МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Чорноморський національний університет
імені Петра Могили
Факультет комп'ютерних наук
Кафедра інтелектуальних інформаційних систем

ДОПУЩЕНО ДО ЗАХИСТУ
В. о. завідувача кафедри інтелектуальних
інформаційних систем, канд. техн. наук, доцент
_____Є. В. Сіденко
«____»_____2023 р.

МАГІСТЕРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

ІНТЕЛЕКТУАЛЬНА СИСТЕМА ПРОГНОЗУВАННЯ ВАРТОСТІ КОМЕРЦІЙНИХ КОМПАНІЙ НА ОСНОВІ МЕТОДІВ МАШИННОГО НАВЧАННЯ

Спеціальність 122 «Комп'ютерні науки»

122 – МКР – 601.21710118

Виконав студент 6-го курсу, групи 601
_____П. О. Мальченко
«___» лютого 2023 р.

Керівник: канд. техн. наук, доцент
_____І. О. Калініна
«___» лютого 2023 р.

Миколаїв – 2023

Чорноморський національний університет ім. Петра Могили

Факультет комп'ютерних наук

Кафедра інтелектуальних інформаційних систем

Освітньо-кваліфікаційний рівень **магістр**
Галузь знань **12 «Інформаційні технології»**
(шифр і назва)
Спеціальність **122 «Комп'ютерні науки»**
(шифр і назва)

ЗАТВЕРДЖУЮ
В. о. завідувача кафедри інтелектуальних інформаційних систем, канд. техн. наук, доцент
_____ Є. В. Сіденко
«____» _____20__р.

ЗАВДАННЯ

на магістерську кваліфікаційну роботу

Мальченку Павлу Олександровичу

1. Тема магістерської кваліфікаційної роботи «Інтелектуальна система прогнозування вартості комерційних компаній на основі методів машинного навчання».

Керівник роботи Калініна Ірина Олександрівна, канд. техн. наук, доцент

Затв. наказом Ректора ЧНУ ім. Петра Могили від «03» листопада 2022 р. № 199

2. Строк подання студентом роботи «1 » лютого 2023 р.

3. Вхідні (початкові) дані до роботи: часові ряди, найбільш вживані моделі прогнозування часових рядів, методи машинного навчання для прогнозування.

Очікуваний результат роботи: розроблена інтелектуальна система прогнозування вартості комерційних компаній, що представлені часовими рядами, з використанням декількох моделей та їх комбінації.

4. Зміст пояснювальної записки (перелік питань, які потрібно розглянути): дослідження та аналіз сучасного стану статистики у питанні вартості акцій, аналіз методів машинного навчання для прогнозування, аналіз засобів та ПЗ для

прогнозування; дослідження питань трансформації вхідних даних у необхідні формати, заповнення пропусків, аналіз засобів декомпозиції часових рядів, побудови моделей, прогнозування та оцінювальних метрик для визначення якості як моделей, так і прогнозування.

5. Перелік графічних матеріалів: презентація

6. Завдання до спеціальної частини: розробка практичного заняття з використанням побудови моделі ARIMA та прогнозування на її основі. Опис основних питань охорони праці пов'язаних з професійною діяльністю та використання комп’ютерів для роботи в офісі.

7. Консультанти:

Розділ	**Прізвище, ініціали та посада консультанта**	**Підпис**
Спеціальна частина з охорони праці	Григор’єва Л. І., докт. біол. наук, проф.	
Методична частина	Калініна І. О., канд. техн. наук, доцент	

Керівник роботи _канд. техн. наук, доцент, Калініна І. О._

(наук. ступінь, вчене звання, прізвище та ініціали)

_____

(підпис)

Завдання прийнято до виконання __Мальченко П. О._

(прізвище та ініціали)

_____

(підпис)

Дата видачі завдання « 07 » листопада 2022 р.

КАЛЕНДАРНИЙ ПЛАН
Виконання магістерської кваліфікаційної роботи

Тема: Інтелектуальна система прогнозування вартості комерційних компаній на основі методів машинного навчання

№	Найменування роботи	Початок	Закінчення	Примітки
1	Визначення керівника і теми МКР. Подання заяви на затвердження теми МКР	01.09.2022	20.10.2022	Виконано
2	Отримання завдання на виконання МКР	21.10.2022	10.11.2022	Виконано
3	Складання календарного плану на період виконання МКР	11.11.2022	15.11.2022	Виконано
4	Огляд літератури за темою дослідження	16.11.2022	27.11.2022	Виконано
5	Проходження переддипломної практики, збір та аналіз матеріалів до МКР	28.11.2022	18.12.2022	Виконано
6	Аналіз предметної області та розробка технічного завдання. Моделювання результатів	19.12.2022	12.01.2023	Виконано
7	Опис фахової частини МКР, зокрема аналіз сучасного стану прогнозування часових рядів, огляд існуючих технологій, розробка ІС, реалізація обраних технологій з аналізом отриманих результатів	13.01.2023	25.01.2023	Виконано
8	Розробка спеціальної частини з охорони праці та методичної частини	26.01.2023	02.02.2023	Виконано
9	Попередній захист МКР на засіданні комісії кафедри	03.02.2023	03.02.2023	Виконано
10	Корегування роботи за результатами попереднього захисту	04.02.2023	06.02.2023	Виконано
11	Остаточне оформлення пояснювальної записки та слайдів доповіді для захисту	07.02.2023	09.02.2023	Виконано
12	Подання МКР рецензенту	09.02.2023	10.02.2023	Виконано
13	Рецензування МКР	11.02.2023	12.02.2023	Виконано
14	Подання МКР, її електронної копії та інших документів (відгуку, рецензії) до захисту	15.02.2023	16.02.2023	Виконано
15	Захист МКР перед екзаменаційною комісією (ЕК)	22.02.2023	23.02.2023	Виконано

Розробив студент Мальченко П. О. _____
(прізвище та ініціали) (*підпис*)

Керівник роботи канд. техн. наук, доцент, Калініна І. О. _____
(наук. ступінь, вчене звання, прізвище та ініціали) *(підпис)*

«13» листопада 2022 р.

АНОТАЦІЯ

магістерської кваліфікаційної роботи
студента групи 601 ЧНУ ім. Петра Могили

Мальченка Павла Олександровича

на тему**: «ІНТЕЛЕКТУАЛЬНА СИСТЕМА ПРОГНОЗУВАННЯ ВАРТОСТІ КОМЕРЦІЙНИХ КОМПАНІЙ НА ОСНОВІ МЕТОДІВ МАШИННОГО НАВЧАННЯ»**

Актуальність даного дослідження полягає у необхідності підвищення якості прогнозування, відборі кращих моделей для відповідних наборів даних. Це дозволить зменшити збитки, що несуть компанії у результаті коливання вартості за рахунок підготовки заздалегідь.

Об'єктом дослідження є процес аналізу та прогнозування даних на основі методів машинного навчання.

Предмет дослідження – комбіновані прогнозні моделі на основі ARIMA та GAM.

Метою роботи є підвищення якості прогнозування вартості комерційних компаній за рахунок використання модифікованого методу ARIMA, налаштованих GAM та їх комбінації.

У першому розділі розглядається аналіз наявних робіт на тему прогнозування вартості комерційних компаній методами машинного навчання. У другому розділі розглядаються математичні моделі, методи, інформаційні технології, що використовуються для прогнозування вартості комерційних компаній. У третьому розділі описано проєктування моделей та їх програмна реалізація з подальшим прогнозуваннями та їх результатами. Спеціальна частина представлена двома розділами, що присвячені методичній частині та охороні праці відповідно.

Загальний обсяг роботи – 108 сторінок. Магістерська кваліфікаційна робота містить 18 таблиць, 33 рисунки, 1 додаток і посилання на 47 літературних джерел.

Ключові слова: прогнозування, часові ряди, машинне навчання, декомпозиція, автокореляція, стаціонарність, моделі, метрики, R.

ABSTRACT

to the master's qualification work by the student of the group 601 of Petro Mohyla Black Sea National University

Pavlo Malchenko

“AN INTELLIGENT SYSTEM FOR FORECASTING THE VALUE OF COMMERCIAL COMPANIES BASED ON MACHINE LEARNING METHODS”

A relevance of this study lies in the need to improve the quality of forecasting, to select the best models for the relevant data sets. This will reduce the losses incurred by the company as a result of price fluctuations due to preparation in advance.

An object of research is the process of data analysis and forecasting based on machine learning methods.

A subject of the research is combined predictive models based on ARIMA and GAM.

