конкретної ситуації та властивостей аналізованого часового ряду [23].

2.3 Багатофакторне моделювання динаміки обсягів реалізації товарів

Складні математичні моделі, що зветься нейронними мережами, складаються зі штучних нейронів, які моделюють функціонування людського мозку. Нейронні мережі можуть вивчати складні залежності між вхідними даними та вихідними результатами, використовуючи техніку навчання з учителем або без учителя.

Однією з ключових переваг нейронних мереж є їх здатність до автоматичного виявлення корисних ознак в даних, що знижує необхідність вручну вибирати ознаки для подальшого аналізу даних [24]. Нейронні мережі можуть бути розглянуті як універсальний інструмент для розв'язання різноманітних задач

машинного навчання, зокрема класифікації, регресії, кластеризації, сегментації, генерації тексту та зображень, машинного перекладу, рекомендаційних систем та інших.

У нейронних мережах, навчених з учителем, вхідні дані та відповідні їм правильні відповіді (так звані мітки) використовуються для навчання моделі. Після того, як модель навчилася на тренувальному наборі даних, її можна використовувати для передбачення вихідних значень на нових, раніше невідомих даних.

Кожен нейрон в мережі приймає вхідні дані, виконує деяку операцію на них та видає вихідний результат. Ці операції можуть включати додавання та множення вхідних даних, застосування активаційної функції та інші операції. На основі цих операцій мережа вчиться виконувати певну задачу, шляхом знаходження оптимальних параметрів моделі [25].

LSTM (Довготривала Короткочасна Пам'ять) - це спеціальний тип нейронних мереж, призначений для обробки послідовностей даних. Одна з її особливостей полягає в тому, що вона може запам'ятовувати попередні значення деяких її елементів та використовувати цю інформацію при обробці наступних значень вхідних даних.

LSTM має комплексну структуру, що складається зі спеціальних блоків пам'яті, які дозволяють зберігати та забувати інформацію в процесі обробки даних. Кожен блок містить три гейти (ворота) - "вхідний", "вихідний" та "забування", які регулюють потік інформації в середині блоку.

1. Вхідний гейт визначає, які значення вхідних даних потрібно зберегти в блоку пам'яті.
2. Забувальний гейт відповідає за видалення певної інформації з блоку пам'яті.
3. Вихідний гейт контролює, які значення потрібно використовувати для виведення результату на вихід.

Таким чином, за допомогою цих гейтів LSTM може вирішувати проблему зниклого градієнту, що виникає у звичайних рекурентних мережах під час тренування. Застосування LSTM дозволяє розв'язувати завдання, пов'язані з

аналізом тексту та голосу, такі як автоматичний переклад, генерація тексту, розпізнавання мови та синтез мови, LSTM може бути використана в сферах обробки медичних даних, фінансових прогнозів та багатьох інших галузях [30].

Нейронні мережі, зокрема LSTM, можуть бути дуже корисні для прогнозування. Один з популярних прикладів застосування LSTM - прогнозування цін на акції. Вхідні дані можуть включати ціну акцій в минулому, індекси фондового ринку, новини, які можуть вплинути на ціну акцій, та інші фактори. Нейронна мережа з LSTM шарами зможе вивчити складні залежності між цими вхідними даними та змінами в ціні акцій.

Прогнозування цін на акції - це задача регресії, тому результатом роботи нейронної мережі будуть прогнозовані значення цін. При навчанні нейронної мережі з LSTM шарами для прогнозування цін на акції, можна використовувати різні метрики для оцінки якості прогнозів, такі як середня квадратична помилка, середня абсолютна помилка та інші.

Іншими важливими методами моделювання та прогнозування динаміки обсягів реалізації товарів є лінійна регресія (linear regression) та VAR (Vector Autoregression).

Регресійний аналіз є одним з основних методів прогнозування та моделювання залежностей між змінними. У регресійному аналізі прагнуть побудувати лінійну або нелінійну модель, що найкраще описує залежність між вхідними факторами та вихідною змінною (обсягами реалізації товарів). У випадку регресії, функціональна залежність виражається у вигляді математичного рівняння, де параметри моделі підбираються таким чином, щоб мінімізувати помилки між спостережуваними та прогнозованими значеннями.

Регресійний аналіз можна використовувати для моделювання динаміки обсягів реалізації товарів на основі різних факторів, таких як ціни, рекламні витрати, попит, сезонність тощо. Після побудови регресійної моделі можна використовувати її для прогнозування майбутніх обсягів реалізації на основі нових значень факторів.

VAR є статистичною моделлю, яка дозволяє моделювати та прогнозувати динаміку кількох залежних змінних одночасно. Модель VAR базується на ідеї, що кожна змінна у системі залежить від попередніх значень всіх змінних. Змінні у VAR моделі утворюють вектор, та за допомогою статистичних методів оцінюються параметри моделі, такі як коефіцієнти регресії. Після оцінки параметрів VAR моделі можна здійснити прогнозування майбутніх значень кожної змінної у системі.

VAR модель особливо ефективна, якщо змінні у системі взаємозалежні між собою. Наприклад, у випадку обсягів реалізації товарів можуть впливати на ціни, рекламні витрати та інші фактори, а також можуть взаємодіяти між собою. Модель VAR дозволяє врахувати цю взаємодію та залежність між змінними, що призводить до більш точних прогнозів.

Процес побудови VAR моделі включає вибір кількості лагів (попередніх значень), визначення оптимальної кількості змінних у системі та оцінку параметрів моделі. Після цього можна використовувати VAR модель для прогнозування майбутніх значень обсягів реалізації товарів на основі попередніх значень змінних.

Ще одним важливим методом моделювання та прогнозування є Random Forest (випадковий ліс). Random Forest є ансамблевим методом машинного навчання, який базується на рішучих деревах. Він працює шляхом побудови багатьох дерев рішень (лісу) із випадково вибраними підмножинами даних та ознак. Кожне дерево у лісі незалежно приймає прогнози, та після цього результуючий прогноз визначається шляхом голосування або усереднення результатів дерев.

Для моделювання динаміки обсягів реалізації товарів за допомогою Random Forest, можна використовувати різні фактори, такі як ціни, рекламні витрати, попит, сезонність і т.д. Random Forest може автоматично виявити та врахувати важливість кожного фактора у прогнозуванні.

Процес побудови Random Forest включає вибір кількості дерев, кількості випадково вибраних ознак та інших параметрів. Після навчання моделі на

наявних даних, можна використовувати її для прогнозування майбутніх обсягів реалізації товарів на основі нових значень факторів [26].

Random Forest є потужним та гнучким методом, який часто використовується в прогнозуванні обсягів реалізації товарів через свою здатність до моделювання складних взаємозв'язків між факторами та точне прогнозування.

РОЗДІЛ 3

ПРАКТИЧНЕ ЗАСТОСУВАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ОБСЯГІВ ПРОДАЖІВ ТОВАРІВ

3.1 Розвідувальний аналіз та декомпозиція системи часових рядів, що визначають динаміку продажів товарів у торгівельній мережі

У цій роботі ми моделюємо динаміку продажів магазинів мережі Walmart. Мережа цих супермаркетів є однією з найбільших торговельних ланцюгів у світі. Компанія працює в різних країнах, включаючи США, Канаду та Мексику. Walmart - це одна з найбільших роздрібних мереж у світі, з більш ніж 11.000 магазинами в різних країнах світу. Велика кількість магазинів, різноманітність товарів та залежність від зовнішніх факторів роблять Walmart прикладом компанії, яка вдало використовує машинне навчання для оптимізації своєї діяльності та є ідеальним об'єктом дослідження для моделювання динаміки продажів. В рамках цього практичного дослідження ми використовуємо датасет, що містить історичні дані про обсяги прибутку з 05.02.2010 до 01.11.2012. Даний датасет складається з різних атрибутів, що включають інформацію про магазин, дату продажу, прибуток, позначку святкового тижня та деякі економічні показники.

Розглянемо деталі датасету:

- Атрибут - "Store" - є номером ідентифікації для кожного окремого магазину. Цей атрибут дозволяє нам відслідковувати продажі для кожного магазину окремо.
- Атрибут - "Date" - вказує на тиждень, коли відбулися продажі. Цей атрибут дозволяє нам аналізувати динаміку продажів протягом певного періоду.
- "Weekly_Sales" є грошовим показником в доларах, який відображає обсяги продажу для кожного магазину протягом певного тижня. Цей атрибут є

цільовою змінною, яку ми намагаємося передбачити за допомогою нашої моделі.

- Атрибут "Holiday_Flag" вказує на особливість тижня. Якщо значення цього атрибуту дорівнює 1, це означає, що цей тиждень є святковим, наприклад, вихідним. Значення 0 вказує на те, що тиждень є звичайним.
- Температура записана у атрибуті "Temperature". Цей атрибут вказує на середню температуру в регіоні, де знаходиться магазин, в той тиждень, коли відбулися продажі.
- "Fuel_Price"- це атрибут, який вказує на вартість пального в регіоні, де розташований магазин. Вартість палива може впливати на споживчу активність та витрати покупців, тому цей атрибут може мати вплив на обсяги продажу.
- Атрибут "CPI" (Переважаючий індекс споживчих цін) надає інформацію про рівень споживчих цін в регіоні, де знаходиться магазин. Цей показник може впливати на споживчу поведінку та прийняття рішень щодо покупок, тому він є важливим для розуміння динаміки обсягів продажу.
- "Unemployment" - це атрибут, який вказує на рівень безробіття в регіоні, де знаходиться магазин. Він може слугувати показником економічної стабільності та споживчої активності, що можуть впливати на обсяги продажу.

Використовуючи цей датасет, ми досліджуємо показник "Weekly_Sales" та вплив інших чинників на нього за допомогою інструментів машинного навчання. Цей аналіз допоможе нам отримати більш глибоке розуміння факторів, що впливають на продажі, і може допомогти управлінцям та бізнес-аналітикам зробити кращі рішення щодо стратегій продажу та планування запасів.

Перевірити наші дані можна за допомогою наступних кроків:

- Використовуючи функцію `head()`, можна переглянути перші кілька рядків датасету. Це дозволяє отримати загальне уявлення про дані і переконатися, що датасет правильно завантажено.

- За допомогою функції `dim()`, можна отримати кількість рядків та стовпців у датасеті. Це дає загальне уявлення про його розмір та обсяг даних.
- Функція `str()` дозволяє переглянути типи даних кожного стовпця у датасеті. Це важливо для подальшого розуміння даних і визначення, які атрибути потребуватимуть подальшої обробки.