A purpose of the work is to improve the quality of forecasting the value of commercial companies due to the use of the modified ARIMA method, adjusted GAMs and their combination.

The first section deals with the analysis of existing works on the topic of forecasting the value of commercial companies using machine learning methods. The second chapter examines mathematical models, methods, and information technologies used to forecast the value of commercial companies. The third section describes the design of the models and their software implementation with further predictions and their results. The special part is represented by two sections devoted to the methodical part and labor protection, respectively.

The overall scope of the work is 108 pages. Thesis contains 18 tables, 33 figures, 1 appendix, 47 sources.

Keywords: forecasting, time series, machine learning, decomposition, autocorrelation, stationarity, models, metrics, R.

ЗМІСТ

Пояснювальна записка

до магістерської кваліфікаційної роботи

на тему:

«ІНТЕЛЕКТУАЛЬНА СИСТЕМА ПРОГНОЗУВАННЯ ВАРТОСТІ КОМЕРЦІЙНИХ КОМПАНІЙ НА ОСНОВІ МЕТОДІВ МАШИННОГО НАВЧАННЯ»

Спеціальність 122 «Комп'ютерні науки»

122 – МКР – 601.21710118

Виконав студент 6-го курсу, групи 601
П. О. Мальченко
«__» лютого 2023 р.

Керівник: канд. техн. наук, доцент
І. О. Калініна
«__» лютого 2023 р.

Миколаїв – 2023

ВСТУП

Фондовий ринок є одним із найважливіших складових ринкової економіки, оскільки він забезпечує компаніям доступ до капіталу, дозволяючи інвесторам купувати частки власності в компанії. Сфера фондового ринку постійно розвивається в процесі удосконалення. Враховуючи варіації, які він приносить щодня, інвесторам потрібно інтенсивно планувати, щоб отримати прибуток. Прогнозування даних фондової біржі включає припущення, що інформація, доступна на даний момент, має певний прогностичний зв'язок із майбутніми прибутками акцій. Прогнозування фондових тенденцій є одним із найскладніших завдань на фінансовому ринку через складність у заплутаному світі фондового ринку. Інвестори на фондовому ринку завжди знаходять техніку, яка може гарантувати легкий прибуток шляхом прогнозування фондових трендів і мінімізувати ризик інвестування. Це мотивує дослідників у цій галузі розробляти нові моделі прогнозування.

Курси акцій не є випадково згенерованими значеннями, а їх можна розглядати як модель дискретного часового ряду, яка базується на наборі чітко визначених числових даних, зібраних у послідовних точках через рівні проміжки часу.

Прогнозування курсу акцій є важливою темою у фінансах та економіці, яка передбачає ефективне прогнозування динаміки акцій, яке може мінімізувати ризик збитків і максимізувати прибуток. Передбачити ціну акцій є складним завданням через складні моделі часових рядів.

Дані часового ряду – це послідовність числових спостережень, природно впорядкованих у часі. Останнім часом стрімкий розвиток інформаційних технологій призвів до ситуації, коли величезні обсяги даних накопичуються з великою швидкістю і фактично утворюють різноманітні часові ряди. Моделювання таких часових рядів є надзвичайно важливим і привертає увагу як практиків, так і дослідників. Однак це також вважається досить складною

проблемою через багато складних характеристик, які часто присутні в часових рядах, таких як нерівномірності, нестабільність, тренди та шуми, тощо. Для моделювання часових рядів на основі їх поточної та минулої поведінки було розроблено низку методів.

Модель інтегрованої авторегресії середнього ковзного (ARIMA) і узагальнена адитивна модель (GAM) широко використовувалися для прогнозування часових рядів у сфері цін на біржі М’янми. Дані готуються для аналізу часових рядів шляхом виконання етапів попередньої обробки даних, таких як перетворення позначок часу, стаціонарна ідентифікація та стаціонарна обробка. Щоб знайти найточнішу модель прогнозу та найбільш відповідний період прогнозування, виконується аналіз помилок методів GAM і ARIMA та порівнюється на одному наборі даних. Вони надають деякі варіанти обробки сезонності набору даних. Це варіанти річної, тижневої та щоденної сезонності. Завдяки наданню цих параметрів аналітик даних може вибрати доступну деталізацію часу для моделі прогнозу на наборі даних.

У сучасних організаціях, які піддаються різким і величезним змінам, які впливають навіть на найбільш усталені структури, і де всі вимоги бізнес-сектору потребують точного та практичного прочитання в майбутньому, прогнози стають дуже важливими, оскільки вони є ознакою виживання та мовою бізнесу у світі.

Метою роботи є підвищення якості прогнозування вартості комерційних компаній за рахунок використання модифікованого методу ARIMA, налаштованих GAM та їх комбінації.

Об’єктом дослідження є процес аналізу та прогнозування даних на основі методів машинного навчання.

Предметом дослідження постають комбіновані прогнозні моделі на основі ARIMA та GAM.

1 АНАЛІЗ ПРОГНОЗУВАННЯ ВАРТОСТІ КОМЕРЦІЙНИХ КОМПАНІЙ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

1.1 Опис предметної сфери

Для початку з’ясуємо що ж таке акції. Капітал компанії поділений на акції [1]. Кожна акція утворює одиницю власності компанії та пропонується для продажу з метою залучення капіталу для компанії.

Акції можна умовно розділити на дві категорії – власний капітал і привілейовані акції. Акції дають їхнім власникам право ділитися доходами/прибутками компанії, а також право голосу на загальних зборах компанії. Такий акціонер повинен ділитися прибутком, а також нести збитки, понесені компанією.

З іншого боку, привілейовані акції приносять своїм власникам лише дивіденди, які є фіксованими, не дають права голосу. Акціонери вважаються справжніми власниками компанії. Коли акції виставляються на продаж безпосередньо компанією вперше, вони пропонуються на первинному ринку, тоді як торгівля акціями відбувається на вторинному ринку.

Акції оцінюються відповідно до різноманітних принципів на різних ринках, але основною передумовою є те, що акція коштує тієї ціни, за якою, ймовірно, відбулася б угода, якби акції продавалися. Ліквідність ринків є основним критерієм щодо того, чи можна продати акцію в будь-який момент часу. Зазвичай вважається, що фактична операція продажу акцій між покупцем і продавцем забезпечує найкращий ринковий показник prima facie щодо «справжньої вартості» акцій у цей конкретний час.

В свою чергу прогнозування є на сьогоднішній день однією з важливих задач планування бізнес-процесів [2]. Від якості прогнозування залежить якість прийнятих рішень щодо планування виробництва, транспортування та персоналу. Прогнозування має бути невід’ємною частиною процесу прийняття стратегічних рішень, оскільки воно може відігравати важливу роль у багатьох сферах

діяльності компанії. Завжди є потреба у короткострокових, середньострокових і довгострокових прогнозах в залежності від конкретного завдання. Деякі речі легше передбачити, ніж інші. Час сходу сонця завтра вранці можна передбачити точно. З іншого боку, цифри завтрашнього лотереї неможливо передбачити з точністю. Передбачуваність події або кількості залежить від кількох факторів, зокрема:

а) наскільки добре ми розуміємо фактори, що цьому сприяють;

б) скільки даних доступно;

в) наскільки майбутнє схоже на минуле;

г) чи можуть прогнози вплинути на те, що ми намагаємося спрогнозувати.

Часто в прогнозуванні ключовим кроком є знання того, коли щось можна спрогнозувати точно, а коли прогнози будуть не кращими, ніж підкидання монети. Хороші прогнози враховують справжні моделі та зв’язки, які існують в історичних даних, але не повторюють минулі події, які більше не повторяться.

Багато людей помилково вважають, що прогнози неможливі в мінливому середовищі. Будь-яке середовище змінюється, і хороша модель прогнозування відображає те, як все змінюється. Прогнози рідко припускають, що середовище залишається незмінним. Зазвичай припускають, що навколишнє середовище змінюватиметься й у майбутньому. Тобто, дуже мінливе середовище і надалі залишатиметься дуже мінливим; бізнес із коливаннями продажів матиме коливання продажів; і економіка, яка пройшла через буми та спади, продовжуватиме проходити через буми та спади. Модель прогнозування призначена для того, щоб фіксувати те, як рухаються речі, а не лише те, де вони є. Як сказав Авраам Лінкольн: «Якби ми могли спочатку знати, де ми знаходимося і куди прагнемо, ми могли б краще судити, що і як робити».