Перевірка даних на пропуски та дублікати є важливим кроком при аналізі датасету для забезпечення якості та достовірності даних. Перевірка даних на пропуски зазвичай включає обчислення кількості пропусків у кожному стовпці датасету. Це може бути зроблено за допомогою функцій, які перевіряють, чи є значення відсутніми та підраховують кількість пропусків у кожному стовпці. За допомогою функції `colSums(is.na())`, можна підрахувати кількість пропусків у кожному стовпці датасету. Це допоможе виявити атрибути з відсутніми значеннями, які потребуватимуть обробки [27].

Дублікати можна виявити, порівнюючи кожен запис з усіма іншими записами у датасеті. Якщо існують повтори, то це означає наявність дублікатів. Для перевірки наявності дублікатів зазвичай використовують функцію `duplicated()`, яка повертає логічне значення (`TRUE` або `FALSE`) для кожного запису, вказуючи, чи є він дублікатом. Це важливо, оскільки дублікати можуть спотворити аналіз і призвести до неправильних результатів.

Побудова та опис базових візуалізацій є важливим етапом практичного дослідження для отримання візуального уявлення про дані та їх розподіл. Ми розглянемо побудову базових візуалізацій, які дозволять нам оцінити розподіл продажів серед магазинів протягом періоду дослідження [28].

Досліджуємо наявність сезонності, трендів та інших закономірностей. В рамках параграфу можна побудувати графік залежності тижневих прибутків від часу (рис. 3.1). Це дозволить нам оцінити наявність регулярностей, сезонності та трендів у динаміці продажів.

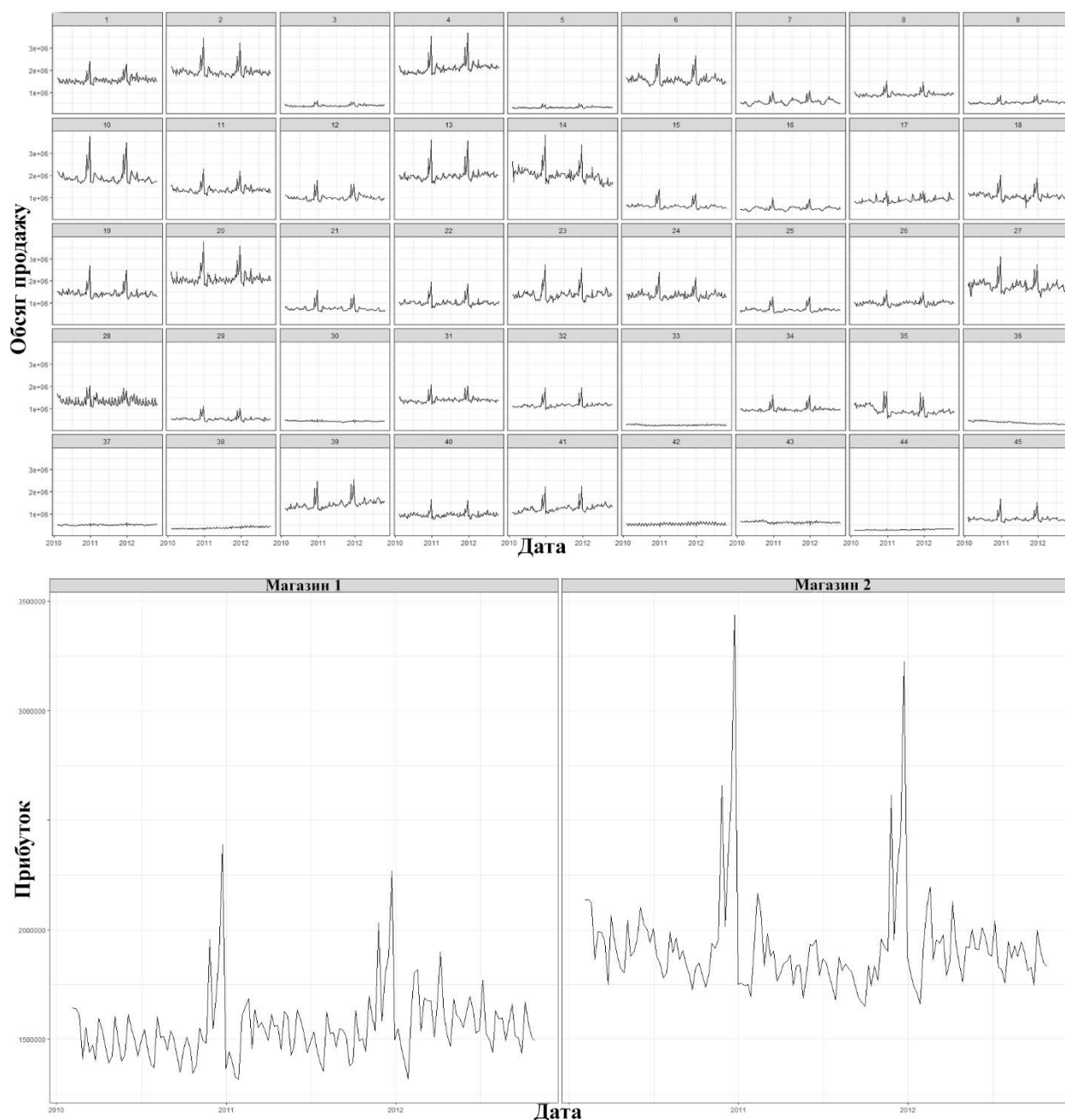


Рисунок 3.1 – Графік часового ряду продажів Walmart

Джерело: розроблено автором

Результатом цього коду, додаток А.1, є графік, який відображає залежність тижневих прибутків від часу. Використовуючи різні шкали, ми маємо змогу оцінити наявність регулярностей та закономірностей у динаміці прибутків. У даному випадку графік свідчить про наявність періодичності та сезонності, оскільки є тенденції, які повторюються на певних інтервалах часу. Також, існують

певні періоди, коли прибутки регулярно зростають або падають, це може бути ознакою сезонності.

Гістограма є графічним інструментом для відображення розподілу числових даних. В нашому випадку, ми будемо будувати гістограму для стовпця "Weekly_Sales", що вказує на прибуток для кожного магазину в певний тиждень. Для побудови гістограми ми використовуємо бібліотеку ggplot2, яка надає потужні інструменти для створення графіків у R. Ми вказуємо Walmart як джерело даних, а Weekly_Sales як змінну, яку ми хочемо відобразити на осі X. Параметр bins визначає кількість стовпчиків у гістограмі. Результатом є побудова гістограми, яка відображає розподіл тижневих прибутків (рис. 3.2). Вона дозволить нам оцінити частоту продажів у різних діапазонах значень та зробити припущення про їх розподіл, додаток А.2.

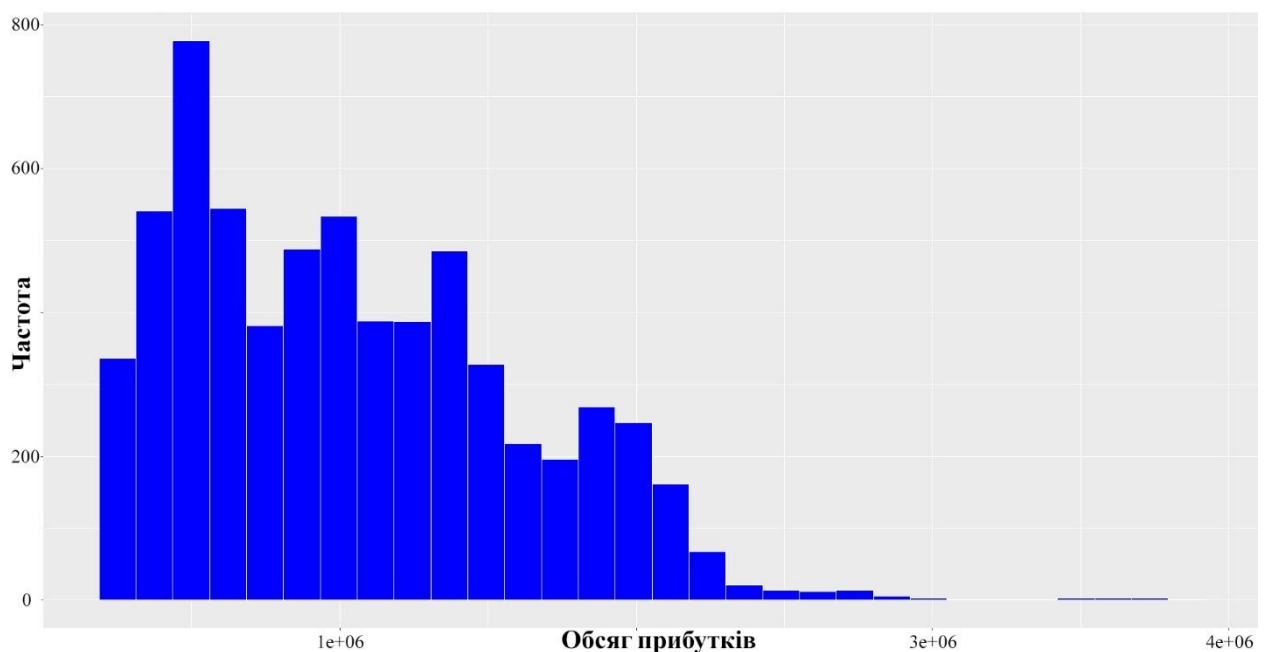


Рисунок 3.2 – Гістограма розподілу тижневих прибутків

Джерело: розроблено автором

Згідно до побудованого графіку, можна спостерігати, що розподіл тижневих прибутків у Walmart має правосторонню скошеність (позитивне скошення). Це означає, що більші значення обсягу продажу відбуваються меншою частотою, а менші значення - більшою частотою. Графік показує, що більшість тижневих

доходів знаходиться в діапазоні від, приблизно, 0 до 1,5 млн. доларів. Доходи, що перевищують цей діапазон, стають все менш поширеними. Деякі тижні мають надзвичайно високі обсяги продажу, що припадають на декілька ізольованих стовпчиків на правому кінці графіка. Це можуть бути тижні зі святковими подіями або спеціальними акціями, що сприяють значному збільшенню продажів.

"Ящик з вусами" (або boxplot) є графічним інструментом, який відображає статистичну інформацію про розподіл числових даних, таких як мінімум, нижній квартиль, медіана, верхній квартиль та максимум. Використання "ящика з вусами" дозволяє нам зрозуміти центральні тенденції та розмах значень у кожному магазині. Розглянемо побудову "ящика з вусами" для кожного магазину, що дозволить нам оцінити розподіл прибутковості серед окремих магазинів протягом усього періоду дослідження, додаток А.3.