Під час прогнозування даних часових рядів мета полягає в тому, щоб оцінити, як послідовність спостережень продовжиться в майбутньому. Найпростіші методи прогнозування часових рядів використовують лише інформацію про змінну, яку потрібно прогнозувати, і не намагаються виявити

фактори, які впливають на її поведінку. Тому вони екстраполюють тренд та сезонні моделі, але ігнорують всю іншу інформацію, таку як маркетингові ініціативи, діяльність конкурентів, зміни в економічних умовах тощо. Всі особливості завдань прогнозування відображаються в аналітичних моделях прогнозування. Сучасні методи створення аналітичних моделей прогнозування часових рядів надають можливість налаштовувати параметри кожної моделі відповідно до обраного набору даних. Тому можна заключити, що існує потреба у системному підході при розробці інформаційної системи прогнозування, яка передбачає комбінування кількох прогнозних моделей для отримання якісних прогнозів. Такі системи вимагають точного визначення параметрів моделей, базових методів прогнозування, а також вибору відповідної комбінації методів для вирішення проблеми.

Відповідні методи прогнозування значною мірою залежать від наявних даних [3]. Якщо немає доступних даних або якщо наявні дані не відповідають прогнозам, то необхідно використовувати якісні методи прогнозування.

Кількісне прогнозування може бути застосоване при виконанні двох умов:

а) доступна чисельна інформація про минуле;

б) розумно припустити, що деякі аспекти минулих моделей триватимуть і в майбутньому.

Існує широкий спектр кількісних методів прогнозування, які часто розробляються в рамках конкретних дисциплін для певних цілей. Кожен метод має свої властивості, точність і вартість, які необхідно враховувати при виборі конкретного методу.

У більшості задач кількісного прогнозування використовуються або дані часових рядів (зібрані через регулярні проміжки часу), або дані перехресного перерізу (зібрані в один момент часу).

Завдання прогнозування зазвичай включає п'ять основних кроків [4] (табл. 1.1).

Таблиця 1.1 – Основні етапи завдання прогнозування

Визначення кроку	Опис кроку
Визначення проблеми	Часто це найскладніша частина прогнозування. Ретельне визначення проблеми вимагає розуміння того, як будуть використовуватися прогнози, кому потрібні прогнози та як функція прогнозування вписується в організацію, яка потребує прогнозів. Прогнозисту потрібно витрачати час на спілкування з усіма, хто буде залучений до збору даних, підтримки баз даних і використання прогнозів для майбутнього планування.
Збір інформації	Завжди потрібні принаймні два види інформації: (а) статистичні дані та (б) накопичений досвід людей, які збирають дані та використовують прогнози. Часто буде важко отримати достатньо історичних даних, щоб підібрати хорошу статистичну модель. Іноді старі дані будуть менш корисними через структурні зміни в прогнозованій системі; тоді ми можемо використовувати лише найновіші дані. Однак пам’ятайте, що хороші статистичні моделі впораються з еволюційними змінами в системі; не викидайте хороші дані без потреби.
Попередній (пошуковий) аналіз	Завжди починайте з побудови даних у графіку. Чи є послідовні закономірності? Чи є суттєвий тренд? Чи важлива сезонність? Чи є докази наявності бізнес-циклів? Чи є в даних якісь викиди, які потрібно пояснити фахівцям? Наскільки сильні зв’язки між змінними, доступними для аналізу? Для цього аналізу розроблено різні інструменти.
Вибір моделей	Найкраща модель для використання залежить від наявності історичних даних, міцності зв’язків між змінною прогнозу та будь-якими пояснювальними змінними, а також від способу

Визначення кроку	Опис кроку
	використання прогнозів. Зазвичай порівнюють дві-три потенційні моделі. Кожна модель сама по собі є штучною конструкцією, яка базується на наборі припущень (явних і неявних) і зазвичай включає один або більше параметрів, які необхідно оцінити з використанням відомих історичних даних.
Використання та оцінка моделі прогнозування	Після вибору моделі та оцінки її параметрів модель використовується для складання прогнозів. Ефективність моделі можна належним чином оцінити лише після того, як стануть доступними дані за прогнозований період. Було розроблено ряд методів, які допомагають оцінити точність прогнозів. Існують також організаційні питання щодо використання та виконання прогнозів. При використанні моделі прогнозування на практиці виникають численні практичні питання, наприклад, як працювати з відсутніми значеннями та викидами або як мати справу з короткими часовими рядами.

1.2 Огляд та аналіз наявних публікацій

Наявна велика кількість статей та інших наукових робіт чи підручників, що використовують моделі ARIMA та GAM для прогнозування за часовими рядами. Так, наприклад, науковці з Іраку зазначають, що модель ARMA є компонентом двох різних моделей, які пояснюють поведінку ряду з двох різних точок зору: моделі авторегресії (AR) і моделі ковзного середнього (MA) [5-9]. MA поєднує залежність між залишковою помилкою від моделі ковзного середнього та спостереження. Різниця між ARMA та ARIMA полягає в інтеграційному компоненті, який повертає нас до теми стаціонару. Модель ARIMA — це слово, що складається з трьох частин AR, MA та integrate (I). Це стосується різниці

необроблених точок даних, щоб забезпечити збереження стаціонарності. Іншими словами, точка даних буде замінена різницею між двома значеннями. Насправді більшість змінних є нестаціонарними; отже, вони повинні пройти через процес трансформації, який називається диференціацією або інтеграцією, перш ніж вони стануть стаціонарними. Запропонований модифікований метод відрізняється від автоматичного самостійним перебором можливих моделей порівнюючи не лише інформаційний критерій Акайки або Баєсівський інформаційний критерій, а й розподіл залишків та автокореляцію.

Інша робота присвячена також модифікованому методу ARIMA надає змінений підхід [10-14]. Замість побудови однієї моделі ARIMA, яка відповідала б усім доступним часовим рядам, було створено кілька моделей ARIMA, які б відповідали одному ряду, але закінчувалися в різні дати. Прогнози були виконані за допомогою цих історичних моделей. Прогнозовані значення порівнювалися з реальними. Різниці між реальними та прогнозованими значеннями були використані для модифікації результатів загальної моделі ARIMA. Ітераційні обчислення були завершені макросом.

Оскільки поточні значення часового ряду зазвичай мають сильніші кореляції з останніми значеннями ряду, але не з попередніми, номер ітерації для специфікації лагу було призначено 10 (спочатку було призначено 16). Загалом було створено 1800 (60x30) можливих трирівневих комбінацій затримок і типів, щоб моделі відповідали одній серії, але закінчилися через 30 різних тижнів. Для кожної моделі було обрано тип моделі та комбінацію затримки з найменшим SBC. Найкраща модель ARIMA із заданим типом і затримками була автоматично створена макросом.

Використання двох різних підходів зустрічається у роботі за участі опорної векторної регресії та моделі ARIMA, а також їх комбінації [15-18]. Модель опорного вектора регресії (SVR) є багатообіцяючим інструментом для фінансового прогнозування акцій із великими коливаннями, тоді як моделі (ARIMA) чудово підбирають лінійність без шкоди для прогнозів трендів для

довгострокових горизонтів, що має вирішальне значення, наприклад, для оптимізація портфеля. Завдяки своїй ефективності для вирішення проблем нелінійного оцінювання, моделі SVR широко використовувалися для прогнозування часових рядів у різних областях, наприклад, для вітрової енергії та фінансового прогнозування, а також для прогнозування вартості та точності виробництва промислового обладнання. Отже, у цій статті ми використовуємо модель SVR як компонент гібридної моделі, оскільки вона показала свою ефективність для прогнозування часових рядів.

Перевагою використання моделей ARIMA є їх універсальність і те, що їх можна налаштувати для прогнозування різних часових явищ. Можна коригувати різноманітні дані часових рядів за допомогою AR (авторегресійних моделей), MA (ковзних середніх) або ARMA, що є комбінацією AR і MA. Завдяки своїй ефективності як класичного методу прогнозування він широко використовується, наприклад, для прогнозування якості води, робочих навантажень у хмарних додатках, індексу EBITDA для фінансових показників і короткострокових навантажень клієнтів. Оскільки ARIMA є лінійною моделлю, вона не може охопити нелінійні моделі в часовому ряді; отже, для прогнозування сильних коливань у прогнозуванні часових рядів використовуються різні прогностичні моделі, засновані на техніках машинного та глибокого навчання.

Використання гібридних моделей або поєднання пари методів для підвищення продуктивності прогнозу є однією з найпотужніших альтернатив для покращення помилок прогнозування. Гібридні моделі бувають або однорідними, наприклад із використанням різноманітних налаштувань нейронної мережі, або гетерогенними, що використовують як лінійні, так і нелінійні методи. Це може бути дуже корисним для прогнозування часових рядів, оскільки часові ряди можуть демонструвати нелінійну поведінку з часом; однак він може стати лінійним відповідно до вхідних міркувань. Для цих типів гібридних моделей, тоді як модель машинного навчання обробляє нелінійність, ARIMA обробляє лінійну, нестаціонарну частину.