В результаті ми отримуємо графіки "ящиків з вусами" (рис. 3.3), який відображає розподіл тижневих продажів для кожного магазину. Цей графік допоможе нам зрозуміти розмах значень, центральні тенденції та наявність святкових тижнів в продажах для кожного магазину.

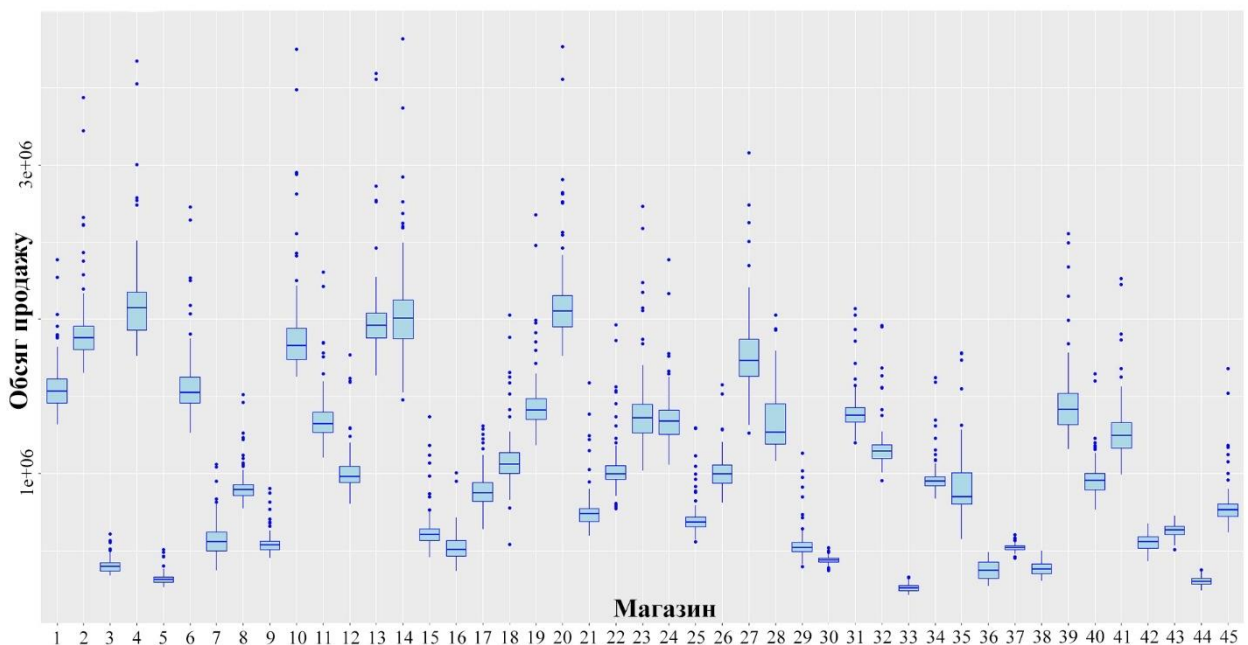


Рисунок 3.3 – Графік розподілу тижневих продажів для кожного магазину

Джерело: розроблено автором

За допомогою ящика з вусами було проаналізовано розподіл доходів для кожного окремого магазину. Зауважується, що медіани для кожного магазину відрізняються значно. Для деяких магазинів, таких як 4, 14 та 21, медіана прибутковості вища, ніж у інших магазинів, що може свідчити про більшу популярність цих магазинів серед споживачів. З іншого боку, для деяких магазинів, таких як 28, 35, медіана значно нижча, що може вказувати на меншу популярність цих магазинів серед споживачів. Також зауважується, що діапазон продажів для кожного магазину значно відрізняється, що може бути пов'язано з різними факторами, такими як розташування магазину, рівень конкуренції та інші.

Дослідимо залежність між температурою та доходами в магазинах, додаток А.4. Якщо побудувати графік кореляції, то можна спостерігати, що існує деяка позитивна кореляція між цими двома факторами, що означає, що високі температури зазвичай супроводжуються більшими доходами. Однак, також можна побачити, що є деякі викиди, де прибуток є високим при низьких температурах і навпаки. Загалом можна зробити висновок, що температура має вплив на продажі, але інші фактори, такі як свята, паливна вартість, безробіття і т.д. також можуть вплинути на продажі і мають бути враховані при аналізі.

Розглянемо графіки для порівняння продажів у святковій та звичайній тиждні. Для початку, ми створимо новий стовпчик "Holiday_Type", що визначатиме тип святкового тижня. Якщо тиждень є святковим, то значення буде "Holiday", в іншому випадку - "Non Holiday". Це допоможе нам категоризувати дані та порівняти прибутковість у різних типах тижнів. Код для створення стовпчика "Holiday_Type". У цьому коді, додаток А.5, ми використовуємо умовне вираз `ifelse()`, який перевіряє значення стовпчика "Holiday_Flag". Якщо значення рівне 1 (святковий тиждень), то присвоюється значення "Holiday", в іншому випадку - "Non Holiday". Будуємо графік порівняння продажів, додаток А.6 (рис. 3.4-3.5):

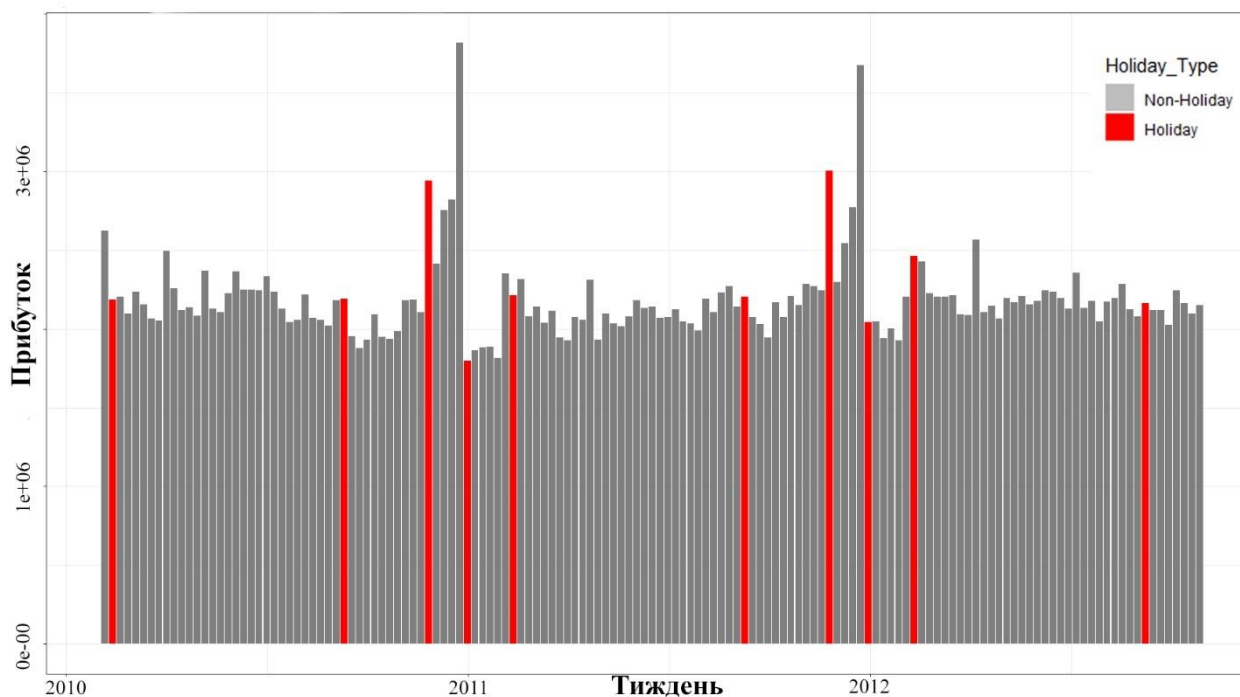


Рисунок 3.4 – Прибутковість у святковій та несвятковій тиждні

Джерело: розроблено автором

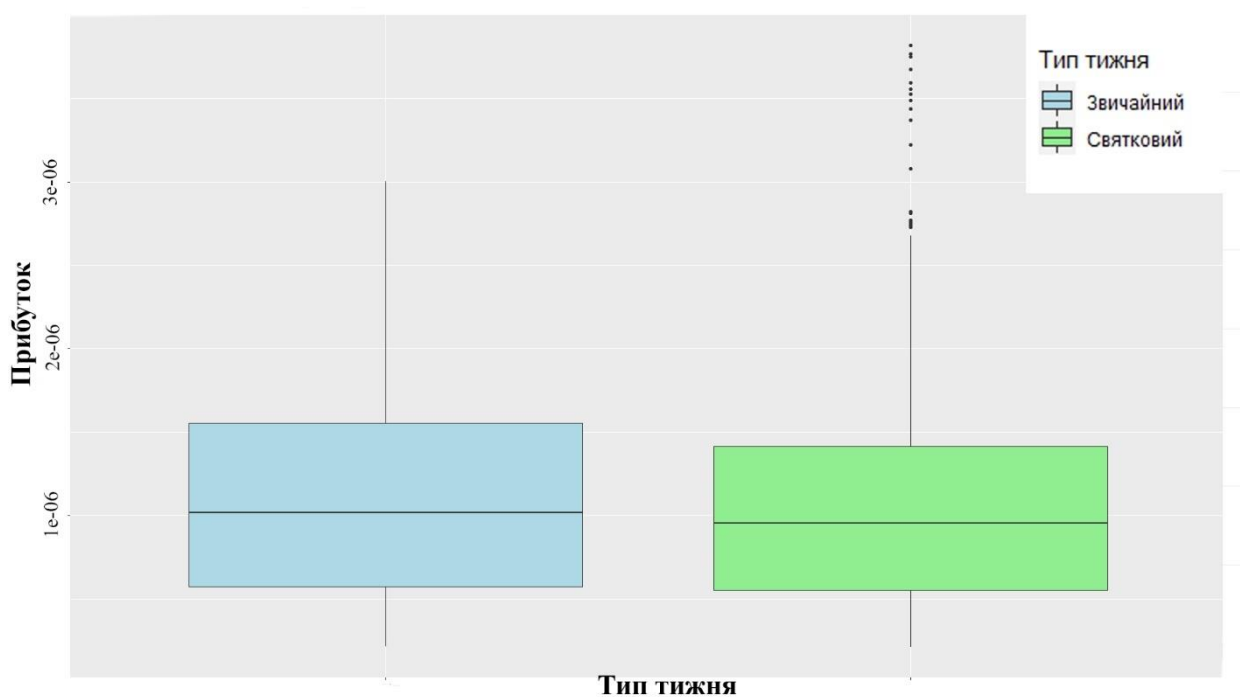


Рисунок 3.5 – Прибутковість у святковій та несвятковій тиждні

Джерело: розроблено автором

Проаналізувавши рис. 3.4 можна зробити висновок, що немає значимих статистичних різниць у тижневих продажах між святковими та несвятковими

періодами. Графік не показує систематичний зріст або спад прибутків у святкові тижні порівняно з несвятковими. Для того щоб достовірніше упевнитися у цьому, розглянемо рис. 3.5. З даного графіку видно, що медіана прибутковості, на не святковому тижні вища, ніж у святкового. Крім того, розмах (відстань між нижнім та верхнім кuartилями) також більший ніж у святкового періоді, що свідчить про більшу варіативність прибутків. Даний аналіз дає нам розуміння, що святкові дні не мають значущого впливу на продажі в магазинах Walmart.