Ще однією роботою, присвяченою прогнозуванню за допомогою порівняння двох моделей, є дослідження [19]. У запропонованій системі ціна акцій М'янми прогнозується за допомогою моделі ARIMA та реалізована за допомогою python. Історичний набір даних про акції М'янми було зібрано з індексу цін на акції М'янми (MYANPIX). Система складається з двох основних компонентів: попередньої обробки та прогнозування фондового ринку.

Автоматично виявлено функцію оптимального порядку: auto_arima використовується для побудови моделі ARIMA. Функція auto_arima намагається визначити найбільш оптимальні параметри для моделі ARIMA та повертає підігнану модель ARIMA. Ця функція працює за допомогою розширеного методу Дікі-Фуллера (ADF) для визначення порядку розрізнення, d, а потім підгонки моделей у межах визначених діапазонів start_p, max_p, start_q і max_q. Подальше створення моделей поділяється на 2 етапи: ARIMA та GAM.

ARIMA. Щоб побудувати модель, дані розділяються на дані навчання та тестування. Відповідний порядок з кроку 1 побудови моделі ARIMA використовувався для побудови трьох моделей ARMA на навчальних даних (щоденні, місячні, щотижневі дані) і часових рядах прогнозування (щоденні, місячні, щотижневі тестові дані). У цій роботі альфа 0,05 означає 95% достовірності для прогнозування.

GAM (Prophet). Для побудови моделі використовуються два розділених даних: навчальні та тестові дані. Модель відповідає власному спеціальному набору даних для обробки часових рядів і сезонності. Набір даних містить два основні стовпці: стовпець «ds», у якому зберігаються часові ряди дат, і інший стовпець «y», у якому зберігаються відповідні значення часових рядів у наборі даних. Три різні параметри для періоду в моделі використовуються для прогнозування даних часового ряду (щодня, щотижня, щомісяця).

1.3 Постановка задачі

Для вирішення задачі вибору кращої моделі з подальшим прогнозуванням вартості акцій компанії треба виконати певні задачі. Перш за все, потрібно завантажити дані, трансформувати та візуалізувати їх. Після візуальної оцінки, потрібно провести перевірки на викиди та пропуски. У разі виявлення будь-якої з вказаних проблем – вирішити її. Наступним кроком є візуалізація вже редагованої вибірки. Після оцінки нормальності отриманих значень – розділити набір даних на вибірки: навчальна та тестова. Далі треба виконати тести на нелінійність, автокореляцію, нестаціонарність даних. Також треба виконати декомпозицію часового ряду. Після декомпозиції, обов’язковим є перевірка кожної компоненти та визначення наявності впливу на дані. Для моделі ARIMA потрібно провести тести на кількість необхідних трендових та сезонних диференціацій, візуалізувати дані, а також переглянути автокореляцію та часткову автокореляцію. На основі отриманих даних – визначаються параметри p, d, q та P, D, Q. З вибірки параметрів створюється вибірка моделей, які будуть проходити модифікований метод ARIMA. Також, необхідно виконати автоматичний підбір параметрів моделі, глибинний підбір параметрів моделі та згладження вибірки з автоматичним підбором параметрів. Останні три моделі є важливими, оскільки вони представляють собою виключно комп’ютеризований(математичний) підхід, який дозволяє оцінити моделі виключно з числовими показниками. Після ітераційного відбору моделей за модифікованим методом ARIMA, отримана модель додається до автоматичних моделей і виконується тестування якостей моделей. Тестування включає в себе перевірку таких показників:

а) AIC;

б) BIC;

в) R^2;

г) Durbin-Watson (DW).

Якщо якась з моделей сильно відрізняється у гіршу сторону за цими показниками – її можна відкинути. Після цього відбувається обов'язкова перевірка залишків на автокореляцію та нормальність розподілу за допомогою візуалізації. Також треба провести тест портманто.

Далі відбувається перехід до моделей типу GAM. Використовуючи ручні налаштування моделей, треба створити декілька моделей, кожна з яких використовуватиме окрему складову [20]. Наприклад, одна модель використовує тренд, друга сезон з мультиплікативною складовою, а третя комбінацію складових. Проводячи тестування моделей з різними значеннями параметрів потрібно обрати декілька найкращих, що найточніше представлятимуть дані вибірки. Знову ж таки, обов'язкова перевірка залишків.

Після перевірки моделей – прогнозування. Перш за все після прогнозування потрібно візуалізувати отримані значення. Оцінивши графічно отримані дані, інтервали довіри треба провести тести на точність. Ці тести включають в себе основні метрики похибок, такі як: ME, MASE, RMSE, тощо. Додатково потрібно провести перевірку за метриками Вінклера та CRPS. Після перевірки результатів, оцінити наскільки сильно найкращі дві моделі відрізняються від ідеальних значень (значень тестової вибірки) і на їх основі створити комбіновану модель, з подальшою її перевіркою, прогнозом на її основі та перевіркою самого прогнозу.

Пройшовши усі вищенаведені етапи, потрібно зробити висновки по проведеній роботі і визначити, яка з моделей показала себе найкраще, проаналізувати її роботу на вибірці. За необхідності – провести покращення моделі або шляхом модифікації, або шляхом комбінації з іншою моделлю. Таким чином, використання різних підходів до розробки моделей прогнозування в єдиній інформаційній системі дозволяє отримати ефективний прогноз вартості комерційної компанії.

Висновки до розділу 1

Було визначено, що акції – це частки капіталу компанії. Кожна акція утворює одиницю власності компанії та пропонується для продажу з метою залучення капіталу для компанії. Акції оцінюються відповідно до різноманітних принципів на різних ринках, але основною передумовою є те, що акція коштує тієї ціни, за якою, ймовірно, відбулася б угода, якби акції продавалися.

Прогнозування є на сьогоднішній день однією з важливих задач планування бізнес-процесів. Від якості прогнозування залежить якість прийнятих рішень щодо планування виробництва, транспортування та персоналу. Прогнозування має бути невід'ємною частиною процесу прийняття стратегічних рішень, оскільки воно може відігравати важливу роль у багатьох сферах діяльності компанії.

Сучасні методи створення аналітичних моделей прогнозування часових рядів надають можливість налаштовувати параметри кожної моделі відповідно до обраного набору даних. Тому можна заключити, що існує потреба у системному підході при розробці інформаційної системи прогнозування, яка передбачає комбінування кількох прогнозних моделей для отримання якісних прогнозів. Такі системи вимагають точного визначення параметрів моделей, базових методів прогнозування, а також вибору відповідної комбінації методів для вирішення проблеми.

Наявні публікації на схожу тематику пропонують використання методів машинного навчання (моделі ARIMA, GAM, SVR, модифіковані та комбіновані моделі), штучного інтелекту, тощо. Пропонуються різні підходи, але всі вони орієнтуються на наближення значень згенерованої моделі до фактичних значень, зменшення значення похибок, збільшення точності. Перевагою використання моделей ARIMA є їх універсальність і те, що їх можна налаштувати для прогнозування різних часових явищ. Можна коригувати різноманітні дані часових рядів за допомогою AR (авторегресійних моделей), MA (ковзних середніх) або ARMA, що є комбінацією AR і MA. Завдяки своїй ефективності як класичного

методу прогнозування він широко використовується, наприклад, для прогнозування якості води, робочих навантажень у хмарних додатках, індексу EBITDA для фінансових показників і короткострокових навантажень клієнтів. Оскільки ARIMA є лінійною моделлю, вона не може охопити нелінійні моделі в часовому ряді; отже, для прогнозування сильних коливань у прогнозуванні часових рядів використовуються різні прогностичні моделі, засновані на техніках машинного та глибокого навчання. Модель GAM відповідає власному спеціальному набору даних для обробки часових рядів і сезонності. Набір даних містить два основні стовпці: стовпець «ds», у якому зберігаються часові ряди дат, і інший стовпець «y», у якому зберігаються відповідні значення часових рядів у наборі даних.

Таким чином, використання різних підходів до розробки моделей прогнозування в єдиній інформаційній системі дозволяє отримати ефективний прогноз вартості комерційної компанії.

2 МАТЕМАТИЧНІ МОДЕЛІ, МЕТОДИ, ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ, ЩО ВИКОРИСТОВУЮТЬСЯ ДЛЯ ПРОГНОЗУВАННЯ ВАРТОСТІ КОМЕРЦІЙНИХ КОМПАНІЙ

2.1 Машинне навчання

Машинне навчання (ML) прагне автоматично вивчати значущі зв’язки та шаблони на прикладах і спостереженнях [21, 22]. Досягнення в ML сприяли нещодавньому зростанню інтелектуальних систем з людськими когнітивними можливостями, які проникають у наше ділове та особисте життя та формують мережеву взаємодію на електронних ринках усіма мислимими способами, завдяки чому компанії покращують процес прийняття рішень для продуктивності, залученості та утримання співробітників, системи помічників, які можна навчити, адаптуються до індивідуальних уподобань користувачів, а торгові агенти похитують традиційні ринки фінансової торгівлі.

Крім розкрученого вигляду, науковцям, як і професіоналам, потрібне глибоке розуміння базових концепцій, процесів, а також викликів для впровадження такої технології. На цьому фоні ідея полягає в тому, щоб передати розуміння машинного навчання та глибокого навчання (DL) у контексті електронних ринків. Таким чином, спільнота може отримати вигоду з цих технологічних досягнень — чи то для вивчення великих і високорозмірних активів даних, зібраних у цифрових екосистемах, чи для розробки нових інтелектуальних систем для електронних ринків. Щоб забезпечити розуміння галузі, необхідно розрізняти кілька відповідних термінів і понять один від одного (див. рис. 2.1) [23].

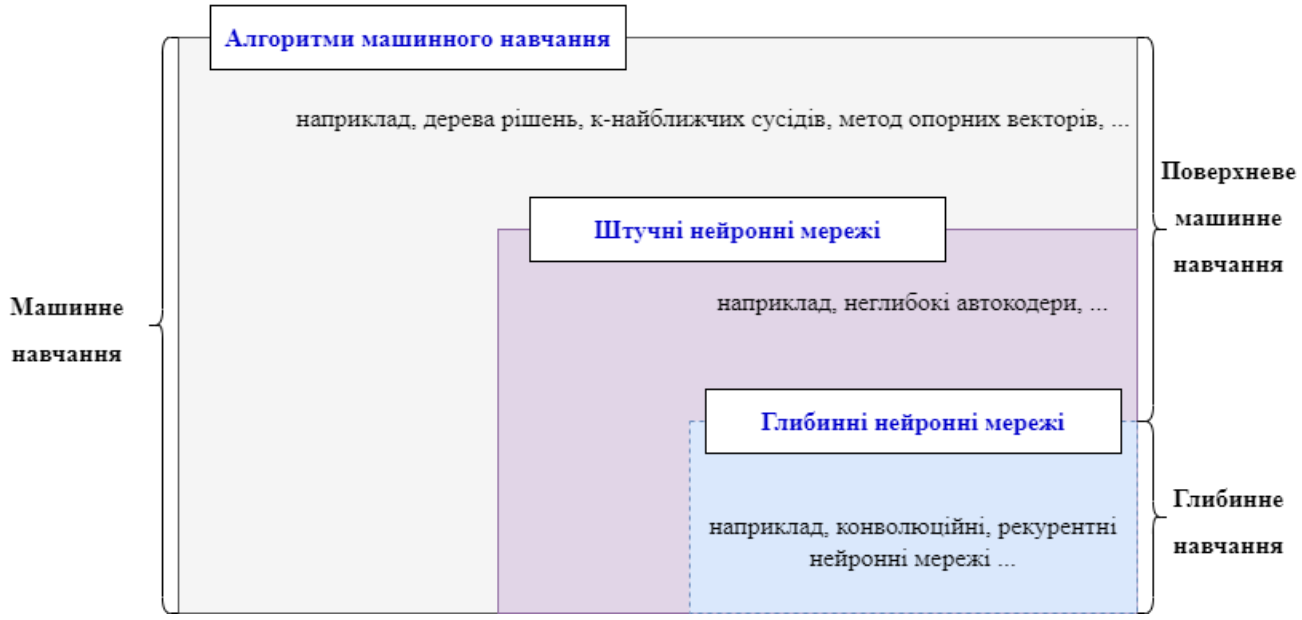

Рисунок 2.1 – Діаграма Венна концепцій і класів машинного навчання

У широкому розумінні штучний інтелект включає в себе будь-яку техніку, яка дозволяє комп’ютерам імітувати людську поведінку та відтворювати або перевершувати прийняття людських рішень для вирішення складних завдань самостійно або з мінімальним людським втручанням [24]. Як таке, воно пов’язане з різними центральними проблемами, включаючи представлення знань, міркування, навчання, планування, сприйняття та комунікацію, і стосується різноманітних інструментів і методів (наприклад, міркування на основі випадків, системи на основі правил, генетичні алгоритми, нечіткі моделі, мультиагентні системи) [25]. Ранні дослідження штучного інтелекту зосереджувались насамперед на жорстко закодованих висловлюваннях у формальних мовах, які потім комп’ютер може автоматично вираховувати на основі правил логічного висновку. Ця подія також відома як підхід бази знань [23]. Однак парадигма стикається з кількома обмеженнями, оскільки людям, як правило, важко пояснити всі свої неявні знання, необхідні для виконання складних завдань [26].

Машинне навчання долає такі обмеження. Загалом, ML означає, що продуктивність комп’ютерної програми покращується з досвідом щодо певного класу завдань і показників ефективності [27-29]. Таким чином, він спрямований

на автоматизацію завдання побудови аналітичної моделі для виконання когнітивних завдань, таких як виявлення об’єктів або переклад природною мовою. Це досягається шляхом застосування алгоритмів, які ітеративно вивчають дані навчання, що стосуються конкретної проблеми, що дозволяє комп’ютерам знаходити приховані ідеї та складні шаблони без явного програмування [30]. Особливо в задачах, пов’язаних з великомірними даними, такими як класифікація, регресія та кластеризація, ML демонструє хорошу застосовність. Вивчаючи попередні обчислення та витягуючи закономірності з масивних баз даних, це може допомогти отримати надійні та повторювані рішення. З цієї причини алгоритми ML були успішно застосовані в багатьох сферах, таких як виявлення шахрайства, оцінка кредитоспроможності, аналіз наступної найкращої пропозиції, розпізнавання мови та зображень або обробка природної мови (NLP).

Виходячи з поставленої проблеми та наявних даних, ми можемо виділити три типи ML: навчання з вчителем, навчання без вчителя та навчання з підкріпленням (табл. 2.1) [21].

Таблиця 2.1 – Огляд типів машинного навчання

Тип	Опис
Навчання з вчителем	Навчання з вчителем вимагає навчального набору даних, який охоплює приклади для вхідних даних, а також відповіді з мітками або цільові значення для вихідних даних. Прикладом може бути передбачення кількості активних користувачів, підписаних на ринкову платформу протягом місяця, як результат (вважається цільовою змінною або змінною y) на основі різних вхідних характеристик, таких як кількість проданих продуктів або позитивні відгуки користувачів (часто згадуються як вхідні функції або змінні x). Потім пари вхідних і вихідних даних у навчальному наборі використовуються для калібрування відкритих параметрів моделі ML. Коли модель успішно

Тип	Опис
	навчена, її можна використовувати для прогнозування цільової змінної y з урахуванням нових або невидимих точок даних вхідних функцій x. Що стосується типу навчання з вчителем, ми можемо додатково розрізнити проблеми регресії, де прогнозується числове значення (наприклад, кількість користувачів), і проблеми класифікації, де результатом передбачення є категорична класова приналежність, наприклад «глядачі» або «покупці». ”.
Навчання без вчителя	Навчання без вчителя має місце, коли система навчання має виявляти шаблони без будь-яких попередніх позначок чи специфікацій. Таким чином, навчальні дані складаються лише зі змінних x з метою пошуку цікавої структурної інформації, такої як групи елементів, які мають спільні властивості (відомі як кластеризація), або представлення даних, які проектуються з простору великої розмірності в нижчий (відоме як зменшення розмірності). Яскравим прикладом неконтрольованого навчання на електронних ринках є застосування методів кластеризації для групування клієнтів або ринків у сегменти з метою більш конкретного спілкування з цільовою групою.
Навчання з підкріпленням	У системі навчання з підкріпленням замість того, щоб надавати пари вхідних і вихідних даних, ми описуємо поточний стан системи, визначаємо мету, надаємо перелік дозволених дій та їх обмежень середовища для їх результатів, і дозволяємо моделі ML випробувати процес самостійне досягнення мети за принципом проб і помилок для максимізації винагороди. Моделі навчання з підкріпленням з великим успіхом застосовуються в середовищах закритого світу, таких як ігри, але вони також актуальні для багатоагентних систем, таких як електронні ринки.