Проведемо аналіз залежності між безробіттям та обсягами продажів, додаток А.7. Для проведення аналізу потрібно побудувати діаграму розсіювання, щоб візуалізувати вплив безробіття на продажі.

Завдяки побудованому графіку, можна визначити вплив одного фактора на інший, цей метод є доволі ефективний щоб спостерігати залежність. У даному випадку діаграма показала, що немає чіткої залежності між безробіттям та прибутковістю, але можна помітити, що є декілька великих значень прибутковості в тих тижнях, коли безробіття є найнижчим, а також є кілька тижнів з високим рівнем безробіття, де прибутковість також висока. Це може свідчити про те, що безробіття не є головним фактором, який визначає рівень прибутковості.

Дослідимо залежності між вартістю пального та прибутковістю. Побудуємо діаграму розсіювання, щоб візуалізувати цю залежність, код аналогічний попередній візуалізації додаток А.8. Результатом цього коду буде діаграма розсіювання, яка візуалізує залежність між вартістю пального та обсягами продажів.

Аналізуючи цю залежність можна стверджувати, що ці показники не залежать один від одного, зі збільшенням ціни на паливо та навпаки, обсяг продажу не змінюється, та ніяк не впливає на прибуток магазинів.

3.2 Аналіз результатів згладжування та ARIMA - моделювання динаміки продажів товарів

Для дослідження більш детальних закономірностей можна також використовувати інші методи аналізу часових рядів, наприклад ARIMA. Щоб застосувати ARIMA, нам потрібно спочатку побудувати графік часового ряду продажів. З його допомогою можна отримати загальне уявлення про характеристики ряду і визначити наявність тренду, сезонності або інших закономірностей, які потребують подальшого дослідження з використанням ARIMA та інших методів аналізу часових рядів.

Для застосування моделі ARIMA нам потрібно спочатку визначити значення параметрів p , d та q . Ми можемо визначити значення параметрів, використовуючи автоматичне підбір значень за допомогою функції `auto.arima`, додаток Б.1. Ми перетворюємо стовпець `Weekly_Sales` в часовий ряд `ts_sales` з частотою 52 (оскільки в одному році є приблизно 52 тижні). Використовуємо функцію `auto.arima()` для автоматичного підбору оптимальної моделі ARIMA.

Наступним кроком, ми можемо розглянути наступні підходи:

1. Аналіз компонентів. Використовуючи функцію `decompose()`, можна розкрити часовий ряд на його основні компоненти. Це дозволить окремо досліджувати кожну складову та їх взаємодію.
2. Аналіз автокореляції. Використовуючи функцію `acf()` (рис. 3.7) для вивчення автокореляційних зсувів у часовому ряді, можна виявити наявність сезонності та інших взаємозв'язків між значеннями ряду у різних зсувах часу.

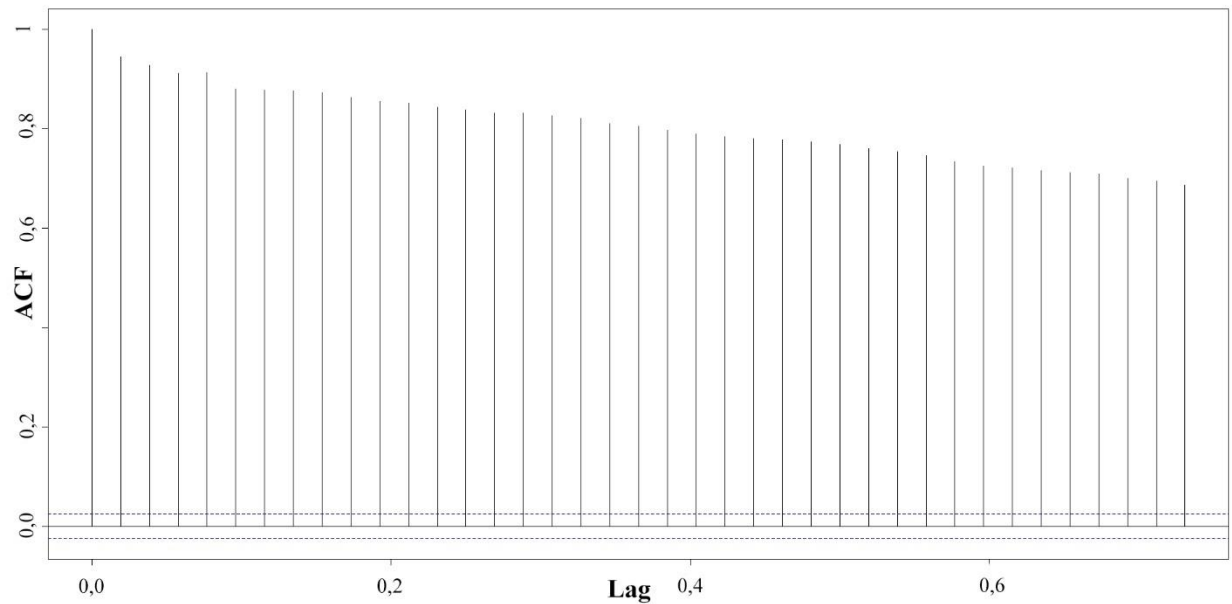


Рисунок 3.7 – Аналіз автокореляції

Джерело: розроблено автором

Загалом, графік демонструє наявність статистично значущої позитивної автокореляції. Це означає, що значення ряду з певним зсувом назад у часі корелюють із значеннями, що знаходяться дещо далі в минулому. Також, не спостерігається сильного зворотного зв'язку або негативної автокореляції між значеннями ряду та їх зсувами у часі.

3. Параметри моделі. При розгляді параметрів ARIMA, ми можемо варіювати значення p , d та q та спостерігати за змінами в моделі та її прогнозах.
4. Аналіз часткових автокореляційних зсувів. Функцію `pacf()` для побудови графіку часткових автокореляційних зсувів. Цей графік допомагає виявити наявність авторегресії та допоміжних лагів, що вказують на тренди та інші закономірності (рис. 3.8).

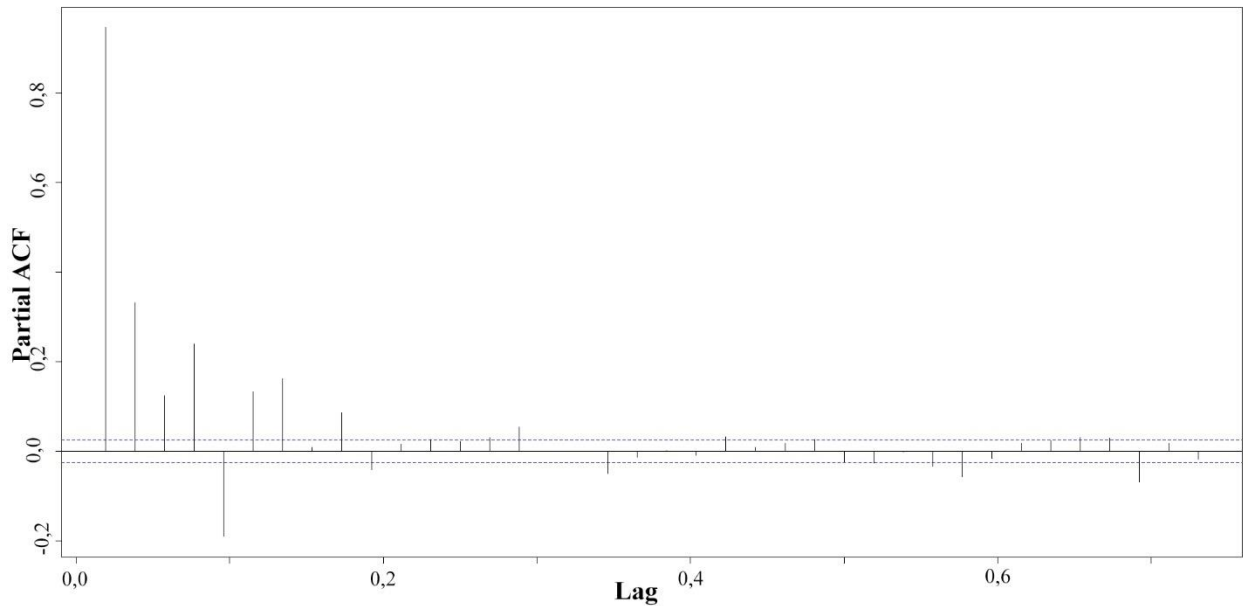


Рисунок 3.8 – Аналіз часткових автокореляційних зсувів

Джерело: розроблено автором

5. Перевірка ряду на стаціонарність. Перед застосуванням моделі ARIMA варто додати перевірку на стаціонарність часового ряду. Для перевірки коду на стаціонарність скористаємося тестом Дікі-Фуллера, додаток Б.2. У нашому випадку (рис. 3.9), згідно з результатами тесту Дікі-Фуллера нульова гіпотеза про нестаціонарність ряду була відхилена на рівні значущості 0.01. Отже, на основі цього тесту можна припустити, що ряд є стаціонарним.

Augmented Dickey-Fuller Test

```
data: ts_sales
Dickey-Fuller = -5.1728, Lag order = 18, p-value = 0.01
alternative hypothesis: stationary
```

Рисунок 3.9 – Перевірка на стаціонарність ряду

Джерело: розроблено автором

Для оцінки якості моделі ARIMA та її результатів, можна використати деякі загальні метрики, а саме:

- Середню абсолютну помилку – це метрика, яка вимірює середню абсолютну різницю між прогнозованими значеннями моделі та спостережуваними

значеннями в часовому ряді. Нижнє значення вказує на кращу точність моделі – 0 означає ідеальне співпадіння прогнозованих і спостережуваних значень.

- Середня квадратична помилка – це метрика, яка вимірює середню квадратичну різницю між прогнозованими та спостережуваними значеннями. Чим нижче значення – тим краще модель.
- Коефіцієнт детермінації – це показник, який вказує на частку варіації залежної змінної, яку можна пояснити моделлю. R^2 варіюється від 0 до 1, де значення ближче до 1 означає кращу відповідність моделі спостережуваним даним.
- Середня абсолютна відсоткова помилка - це метрика, яка вимірює середню відсоткову різницю між прогнозованими та спостережуваними значеннями, використовується для вимірювання відносної точності моделі, особливо коли масштаби даних відрізняються. Чим нижче значення – тим краще модель.
- Інформаційний критерій Акаїке - це метрика, яка використовується для порівняння моделей ARIMA з різними значеннями параметрів. Модель з меншим значенням вважається кращою, оскільки вона краще враховує складність моделі та робить компроміс між точністю та складністю.

Перевіримо нашу модель (рис. 3.10):

```
Series: ts_sales
ARIMA(5,1,5)(1,0,2)[52] with drift

Coefficients:
      ar1      ar2      ar3      ar4      ar5      ma1      ma2      ma3      ma4      ma5      sar1      sma1      sma2
s.e.    0.1509    0.226    0.0723    0.1326    0.0250    0.2158    0.1166    0.1270    0.0476    0.2334    0.4717    0.0936    0.2155
      drift
      -145.7323
s.e.    1113.1551

sigma^2 = 2.255e+10: log likelihood = -85817.73
AIC=171665.5   AICc=171665.5   BIC=171767

Training set error measures:
      ME      RMSE      MAE      MPE      MAPE      MASE      ACF1
Training set 11.90817 150004.1 80827.45 -1.171974 8.288598 0.3024579 0.000283451
>
```

Рисунок 3.10 – Перевірка моделі за метриками

Джерело: розроблено автором

- Коефіцієнти авторегресії та коефіцієнти рухомого середнього;
- Сезонні AR та МА-коефіцієнти (sar1 і sma1) вказують на наявність сезонної складової з періодичністю 52 (щотижневі дані);

- Модель також включає зрушення (drift), що враховує постійну тенденцію з часом;
- Оцінена дисперсія (σ^2) становить $2.255e+10$;
- Логарифмічна подібність моделі (log likelihood) дорівнює -85817.73;
- Критерій інформаційної Акаїке (AIC) становить 171665.5, скоригований AIC (AICc) також 171665.5, і критерій Байєса (BIC) становить 171767. Ці критерії можна використовувати для вибору моделі, де менші значення вказують на кращу адаптацію моделі до даних.

В числовому вигляді запис моделювання моделі ARIMA наведена у формулі (3.1) [29].

(3.1)

ARIMA(5, 1, 5):

$$AR(5): Y_{(t)} = c + \sum(\varphi_{(i)} * Y_{(t-i)}) + \varepsilon_{(t)}$$

$$I(1): \Delta Y_{(t)} = Y_{(t)} - Y_{(t-d)}$$

$$MA(5): Y_{(t)} = \mu + \sum(\theta_{(i)} * \varepsilon_{(t-i)}),$$

де: $Y_{(t)}$ - значення ряду в момент часу t ;

c - константа;

$\varphi_{(i)}$ - коефіцієнти авторегресії (AR) з лагами i ;

$\varepsilon_{(t)}$ - шумовий член (випадкова помилка) в момент часу t ;

$\Delta Y_{(t)}$ - зміна ряду (диференційований ряд);

μ - середнє значення ряду;

$\theta_{(i)}$ - коефіцієнти ковзного середнього (MA) з лагами i .

Використаємо ARIMAX. Для застосування цієї моделі в мові R, знадобиться пакет forecast. Далі, нам потрібно підключити екзогенні змінні, для цього потрібно використовувати функцію `arima()` з аргументом `xreg`. Цей код, додаток Б.3, використовує екзогенні змінні `Temperature`, `Fuel_Price`, `CPI` та `Unemployment` у моделі ARIMA для кращого моделювання часового ряду. Наступним кроком, нам потрібно витягнути параметри p , d , q , для того щоб дізнатися ці параметри подивимося функцію `summary()` для об'єкта моделі ARIMA. Отже, в даному випадку:

$p = 5$: Це показує, що модель використовує 5 попередніх значень авторегресії для прогнозування поточного значення.

$d = 1$: Це показує, що часовий ряд був підданий одноразовому диференціюванню, щоб зробити його стаціонарним.

$q = 5$: Це показує, що модель використовує 5 попередніх значень ковзаючого середнього для прогнозування поточного значення.

Для оцінки точності моделі ARIMAX використовується функція `assurasy()`, що обчислює різні статистичні показники, такі як середня абсолютна помилка, середня квадратична помилка, середня абсолютна відсоткова помилка тощо, для оцінки якості прогнозування моделі.

`residuals()` - отримує залишкові (рештки) моделі ARIMAX. Залишкові є різницею між фактичними значеннями часового ряду і прогнозованими значеннями, отриманими в результаті застосування моделі.

Для оцінки якості моделі ARIMAX використовується також метрики як і у ARIMA, але ще додається показник залишколовсті моделі.

3.3 Аналіз результатів багатofакторного моделювання обсягів продажів товарів

Щоб оцінити коефіцієнти моделі багатofакторної регресії, потрібно використати метод найменших квадратів. В R можна використати функцію `lm()` для побудови моделі регресії та отримання оцінок коефіцієнтів. У нашому випадку модель регресії побудована залежно від змінних `Temperature`, `Fuel_Price`, `CPI` та `Unemployment`, а залежна змінна - `Weekly_Sales`. Функція `lm()` повертає об'єкт моделі, і за допомогою `coef()` ми маємо оцінки коефіцієнтів моделі (рис. 3.11).

```
(Intercept)  Temperature  Fuel_Price      CPI  Unemployment
1743607.6200    -885.6699  -12248.4245   -1585.8180   -41214.9873
> |
```

Рисунок 3.11 – Отримані оцінки коефіцієнтів моделі

Джерело: розроблено автором

Ці значення вказують на очікувані зміни у залежній змінній `Weekly_Sales` при зміні відповідних незалежних змінних. Наприклад, кожне одиниця збільшення значення `Temperature` пов'язане зі зниженням середнього тижневого обсягу доходу на 885.6699 одиниць. Аналогічно, збільшення `Fuel_Price`, `CPI` та `Unemployment` пов'язані зі зменшенням середнього тижневого обсягу прибутковості.

Для визначення значимості факторів у моделі, можна використати статистичні тести, такі як t-тест або аналіз дисперсії (ANOVA). Т-тест застосовується щоб перевірити значимість окремого фактора. Аналіз дисперсії (ANOVA) – якщо потрібно оцінити значимість декількох факторів одночасно та їх взаємодію. Застосуємо аналіз дисперсії (ANOVA). Наприклад, візьмемо результати для фактора "Температура" (`Temperature`) у моделі залежності від "Тижневих продажів" (`Weekly_Sales`) (рис. 3.12).

```
> print(anova_temperature)
Call:
  aov(formula = weekly_sales ~ Temperature, data = walmart)

Terms:
              Temperature      Residuals 
Sum of Squares  8.344136e+12  2.040947e+15 
Deg. of Freedom           1           6433 

Residual standard error: 563260.3
Estimated effects may be unbalanced
> |
```

Рисунок 3.12 – Отримані результати ANOVA

Джерело: розроблено автором

Загальна сума квадратів між групами, пов'язаними з фактором "Температура" (`Temperature`), становить $8.344136e+12$, а сума квадратів залишків (`Residuals`) - $2.040947e+15$. Також надається інформація про ступені свободи та оцінку стандартної помилки залишкових величин. Для отримання значень F-