2.2 Статистичний аналіз

Статистичний аналіз – це процес збору та аналізу даних з метою виявлення закономірностей і тенденцій [31]. Це компонент аналітики даних. Це метод для усунення упередженості в оцінюванні даних за допомогою чисельного аналізу. Ця техніка корисна для збору інтерпретацій досліджень, розробки статистичних моделей і планування опитувань і досліджень. Статистичний аналіз можна використовувати в таких ситуаціях, як збір інтерпретацій досліджень, статистичне моделювання або планування опитувань і досліджень. Це також може бути корисним для організацій бізнес-аналітики, яким доводиться працювати з великими обсягами даних.

Статистичний аналіз – це науковий інструмент, який допомагає збирати та аналізувати великі обсяги даних, щоб визначити загальні закономірності, тенденції та перетворити їх на значущу інформацію. Простими словами, статистичний аналіз – це інструмент аналізу даних, який допомагає зробити важливі висновки з необроблених і неструктурованих даних.

Висновки зроблені за допомогою статистичного аналізу, який полегшує прийняття рішень і допомагає підприємствам робити прогнози на майбутнє на основі минулих тенденцій. Його можна визначити як науку про збір і аналіз даних для виявлення тенденцій і закономірностей і їх представлення. Статистичний аналіз передбачає роботу з числами та використовується підприємствами та іншими установами для використання даних для отримання суттєвої інформації.

Нижче наведено 6 типів статистичного аналізу (табл. 2.2) [32]:

Таблиця 2.2 – Типи статистичного аналізу

Тип	Опис
Описовий аналіз	Описовий статистичний аналіз передбачає збір, інтерпретацію, аналіз та узагальнення даних для представлення їх у формі діаграм, графіків і таблиць. Замість

Тип	Опис
	того, щоб робити висновки, він просто робить складні дані легкими для читання та розуміння.
Інференційний аналіз	Інференційний статистичний аналіз спрямований на отримання значущих висновків на основі проаналізованих даних. Він вивчає зв’язок між різними змінними або робить прогнози для всієї сукупності.
Прогнозний аналіз	Прогнозний статистичний аналіз – це тип статистичного аналізу, який аналізує дані для визначення минулих тенденцій і прогнозування майбутніх подій на їх основі. Він використовує алгоритми машинного навчання, аналіз даних, моделювання даних і штучний інтелект для проведення статистичного аналізу даних.
Наказовий аналіз	Наказовий аналіз проводить аналіз даних і призначає найкращий курс дій на основі результатів. Це тип статистичного аналізу, який допомагає прийняти обґрунтоване рішення.
Дослідницький аналіз даних	Дослідницький аналіз схожий на інференційний аналіз, але відмінність полягає в тому, що він включає дослідження невідомих асоціацій даних. Він аналізує потенційні зв’язки в даних.
Причинно-наслідковий аналіз	Причинно-наслідковий статистичний аналіз зосереджується на визначенні причинно-наслідкового зв’язку між різними змінними в необроблених даних. Простими словами, він визначає, чому щось відбувається, і його вплив на інші змінні. Ця методологія може бути використана компаніями для визначення причини невдачі.

Статистичний аналіз можна назвати благом для людства, і він має багато переваг як для окремих осіб, так і для організацій. Нижче наведено деякі з причин, чому вам варто розглянути можливість інвестування в статистичний аналіз:

а) Це може допомогти вам визначити місячні, квартальні, річні показники прибутку від продажів і витрат, що полегшує прийняття рішень;

б) Це може допомогти вам прийняти обґрунтовані та правильні рішення;

в) Це може допомогти вам визначити проблему або причину збою та внести виправлення. Наприклад, він може визначити причину збільшення загальних витрат і допомогти вам скоротити марнотратні витрати;

г) Це може допомогти вам провести аналіз ринку та скласти ефективну стратегію маркетингу та продажів;

ґ) Це допомагає підвищити ефективність різних процесів.

2.3 Середовище розробки

Для вирішення поставленої задачі було обрано середовище розробки RGui. RGui базується на R, де R – мова програмування і програмне середовище для статистичних обчислень, аналізу та зображення даних в графічному вигляді. Розробка R відбувалась під істотним впливом двох наявних мов програмування: мови програмування S з семантикою успадкованою від Scheme. R названа за першою літерою імен її засновників Роса Іхаки (Ross Ihaka) та Роберта Джентлмена (Robert Gentleman) працівників Оклендського Університету в Новій Зеландії.

Якщо коротко, то [33]:

а) R – це «мова та середовище для статистичних обчислень і графіки»; ви можете думати про це як про комбінацію пакету статистики та мови програмування;

б) R повністю безкоштовне; вам не потрібно платити за це, і ви можете вносити в нього будь-які зміни;

в) R працює на Windows, MacOS, Linux і багатьох варіантах Unix;

г) R не підтримується жодним комерційним підприємством, але він має дуже активну спільноту розробників. Посібники повні, і про це середовище існує багато підручників;

ґ) У R є величезна кількість вбудованих стандартних і найсучасніших статистичних функцій, широкий вибір (безкоштовних) додаткових пакетів, які збільшують функціонал, і ви можете ще розширити їх. Кожен стандартний статистичний аналіз можна виконати в R;

д) R здебільшого керується командним рядком (хоча були розроблені різні графічні інтерфейси); це ускладнює використання, але забезпечує гнучкість, документування та повторення аналізів.

R має значні можливості для здійснення статистичних аналізів, включаючи лінійну і нелінійну регресію, класичні статистичні тести, аналіз часових рядів (серій), кластерний аналіз і багато іншого. R легко розбудовується завдяки використанню додаткових функцій і пакетів доступних на сайті Comprehensive R Archive Network.

Багато з вищевказаних особливостей середовища розробки/мови програмування можна приписати й іншим представникам тієї чи іншої сфери, тому необхідним є вказати саме ті переваги, які відрізняють R від усіх своїх конкурентів. Перш за все – R це мова створена саме для статистичного аналізу, тому в ній вже спрощено роботу з усіма видами числової інформації. Наприклад, в ній прибрано необхідність опрацьовувати базові завдання з масивами через цикли, замість цього є можливість безпосередньої роботи як із стовбцями, так і з рядками даних. По-друге, R – сучасна мова та середовище з великою спільнотою. В історії вже зустрічалися мови, що спеціалізуються саме на математичній та статистичній обробці даних, але вони або застарілі, або не мають можливості використовувати сучасні засоби обробки типів наборів даних. По-третє, в R реалізована потужна графічна база, що дозволяє ефективно візуалізувати дані та їх окремі компоненти. Якщо потужності замало, то на допомогу приходять

безкоштовні бібліотеки, що на додачу до потужності надають і зручність візуалізації, за необхідності.

Саме завдяки такому набору характеристик, середовище та мова R були обрані для виконання поставленої задачі. У наступних розділах буде присвячено більше уваги бібліотекам, що дозволяють легко та якісно налаштовувати моделі для прогнозування. Вибір бібліотек також заснований на сучасних методах обробки та перетворення даних, в тому числі tidyverse. Перетворення даних у цей формат вимагає певної попередньої роботи, але ця робота окупається в довгостроковій перспективі. Отримавши чіткі дані та акуратні інструменти, надані пакетами в tidyverse, витрати часу в майбутньому на переміщення даних з одного представлення в інше зменшуються, що дозволить вам витрачати більше часу на наявні аналітичні запитання.

2.4 Математичні моделі

2.4.1 ARIMA

У статистиці та економетриці, зокрема в аналізі часових рядів, модель авторегресійної інтегрованої ковзної середньої (ARIMA) є узагальненням моделі авторегресійної ковзної середньої (ARMA). Обидві ці моделі адаптуються до даних часових рядів або для кращого розуміння даних, або для прогнозування майбутніх точок у ряді (прогнозування) [34, 35]. Експоненціальне згладжування та моделі ARIMA є двома найбільш широко використовуваними підходами до прогнозування часових рядів, які забезпечують додаткові підходи до проблеми [36]. У той час як моделі експоненціального згладжування базуються на описі тенденції та сезонності в даних, моделі ARIMA спрямовані на опис автокореляції в даних.