статистики та р-значення нам потрібно використати функцію `summary()` на результаті аналізу ANOVA (рис. 3.13). У кожному результаті виводяться значення F-статистики та р-значення. Ці значення вказують на статистичну значимість впливу факторів на обсяги продажів у Walmart. Якщо р-значення менше встановленого порогового значення альфа (зазвичай 0.05), це вказує на статистичну значимість впливу фактора. У нашому випадку, всі фактори, крім `anova_fuel_price`, мають дуже низькі р-значення, що означає, що вони мають статистично значущий вплив на обсяги продажів.

```
> summary(anova_temperature)
      Df Sum Sq Mean Sq F value Pr(>F)
Temperature 1 8.344e+12 8.344e+12  26.3 3.01e-07 ***
Residuals 6433 2.041e+15 3.173e+11
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(anova_fuel_price)
      Df Sum Sq Mean Sq F value Pr(>F)
Fuel_Price 1 1.835e+11 1.835e+11  0.576 0.448
Residuals 6433 2.049e+15 3.185e+11
> summary(anova_cpi)
      Df Sum Sq Mean Sq F value Pr(>F)
CPI 1 1.081e+13 1.081e+13  34.12 5.44e-09 ***
Residuals 6433 2.038e+15 3.169e+11
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(anova_unemployment)
      Df Sum Sq Mean Sq F value Pr(>F)
Unemployment 1 2.310e+13 2.31e+13  73.35 <2e-16 ***
Residuals 6433 2.026e+15 3.15e+11
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Рисунок 3.13 – Результати тесту ANOVA

Джерело: розроблено автором

Оцінемо важливість факторів за допомогою різних методів, включаючи алгоритми машинного навчання. Напишемо код з використанням алгоритму Random Forest, додаток В.1. У цьому прикладі ми побудували модель Random Forest залежності між факторами (усі доступні фактори в датасеті Walmart) та змінною відгуку "Weekly_Sales". Після натренування моделі ми використовуємо функцію `importance()`, щоб отримати показники важливості факторів. Отримані показники важливості факторів представлені у вигляді коефіцієнта важливості залежності (`IncNodePurity`) для кожного фактора. Чим більше значення коефіцієнта,

тим важливіший фактор в моделі. За наведеними даними, можна зробити наступні припущення щодо важливості факторів у моделі:

- Store: 9.803833e+14
- Date: 8.385222e+13
- Holiday_Flag: 4.467597e+12
- Temperature: 1.057246e+14
- Fuel_Price: 9.133224e+13
- CPI: 3.030426e+14
- Unemployment: 2.735363e+14
- Holiday_Type: 4.137377e+12

За отриманими значеннями, можна припустити, що фактор "Store" має найвищу важливість у моделі, оскільки має найбільше значення коефіцієнта важливості.

Для оцінки якості прогнозу моделі Random Forest, потрібно обчислити основні метрики, для цього скористаємося даним кодом, додаток В.2, він видає три параметри:

- MAE (Mean Absolute Error): 79292.18
- RMSE (Root Mean Squared Error): 113474.66
- MAPE (Mean Absolute Percentage Error): 10.86%

Ці метрики вказують на те, що середня абсолютна помилка прогнозу становить близько 79292.18, середня квадратична помилка - близько 113474.66 і середня абсолютна відносна помилка - близько 10.86%.

Проведемо аналіз системи часових рядів на базі VAR. Аналіз системи часових рядів на базі VAR моделей є популярним підходом для моделювання та прогнозування взаємозв'язків між кількома часовими рядами [30]. Для початку створюємо часовий ряд `ts_data` на основі двох змінних 'Weekly_Sales' і 'Temperature' з датафрейму Walmart. Будуємо модель з використанням функції VAR. Ми встановили параметр $p = 2$, що означає, що модель VAR буде мати два запізнення. Далі ми отримали оцінки коефіцієнтів моделі за допомогою функції

coef і зберегли їх у змінній var_coef. За допомогою функції summary отримали статистичний опис моделі VAR це надає загальну інформацію про модель, таку як коефіцієнти, стандартні помилки, значення t-статистики та значення p-статистики для кожного коефіцієнта.

Аналізуючи результати оцінки VAR-моделі, можна спостерігати:

- Кількість спостережень (Sample size) складає 6433.
- Логарифмічна правдоподібність (Log Likelihood) становить -107549.54.
- Корені характеристичного полінома рівні 0.9727, 0.9585, 0.3405, 0.1831. Ці корені вказують на стаціонарність моделі.
- Залишкова стандартна похибка (Residual standard error) для змінної Weekly_Sales становить 173900, а для змінної Temperature - 6.142.
- Значення Multiple R² та Adjusted R² свідчать про високу пояснювальну здатність моделі: 0.9051 для Weekly_Sales та 0.8892 для Temperature.
- Значущість моделі перевіряється за допомогою F-статистики. Для змінної Weekly_Sales отримано F-статистику 1.533e+04 зі значенням p-value менше 2.2e-16, що свідчить про статистичну значущість моделі. Аналогічно, для змінної Temperature отримано F-статистику 1.289e+04 зі значенням p-value менше 2.2e-16.

Для оцінки якості моделі VAR потрібно звертати увагу на наступні метрики:

- Логарифмічна правдоподібність – це показник, який відображає максимальне значення логарифмічної функції правдоподібності. Чим більше значення логарифмічної правдоподібності, тим краще модель відповідає спостережуваним даним.
- Корені характеристичного полінома – ці корені вказують на стаціонарність моделі VAR. Якщо всі корені менше за модулем одиниці, це свідчить про стаціонарність моделі та добру апроксимацію даних.
- Залишкова стандартна похибка - це показник, який відображає стандартну похибку залишків моделі VAR для кожної змінної. Чим менше значення

залишкової стандартної похибки, тим краще модель підходить до спостережуваних даних.

- Коефіцієнт детермінації – це показники, що вказують на частку варіації залежних змінних, яку можна пояснити моделлю VAR. Високі значення R^2 свідчать про високу пояснювальну здатність моделі, адже вона відображає велику частку варіації залежних змінних.
- Значущість моделі - F-статистика та відповідне p-значення використовуються для визначення статистичної значущості моделі VAR в цілому. Низьке значення p-значення (зазвичай менше 0.05) свідчить про статистичну значущість моделі.

Моделі VAR дозволяє моделювати залежності між кількома часовими рядами, використовуючи їх попередні значення. У випадку VAR(p), попередні значення рядів використовуються з лагами від 1 до p. Це загальне числове представлення моделі VAR, де значення параметра p визначається для конкретної задачі моделювання.

Побудуємо прогнози за кожною з моделей. Зараз ми продемонструємо прогноз за допомогою трьох різних моделей: ARIMA, VAR та Random Forest, додаток В.3.

Ми використали функцію `forecast()` з моделлю `arima_model` та встановили параметр $h = 10$, що означає прогнозування на 10 майбутніх спостережень. Функція `print()` виводить прогнозовані значення на екран, а функція `plot()` дозволяє візуалізувати прогноз (рис. 3.14).

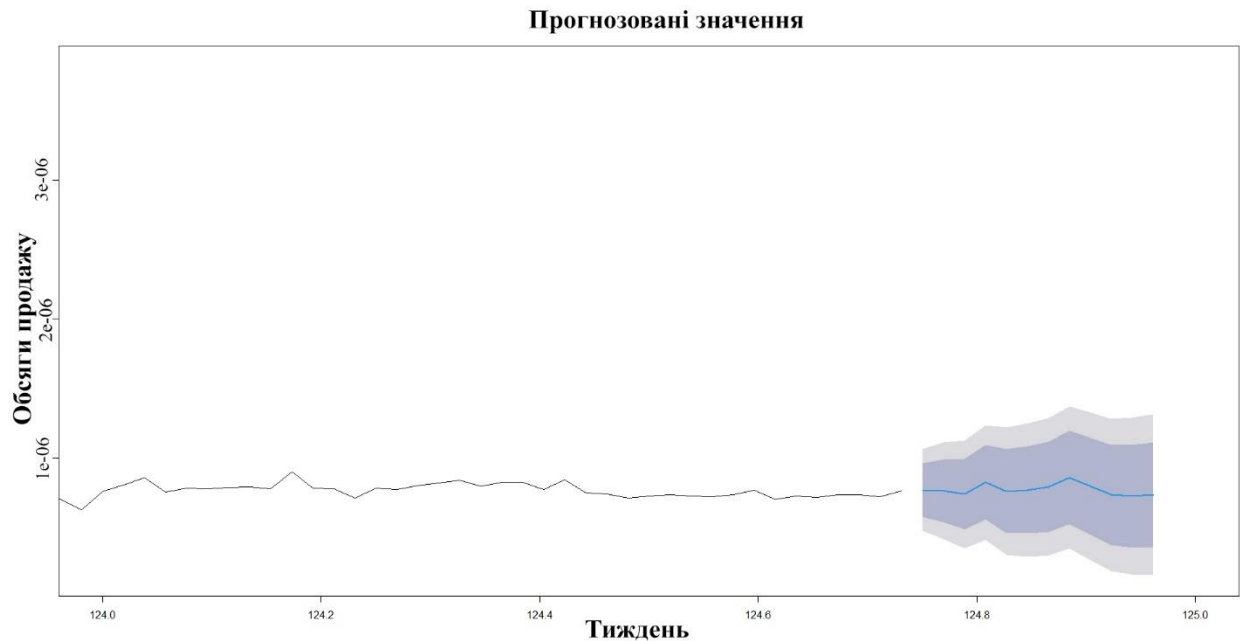


Рисунок 3.14 – Прогноз за допомогою моделі ARIMA

Джерело: розроблено автором

Для оцінки якості прогнозу моделі ARIMA, ми можемо використовувати функцію `assigasy()`, вона автоматично обчислює різні метрики прогнозу для порівняння прогнозованих значень з фактичними значеннями. З основних метрик прогнозу ARIMA можна зробити наступні висновки:

- RMSE для тестового набору даних становить 69252.18, що вказує на середню абсолютну помилку прогнозу на рівні близько 69,252 одиниць. Чим нижче значення RMSE, тим кращий прогноз.
- MAE для тестового набору даних дорівнює 60290.32. Це означає, що середня абсолютна помилка прогнозу становить близько 60,290 одиниць. Чим нижче значення MAE, тим кращий прогноз.
- MAPE для тестового набору даних складає 8.32%. Це вказує на середню відсоткову помилку прогнозу, яка становить близько 8.32%. Чим нижче значення MAPE, тим кращий прогноз.

На основі наведених метрик можна сказати, що прогноз моделі ARIMA для тестового набору даних має помірну точність. Далі ми здійснили прогнозування на основі моделі VAR за допомогою функції `predict`, встановили параметр `n.ahead =`

10, що означає, що прогнози на 10 майбутніх спостережень. Виводимо на екран отримані прогнози за допомогою функції `print()`. Це дозволяє переглянути прогнозовані значення для кожної змінної на майбутні спостереження (рис. 3.15).

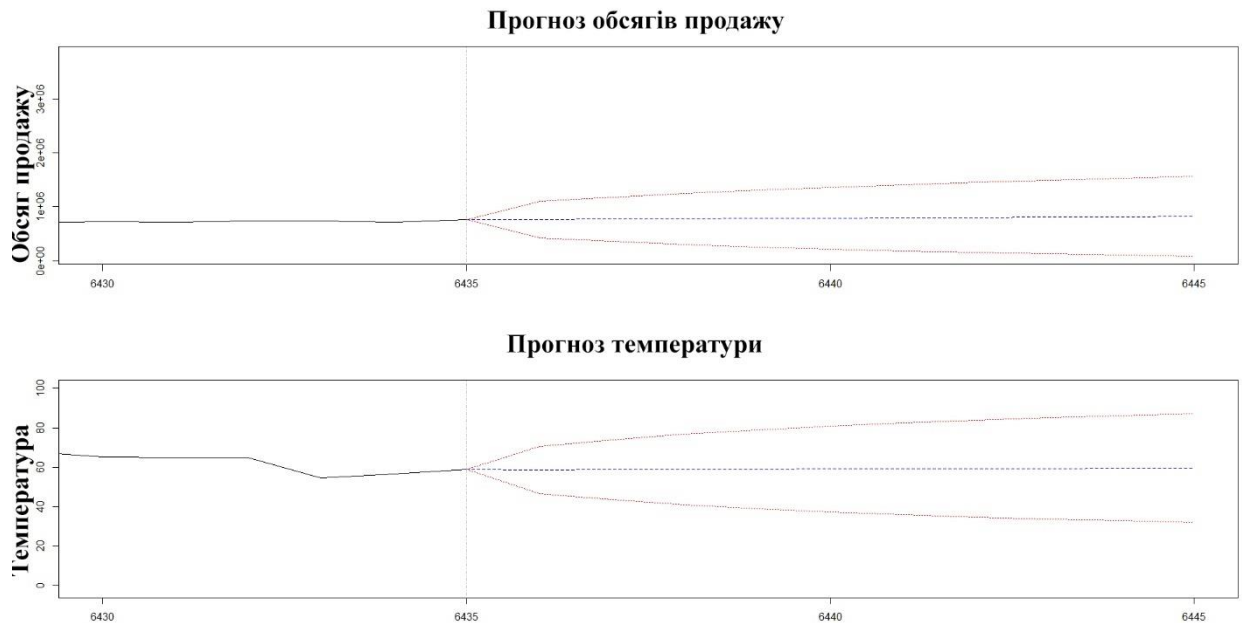


Рисунок 3.15 – Прогноз за допомогою моделі VAR

Джерело: розроблено автором

Для оцінки якості прогнозу моделі VAR, потрібно обчислити основні метрики, для цього скористаємося даним кодом, додаток B.4, він видає три параметри:

- MAE (середня абсолютна помилка): 340,744.68
- RMSE (корень середньоквадратичної помилки): 429,724.83
- MAPE (середня відносна помилка в процентах): 46.67

Ці метрики допомагають оцінити точність прогнозу моделі VAR. Більш низькі значення MAE та RMSE, а також нижчий відсоток MAPE вказують на кращу точність прогнозу.

Ми використали функцію `predict()` з моделлю `rf_model` та передали `newdata` зі значеннями змінних з датафрейму `Walmart`, щоб зробити прогноз на майбутні спостереження (рис. 3.18). Використання функції `predict()` для прогнозу Random Forest у даному випадку не дає очікуваних результатів, оскільки прогнозується вся

наявна вибірка, включаючи відомі значення цільової змінної. Це може призводити до відображення лише наявних значень на графіку, без чіткого прогнозу.

Отже, за метриками RMSE та MAE, кращим прогнозом є ARIMA, оскільки він має найменші значення помилок порівняно з іншими моделями. Однак, за метрикою MAPE, кращим прогнозом є Random Forest, оскільки він має найменший відсоток відносної помилки. Вибір найкращої моделі залежить від контексту та важливості конкретних метрик. Якщо важлива точність без врахування відсоткової помилки, то ARIMA є кращим варіантом. Якщо важливо мінімізувати відсоткову помилку, то Random Forest може бути кращим вибором.

ВИСНОВКИ

У даній кваліфікаційній бакалаврській роботі були вирішені такі завдання:

Дослідили принципи, предметну область та функціонування ринку предметів споживання як об'єкту моделювання. Оглянувши тенденції динаміки продажів та факторів, що впливають на них, у контексті торговельної діяльності можна дійти висновку, що ринок постійно змінюється, а споживачі стають все вибагливішими. Для успішної конкуренції в цій галузі, компаніям необхідно бути в курсі останніх тенденцій та пристосовуватися до них. Аналіз теоретичних аспектів показав, що останні тенденції це: інформатизація, впровадження ІТ в сферу торгівлі, та перехід від застарілих традиційних магазинів до онлайн продажів. Далеко не останню роль грає саме цифровізація, та впровадження ІТ технологій. Застосування різних методів та інструментів машинного навчання підтвердило важливість використання новітніх інформаційних систем для аналізу та прогнозування динаміки продажів. Це дозволяє компаніям отримувати цінну інформацію, виявляти незрозумілі залежності та приймати обґрунтовані рішення для підвищення ефективності бізнесу.

Проаналізували наукові джерела з напряду тенденцій розвитку ринку в умовах нестабільної економічної ситуації. Було розглянуто низку проблем, що прямо впливають на розвиток ринку, зокрема, такі фактори як криза, війни, та як наслідок - бідність населення. Ми розглянули, як бізнес перелаштовується задля виживання у такі скрутні часи, якими ІТ методами вони користуються та порівняли їх з світовими лідерами. Як підсумок, можна сказати, що компанії мають спільну мету - ефективне управління запасами та підвищення рівня продажів, для цього вони впроваджують машинне навчання, та вкладають великий людський ресурс, щоб знаходити нові рішення та мати неабияку перевагу над своїми конкурентами на ринку.

Розглянули методи та інструменти моделювання. В дослідженні були розглянуті різноманітні методи та інструменти моделювання, а саме одномірний

ряд, моделі ARIMA та VAR, нейронні мережі, зокрема LSTM, а також метод Random Forest, який базується на ансамблюванні дерев рішень. Узагальнюючи, кожен з цих методів має свої переваги та обмеження. Використання різних методів моделювання може допомогти досягти більш точних та надійних прогнозів у різних контекстах та завданнях прогнозування.

Розробили комплекс моделей для аналізу та прогнозування обсягів продажів товарів. Опираючись на проблематику даної роботи, був запропонований комплекс підходів, алгоритмів та моделей, що дозволяють у своїй сукупності аналізувати та прогнозувати дані. Була описана поетапна послідовність кроків, що формулюють загальне розуміння для аналізу та прогнозування, з потенційним впровадженням цих методів у реальну ситуацію у сфері торгівлі.

Виконали експериментальні дослідження на основі запропонованих методів. В практичній частині кваліфікаційної бакалаврської роботи був розглянутий аналіз даних з продажів магазинів Walmart. Ми використали різноманітні інструменти машинного навчання, такі як ARIMA, експоненційне згладжування та інші, для аналізу цих даних та прогнозування майбутніх продажів. Аналіз різних чинників, що впливають на продажі в торговельній галузі, показав, що успішність компаній залежить від багатьох факторів, навіть якщо цей фактор не є очевидним. Наприклад, святкові тижні, погода, вартість пального та рівень безробіття виявилися важливими чинниками, які впливають на продажі. Ми з'ясували в контексті даного дослідження, що святкові тижні не несуть за собою позитивного ефекту, тоді як погода, вартість пального та безробіття можуть мати як позитивний, так і негативний вплив. Ми спостерігали виразну сезонну динаміку, яка повторюється щорічно. Було проведено аналіз розподілу продажів для кожного окремого магазину. Спостерігається, що деякі магазини мають вищий рівень продажів, ніж інші, що може свідчити про їх більшу популярність серед споживачів. Також помітно, що діапазон продажів варіюється значно для кожного магазину, що може бути пов'язано з різними факторами, такими як розташування магазину чи рівень конкуренції.

Додатково, застосування моделі Random Forest дозволило нам визначити важливість окремих факторів у прогнозуванні продажів. Це дозволило нам зрозуміти, які змінні найбільше впливають на обсяги продажів. Отримані результати та методи, які ми використали, можуть бути корисними для багатьох галузей бізнесу, де аналіз динаміки продажів є важливим аспектом. Моделі прогнозування, такі як ARIMA, VAR і Random Forest, можуть допомогти у плануванні та управлінні запасами, а також у прогнозуванні майбутніх продажів. ARIMA і Random Forest показали добрі результати в прогнозуванні для даного набору даних, залежно від вибраної метрики.

Врахування екзогенних факторів може покращити точність прогнозування. Моделі ARIMA з екзогенними змінними (ARIMAX) і VAR дозволяють враховувати ці фактори. Аналіз ANOVA та важливості факторів у моделі Random Forest допомагають виокремити ключові економічні фактори, які мають найбільший вплив на продажі. Ці відомості можна використовувати для прийняття рішень щодо стратегії ціноутворення, маркетингових кампаній та управління асортиментом. Врахування цих чинників може допомогти компаніям оптимізувати свою стратегію продажів і досягти більшої ефективності.

Загалом, ця робота демонструє, що дослідження динаміки продажів та використання інструментів машинного навчання є важливими для вдосконалення бізнес-процесів та досягнення успіху у сучасному конкурентному середовищі.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2). Article 3–28.
<https://www.sjsu.edu/people/tom.means/courses/bigdata/s1/Big-Data-New-Tricks-for-Econometrics---jep.28.2.3.pdf>
2. Mosavi, A.; Faizollahzadeh Ardabili, S.; Shamshirband, S. (2019) Demand Prediction with Machine Learning Models; State of the Art and a Systematic Review of Advances. 2019050175. <https://doi.org/10.20944/preprints201905.0175.v1>
3. Економічна теорія: макро- і мікроекономіка : Навч. посіб. З. Г. Ватаманюк, С. М. Панчишин, С. К. Ревенчук, В. Г. Буняк, Н. І. Гнатюк, І. М. Грабинський, М. І. Крупка, Є. Й. Майовець, Р. В. Михайлишин; ред.: З. Г. Ватаманюк, С. М. Панчишин. - К. : Вид. дім "Альтернативи", 2001. - 606 с. - укр.
4. Економіка підприємства : підручник. Під заг. ред. д.е.н., проф. Ковальської Л.Л. та проф. Кривов'язюка І.В. Київ : Видавничий дім «Кондор», 2020. 700 с.
5. Sharda, R., Delen, D., & Turban, E. (2020). *Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support (11th Edition)*. Pearson.
6. Macrotrends | The Long Term Perspective on Markets. Macrotrends | The Long Term Perspective on Markets. URL: <https://www.macrotrends.net/countries/UKR/ukraine/poverty-rate> (date of access: 17.06.2023).
7. Економіка України: пандемія COVID-19, нова економічна реальність, рекомендації : дослідження для HannsSeidel-Stiftung. URL: http://hss.kiev.ua/files/Doslidshennja_HSS_Galasjuk_UA.pdf
8. Zhang, G. P., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14(1). Article 35-62. https://www.researchgate.net/publication/222489987_Forecasting_With_Artificial_Neural_Networks_The_State_of_the_Art

9. Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine*. *Annals of Statistics*, 29(5). Article 1189-1232.
<https://jerryfriedman.su.domains/ftp/trebst.pdf>
10. Wu, W., Liu, Y., & Hu, M. (2022). *Geo-Information Technology and Its Applications*. https://www.researchgate.net/publication/366190771_Geo-Information_Technology_and_Its_Applications
11. Burkov, A. (2019). *The Hundred-Page Machine Learning Book*. Andriy Burkov.
12. Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts. <https://otexts.com/fpp2/>
13. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
<https://www.deeplearningbook.org/>
14. Wickham, H., & Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media. <https://r4ds.had.co.nz/>
15. Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*. O'Reilly Media.
https://www.researchgate.net/publication/256438799_Data_Science_for_Business
16. Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2015). *Time series analysis: forecasting and control* (5th ed.). *John Wiley & Sons*.
17. Brockwell, P. J., & Davis, R. A. (2016). *Introduction to Time Series and Forecasting* (3rd ed.). *Springer*.
https://warin.ca/ressources/books/2016_Book_IntroductionToTimeSeriesAndFor.pdf
18. Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2015). *Time series analysis: forecasting and control* (5th ed.). *John Wiley & Sons*.
19. Грабовецкий Б. Є. Основы економічного прогнозування : навч. посібник. – Вінниця : ВФ ТАНГ, 2000.
20. Chatfield, C. (2019). *The Analysis of Time Series: An Introduction* (7th ed.). CRC Press.
http://site.iugaza.edu.ps/nbarakat/files/2010/02/Analysis_of_Time_Series_An_Introduction.pdf

21. Enders, W. (2014). *Applied Econometric Time Series*. Wiley.
https://new.mmf.lnu.edu.ua/wp-content/uploads/2018/03/enders_applied_econometric_time_series.pdf
22. Євсєєв Є.А. Економетрика: навчальний посібник для академічного бакалаврату . Є. А. Євсєєв , В. М. Буре. 2-е вид., виправлене і доповнене. М.: Юрайт, 2017. 186 с.
23. Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.
<http://mayoral.iae-csic.org/timeseries2021/hamilton.pdf>
24. Nielsen, M. (2015). *Neural Networks and Deep Learning*. Determination Press.
https://jingyuexing.github.io/Ebook/Machine_Learning/Neural%20Networks%20and%20Deep%20Learning-eng.pdf
25. Jurafsky, D., & Martin, J. H. (2020). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd Edition)*. <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>
26. Breiman, L. Random Forests. *Machine Learning* 45, 5–32 (2001).
<https://doi.org/10.1023/A:1010933404324>
27. Wickham, H. (2014). *Tidy Data*. *Journal of Statistical Software*, 59(10). Article 1-23. https://www.researchgate.net/publication/215990669_Tidy_data
28. Cowpertwait, P. S., & Metcalfe, A. V. (2009). *Introductory Time Series with R*. Springer.
https://michaela-kratofil.com/files/2009_Book_IntroductoryTimeSeriesWithR.pdf
29. Shumway R. H., Stoffer D. S. *Time Series Analysis and Its Applications*. New York, NY : Springer New York, 2011. URL: <https://doi.org/10.1007/978-1-4419-7865-3>
30. Lütkepohl, H. (2006). *New Introduction to Multiple Time Series Analysis*. Springer-Verlag Berlin Heidelberg.
https://doc.lagout.org/science/0_Computer%20Science/3_Theory/Mathematics/New%20Introduction%20to%20Multiple%20Time%20Series%20Analysis.pdf

ДОДАТКИ

Додаток А

Графік часового ряду прибутковості Walmart (A.1)