Моделі ARIMA застосовуються в деяких випадках, коли дані показують докази нестаціонарності в сенсі середнього (але не дисперсії/автоковаріації), де початковий крок розрізнення (що відповідає «інтегрованій» частині моделі) може

бути застосований один або більше разів, щоб усунути нестаціонарність середньої функції (тобто тренду) [34]. Коли сезонність відображається в часовому ряді, можна застосувати сезонну різницю, щоб усунути сезонний компонент. Оскільки модель ARMA, згідно з теоремою декомпозиції Уолда, теоретично достатня для опису регулярного (він же чисто недетермінованого) широкого стаціонарного часового ряду, ми мотивовані зробити стаціонарним нестаціонарний часовий ряд, наприклад, використовуючи розрізнення, перш ніж ми зможемо використовувати модель ARMA. Зауважте, що якщо часовий ряд містить передбачуваний підпроцес (він же чистий синус або комплексно-значний експоненціальний процес), передбачуваний компонент розглядається як ненульовий середній, але періодичний (тобто сезонний) компонент у структурі ARIMA, так що це усувається сезонною різницею.

Частина AR ARIMA вказує на те, що змінна, яка цікавить, регресує на її власні відсталі (тобто попередні) значення. Частина MA вказує на те, що помилка регресії насправді є лінійною комбінацією членів помилки, значення яких мали місце одночасно та в різний час у минулому. I (для «інтегрованого») вказує на те, що значення даних було замінено на різницю між їхніми значеннями та попередніми значеннями (і цей процес розрізнення міг виконуватися більше одного разу). Мета кожної з цих функцій полягає в тому, щоб модель якомога краще відповідала даним.

Несезонні моделі ARIMA зазвичай позначаються як ARIMA(p, d, q), де параметри p, d і q є невід’ємними цілими числами, p – порядок (кількість часових лагів) авторегресійної моделі, d – ступінь різницю (кількість разів, коли дані мали віднімання минулих значень), і q є порядком моделі ковзного середнього. Сезонні моделі ARIMA зазвичай позначаються як ARIMA (p, d, q) (P, D, Q)m, де m означає кількість періодів у кожному сезоні, а великі літери P, D, Q позначають авторегресію, різницю, і умови ковзного середнього для сезонної частини моделі ARIMA.

Загалом існує дві моделі для ARIMA залежно від того, стаціонарний чи нестаціонарний набір даних. Несезонну модель ARIMA можна записати як [25]

$$(1 - \phi 1B - \cdots - \phi_p B^p)(1 - B)^d y_t = c + (1 + \theta_1 B + \ldots + \theta_q B^q)\epsilon_t \quad 2.1$$

де $c = \mu(1 - \phi_1 - \cdots - \phi_p)$ і μ середнє $(1 - B)^d y_t$.

Наведене вище рівняння можна записати таким чином:

$$\Phi(B)(1 - B)^d y_t = c + \theta(B)\epsilon_t \quad 2.2$$

де $\Phi(B) = (1 - \phi_1 B - \cdots - \phi_p B^p)$ є поліномом p^{t^h} порядку у B, а $\theta(B) = (1 + \theta_1 B + \theta_q B^q)$ є поліномом q^{t^h} порядку у B. Несезонну модель ARIMA можна записати як ARIMA(p, d, q). Для сезонного набору даних, сезонну модель ARIMA можна сформувати шляхом включення додаткових сезонних термів у модель ARIMA, і її можна записати так: $ARIMA(p, d, q)(P, D, Q)m$ де m = число спостережень на рік. Наприклад, для квартальних даних, коли ($p = 1$, $d = 1$, $q = 1$, $P = 1$, $D = 1$, $Q = 1$) сезонна модель ARIMA може бути записана таким чином [14]:

$$(1 - \phi_1 B)(1 - \Phi_1 B^{12})(1 - B)(1 - B^{12})y_t = (1 + \theta_1 B)(1 + \Theta_1 B^{12})\epsilon_t \quad 2.3$$

Наведене вище рівняння позначено як $ARIMA(1,1,1)(1,1,1)_{12}$. Мультиплікативну сезонну модель ARIMA задано [5]

$$\Phi_P (B^s)\phi(B)\nabla_s{}^D \nabla^d y_t = \delta + \Theta_Q (B^s)\theta(B)\epsilon_t , \quad 2.4$$

де $\Phi_P (B^s)$: сезонна авторегресія порядку P;

$\phi(B)$: звичайна авторегресія порядку p;

$\Theta_Q (B^s)$: компоненти ковзного середнього порядку Q;

$\theta(B)$: компоненти ковзного середнього порядку q;

$\nabla_s^D = (1 - B)^{SD}$: компонент сезонної різниці;

$\nabla^d = (1 - B)^d$: порядковий диференційний компонент;

ϵ_t: звичайний гаусівський процес білого шуму.

Сезонну частину моделі AR або MA можна буде побачити в сезонних лагах PACF і ACF. Наприклад, модель ARIMA$(0,0,0)(0,0,1)_{12}$ покаже:

а) сплеск на лазі 12 у ACF, але немає інших значних сплесків;

б) експоненціальний спад у сезонних лагах PACF (тобто на лагах 12, 24, 36 і т.д.).

Подібним чином модель ARIMA$(0,0,0)(1,0,0)_{12}$ покаже:

а) експоненціальне загасання сезонних лагів ACF;

б) один значний сплеск на лазі 12 у PACF.

Розглядаючи відповідні сезонні послідовності для сезонної моделі ARIMA, треба приділити увагу сезонним затримкам.

Процедура моделювання майже така ж, як і для несезонних даних, за винятком того, що нам потрібно вибрати сезонні умови AR і MA, а також несезонні компоненти моделі.

Модифікована модель ARIMA може бути узагальнена в наступному алгоритмі [5],

Модифікований алгоритм ARIMA

Крок 1. Визначити закономірності, побудувавши дані та виявивши незвичайні спостереження.

Крок 2. Використати перетворення Бокса-Кокса, щоб стабілізувати дисперсію (якщо необхідно).

Крок 3. Диференціювати дані, щоб вони були стаціонарними (за потреби)

Крок 4. Побудувати графік ACF диференційованих даних і спробувати визначити можливі моделі-кандидати.

Крок 5. Використати метрику AIC для визначення кращої моделі.

Крок 6 Перевірте залишки з моделі, побудувавши графік ACF залишків і виконавши тест портманто.

Крок 7. Чи виглядають залишки як білий шум? Якщо ТАК (розрахувати прогнози), НІ (перейти до кроку 4).

2.4.2 GAM

Для багатьох застосувань статистичного аналізу Загальна лінійна модель використовується як ключова модель у прикладних і соціальних дослідженнях. Для інтерпретації та дослідження параметрів статистичні моделі забезпечують математичну основу і, для конкретних процесів, встановлюють ролі та порівняльну важливість різних змінних [37]. Рис. 2.2 дає загальне уявлення про різні доступні моделі на основі регресії та підкреслює основні переваги перед іншими моделями.

Рисунок 2.2 – Загальний огляд і зв'язок між моделями регресії.

Математичне узагальнення множинних регресій можна трансформувати як загальні лінійні моделі та уточнити під ними [38-40]. Зокрема, лінійний метод

найменших квадратів використовується для прогнозування залежної змінної Y для набору предикторів або змінних X у лінійній регресії.

Іншими словами, передбачити найкращу оцінку предикторів і адитивної моделі для невизначеної моделі; непараметрична функція предиктора визначається як сурогат одного коефіцієнта для кожної змінної. Математичне узагальнення множинних регресій можна трансформувати як загальні лінійні моделі та уточнити під ними. Зокрема, лінійний метод найменших квадратів використовується для прогнозування залежної змінної Y для набору предикторів або змінних X у лінійній регресії.

Іншими словами, передбачити найкращу оцінку предикторів і адитивної моделі для невизначеної моделі; непараметрична функція предиктора визначається як сурогат одного коефіцієнта для кожної змінної.

GAM є класичним додатком загальних лінійних моделей [41, 42]. Інша робота показала, що GLM, що має лінійний предиктор, взаємодіє з сумою гладких функцій коваріат [43]. GAM забезпечує структуру для узагальнення загальної лінійної моделі, дозволяючи адитивність нелінійних функцій змінних.

Крім того, перевага GAM полягає в тому, щоб обмежити помилку в передбаченні залежної змінної Y від різних розподілів шляхом оцінки неспецифічних функцій, які пов’язані за допомогою функції зв’язку із залежною змінною.

GAM забезпечує гнучку специфікацію відповіді, визначаючи модель у термінах гладкої функції як заміну детальних параметричних зв’язків на коваріатах. Ця гнучкість і доцільність досягаються ціною представлення гладких функцій у подібному шаблоні та вибору рівня гладкості.