```
ggplot(walmart, aes(x = Date, y = weekly_sales)) +  
  geom_line() +  
  facet_wrap(~ Store, nrow = 5) + # Розбиття на підгрупи за магазинами  
  ggtitle("динаміка продажів walmart за магазинами") +  
  xlab("дата") +  
  ylab("обсяги продажу") +  
  theme_bw()  
  
filtered_data <- walmart[walmart$Store %in% c(1, 2), ]  
  
ggplot(filtered_data, aes(x = Date, y = weekly_sales)) +  
  geom_line() +  
  facet_wrap(~ Store, nrow = 1) +  
  ggtitle("динаміка продажів walmart за магазинами") +  
  xlab("дата") +  
  ylab("прибуток") +  
  theme_bw()
```

Гістограма розподілу тижневих продажів (A.2)

```
ggplot(walmart, aes(x = weekly_sales)) +  
  geom_histogram(fill = "blue", color = "white", bins = 30) +  
  labs(title = "Розподіл тижневих продажів", x = "обсяг продажу", y = "частота")
```

Графік розподілу тижневих продажів для кожного магазину (A.3)

```
ggplot(walmart, aes(x = factor(Store), y = weekly_sales)) +  
  geom_boxplot(fill = "lightblue", color = "blue") +  
  labs(title = "Розподіл тижневих продажів за магазинами", x = "магазин", y = "обсяг продажу")  
  
ggplot(walmart[walmart$Store == 1, ], aes(x = factor(Store), y = weekly_sales)) +  
  geom_boxplot(fill = "lightblue", color = "blue") +  
  labs(title = "Розподіл тижневих продажів за магазинами", x = "магазин", y = "обсяг продажу")
```

Кореляція між температурою та продажами (A.4)

```
ggplot(walmart, aes(x = Temperature, y = weekly_sales)) +  
  geom_point(alpha = 0.5, color = "blue") +  
  ggtitle("кореляція між температурою та продажами") +  
  xlab("температура") +  
  ylab("тижневі продажі")
```

Категоризуємо стовпець "Holiday_Type" (A.5)

```
walmart$Holiday_Type <- ifelse(walmart$Holiday_Flag == 1, "Holiday", "Non Holiday")
```

Продажі у святкові та несвяткові дні (A.6)

```
ggplot(walmart, aes(x = Date, y = weekly_sales, fill = Holiday_Type)) +  
  geom_bar(stat = "identity", position = "dodge") +  
  scale_fill_manual(values = c("Non-Holiday" = "gray",  
                                "Holiday" = "red")) +  
  ggtitle("порівняння тижневих розпродажів у святкові та несвяткові дні") +  
  ylab("weekly sales") +  
  xlab("тиждень") +  
  theme_bw()  
  
ggplot(walmart, aes(x = Holiday_Type, y = weekly_sales, fill = Holiday_Type)) +  
  geom_boxplot() +  
  ggtitle("порівняння продажів у святкові та звичайні тижні") +  
  xlab("тип тижня") +  
  ylab("тижневі продажі") +  
  scale_fill_manual(values = c("lightblue", "lightgreen"), name = "тип тижня", labels = c("звичайний", "святковий"))
```

Залежність між прибутковістю та % безробіття у регіоні (A.7)

```
ggplot(walmart, aes(x = unemployment, y = weekly_sales)) +  
  geom_point() +  
  geom_smooth(method = "lm", se = FALSE) +  
  ggtitle("Залежність між безробіттям та обсягами продажів") +  
  xlab("безробіття") +  
  ylab("обсяги продажу") +  
  theme_bw()
```

Залежність між вартістю пального та обсягами продажів (A.8)

```
ggplot(walmart, aes(x = Fuel_Price, y = weekly_sales)) +  
  geom_point() +  
  geom_smooth(method = "lm", se = FALSE) +  
  ggtitle("Залежність між вартістю пального та обсягами продажів") +  
  xlab("вартість пального") +  
  ylab("обсяги продажу") +  
  theme_bw()
```

Додаток Б

Використовуємо автоматичне підбір значень ARIMA (Б.1)

```
ts_sales <- ts(walmart$weekly_sales, frequency = 52)
arma_model <- auto.arima(ts_sales)
summary(arma_model)
```

Перевірка на стаціонарність (Б.2)

```
adf_result <- adf.test(ts_sales)
print(adf_result)
```

Підключаємо екзогенні змінні, щоб скористатися ARIMAX (Б.3)

```
exogenous_vars <- cbind(walmart$Temperature, walmart$Fuel_Price, walmart$CPI, walmart$Unemployment)
arma_order <- c(5, 1, 5)
arimax_model <- arima(ts_sales, order = arma_order, xreg = exogenous_vars)
```

Будуємо модель *Random Forest* (Б.4)

```
# Встановлення насіння для відтворюваності результатів
set.seed(123)

# Виконання Random Forest моделі
rf_model <- randomForest(weekly_sales ~ ., data = walmart)

# Виведення показників важливості факторів
importance <- importance(rf_model)
print(importance)
```

Оцінка якості прогнозу моделі *Random Forest* (В.2)

```
#прогноз за допомогою моделі Random Forest:
forecast_rf <- predict(rf_model, newdata = walmart, type = "response")
print(forecast_rf)
plot(forecast_rf, type = "l", xlab = "Time", ylab = "Forecast", main = "Random Forest Forecast")

actual_values_rf <- walmart$weekly_sales
forecast_errors_rf <- forecast_rf - actual_values_rf

mae_rf <- mean(abs(forecast_errors_rf))
rmse_rf <- sqrt(mean(forecast_errors_rf^2))
mape_rf <- mean(abs(forecast_errors_rf / actual_values)) * 100
print(paste("MAE:", mae_rf))
print(paste("RMSE:", rmse_rf))
print(paste("MAPE:", mape_rf))
```

Підсумок прогнозу, по кожній моделі (В.3)

```
#прогноз за допомогою моделі ARIMA:
forecast_arima <- forecast(arima_model, h = 10)
actual_values <- ts_sales[(length(ts_sales) - 9):length(ts_sales)]
accuracy_arima <- accuracy(forecast_arima, actual_values)
print(accuracy_arima)

print(forecast_arima)
plot(forecast_arima, type = "l", main = "прогнозовані значення", xlab = "час", ylab = "обсяги продажу", xlim = c(124, 125))

|

#прогноз за допомогою моделі VAR:
forecast_var <- predict(var_model, n.ahead = 10)
print(forecast_var)
plot(forecast_var, type = "l", main = "прогнозовані значення", xlab = "час", ylab = "обсяги продажу", xlim = c(6430, 6445))

forecast_errors <- forecast_var[["fcst"]][["weekly_sales"]] - actual_values

mae <- mean(abs(forecast_errors))
rmse <- sqrt(mean(forecast_errors^2))
mape <- mean(abs(forecast_errors / actual_values)) * 100

print(paste("MAE:", mae))
print(paste("RMSE:", rmse))
print(paste("MAPE:", mape))

#прогноз за допомогою моделі Random Forest:
forecast_rf <- predict(rf_model, newdata = walmart, type = "response")
print(forecast_rf)
plot(forecast_rf, type = "l", xlab = "Time", ylab = "Forecast", main = "Random Forest Forecast")

actual_values_rf <- walmart$weekly_sales
forecast_errors_rf <- forecast_rf - actual_values_rf

mae_rf <- mean(abs(forecast_errors_rf))
rmse_rf <- sqrt(mean(forecast_errors_rf^2))
mape_rf <- mean(abs(forecast_errors_rf / actual_values)) * 100
print(paste("MAE:", mae_rf))
print(paste("RMSE:", rmse_rf))
print(paste("MAPE:", mape_rf))
```

Оцінка якості прогнозу моделі VAR (В.4)

```
#Прогноз за допомогою моделі VAR:
forecast_var <- predict(var_model, n.ahead = 10)
print(forecast_var)
plot(forecast_var)
plot(forecast_var, type = "l", main = "прогнозовані значення", xlab = "час", ylab = "обсяги продажу", xlim = c(6430, 6445))

forecast_errors <- forecast_var[["fcst"]][["weekly_sales"]] - actual_values

mae <- mean(abs(forecast_errors))
rmse <- sqrt(mean(forecast_errors^2))
mape <- mean(abs(forecast_errors / actual_values)) * 100

print(paste("MAE:", mae))
print(paste("RMSE:", rmse))
print(paste("MAPE:", mape))
```

Ім'я користувача:
Математичного моделювання та статистики Даценк...

ID перевірки:
1015410367

Дата перевірки:
04.06.2023 11:08:25 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
04.06.2023 11:16:46 EEST

ID користувача:
100005744

Назва документа: Струк_Катуніна_Плагіат

Кількість сторінок: 64 Кількість слів: 13385 Кількість символів: 101961 Розмір файлу: 627.20 KB ID файлу: 1015073697

2.15% Схожість

Найбільша схожість: 0.51% з джерелом з Бібліотеки (ID файлу: 1005382073)

0.8% Джерела з Інтернету

187

Сторінка 66

2.09% Джерела з Бібліотеки

300

Сторінка 66

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0% Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи

7