Недавньою пропозицією є модель Prophet, доступна через пакет fable.prophet. Ця модель була представлена Facebook (S. J. Taylor & Letham, 2018) спочатку для прогнозування щоденних даних із тижневою та річною сезонністю, а також ефектами свят. Пізніше його було розширено, щоб охопити більше типів

сезонних даних. Він найкраще працює з часовими рядами, які мають сильну сезонність і кілька сезонів історичних даних.

Prophet можна вважати нелінійною регресійною моделлю виду

$$y_t=g(t)+s(t)+h(t)+\varepsilon_t, \quad 2.6$$

де g(t) описує кусково-лінійний тренд (або «член зростання»), s(t) описує різні сезонні закономірності, h(t) фіксує ефект відпустки, а ε_t є членом помилки білого шуму.

а) Вузли (або точки зміни) для кусково-лінійного тренду вибираються автоматично, якщо не вказано явно. За бажанням можна використовувати логістичну функцію для встановлення верхньої межі тенденції.

б) Сезонна складова складається з членів Фур'є відповідних періодів. За замовчуванням порядок 10 використовується для річної сезонності, а порядок 3 використовується для тижневої сезонності.

в) Святкові ефекти додаються як прості фіктивні змінні.

г) Модель оцінюється за допомогою байєсівського підходу, щоб забезпечити автоматичний вибір точок зміни та інших характеристик моделі.

Висновки до розділу 2

Машинне навчання (ML) прагне автоматично вивчати значущі зв’язки та шаблони на прикладах і спостереженнях. Досягнення в ML сприяли нещодавньому зростанню інтелектуальних систем з людськими когнітивними можливостями, які проникають у наше ділове та особисте життя та формують мережеву взаємодію на електронних ринках усіма мислимими способами, завдяки чому компанії покращують процес прийняття рішень для продуктивності, залученості та утримання співробітників, системи помічників, які можна навчити, адаптуються до індивідуальних уподобань користувачів, а торгові агенти похитують традиційні ринки фінансової торгівлі.

Статистичний аналіз – це науковий інструмент, який допомагає збирати та аналізувати великі обсяги даних, щоб визначити загальні закономірності, тенденції та перетворити їх на значущу інформацію. Простими словами, статистичний аналіз — це інструмент аналізу даних, який допомагає зробити важливі висновки з необроблених і неструктурованих даних.

R має значні можливості для здійснення статистичних аналізів, включаючи лінійну і нелінійну регресію, класичні статистичні тести, аналіз часових рядів (серій), кластерний аналіз і багато іншого. R легко розбудовується завдяки використанню додаткових функцій і пакетів доступних на сайті Comprehensive R Archive Network.

Моделі ARIMA застосовуються в деяких випадках, коли дані показують докази нестаціонарності в сенсі середнього (але не дисперсії/автоковаріації), де початковий крок розрізнення (що відповідає «інтегрованій» частині моделі) може бути застосований один або більше разів, щоб усунути нестаціонарність середньої функції (тобто тренду). Коли сезонність відображається в часовому ряді, можна застосувати сезонну різницю, щоб усунути сезонний компонент. Оскільки модель ARMA, згідно з теоремою декомпозиції Уолда, теоретично достатня для опису регулярного (недетермінованого) широкого стаціонарного часового ряду, ми мотивовані зробити стаціонарним нестаціонарний часовий ряд, наприклад, використовуючи розрізнення, перш ніж ми зможемо використовувати модель ARMA.

Перевага GAM полягає в тому, щоб обмежити помилку в передбаченні залежної змінної Y від різних розподілів шляхом оцінки неспецифічних функцій, які пов’язані за допомогою функції зв’язку із залежною змінною.

GAM забезпечує гнучку специфікацію відповіді, визначаючи модель у термінах гладкої функції як заміну детальних параметричних зв’язків на коваріатах.

3 МОДЕЛЮВАННЯ ТА ПРОГРАМНА РЕАЛІЗАЦІЯ МОДЕЛЕЙ І ПРОГНОЗІВ. ДОСЛІДЖЕННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

3.1 Аналіз вхідного набору даних

Набір даних представлений вартістю акцій компанії Amazon за 2016-2019 роки. Завантаживши дані, можна побачити, що вони представляють собою часовий ряд, визначений 3 змінними: дата, назва компанії, ціна.

Першим кроком у прогнозуванні є підготовка даних у правильному форматі. Цей процес може включати завантаження даних, визначення відсутніх значень, фільтрацію часових рядів та інші завдання попередньої обробки. Функціональність, яку надає tsibble та інші пакеті в tidyverse, значно спрощує цей крок.

Багато моделей мають різні вимоги до даних; деякі вимагають, щоб ряди були в порядку часу, інші вимагають відсутності пропущених значень. Перевірка ваших даних є важливим кроком для розуміння їх характеристик, і її слід завжди робити перед оцінкою моделей.

Після перетворення даних у tidy формат формуємо графічне представлення вибірки (див. рис. 3.1). Візуалізація є важливим кроком у розумінні даних. Для даних часових рядів очевидним графіком для початку є часовий графік. Тобто спостереження відображаються відносно часу спостереження, причому послідовні спостереження з’єднуються прямими лініями. Перегляд ваших даних дозволяє визначити загальні закономірності та згодом визначити відповідну модель.

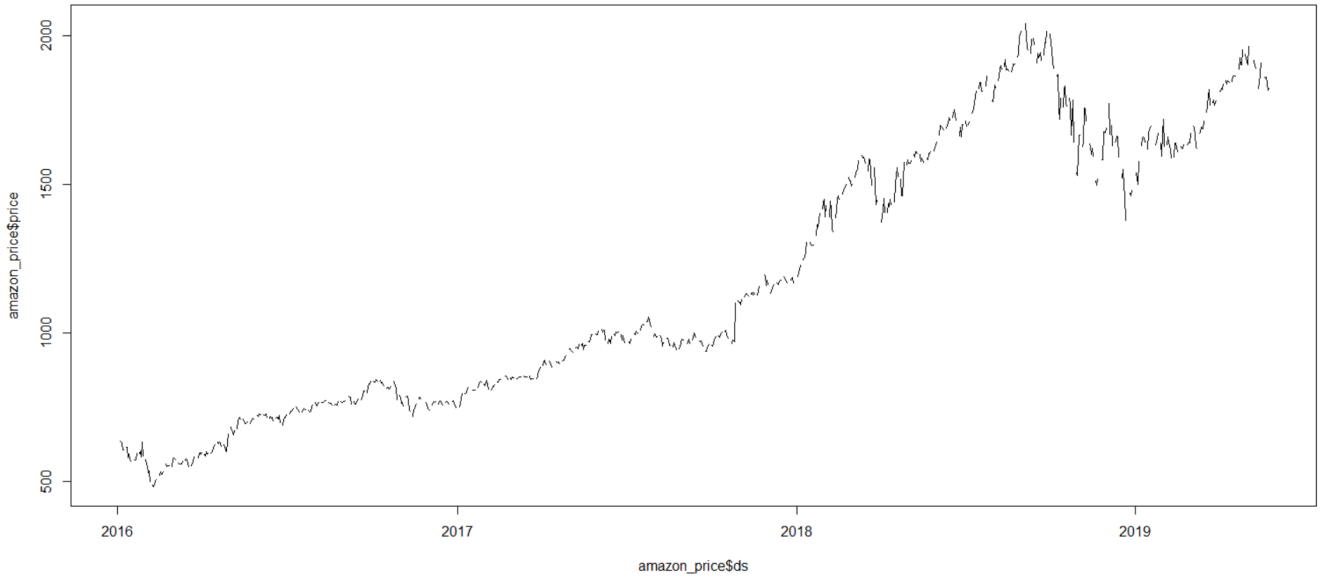

Рисунок 3.1 – Графічне представлення вибірки

Дані не мають чітко вираженої річної сезонності, тренд змінюється з плином часу рухаючись здебільшого вгору до третьої чверті 2018 року, має різкий спдний характер до початку 2019 і знову має направлення вгору до березня 2019 року. Остання частина представлена травнем, що має направлення тренду вниз.

Як видно з графіку, дані мають пропуски і не мають явних викидів. Завдяки існуючим алгоритмам в R, заповнення пропусків для наборів даних такого типу не є проблемою.

Перед проведенням процедури заповнення пропусків потрібно провести їх аналіз (див. рис. 3.2). Для проведення такого аналізу, використовується функція з пакету imputeTS, що виводить повну статистику по пропускам [44].

```
"Length of time series:"
1242
"-------------------------"
"Number of Missing Values:"
394
"-------------------------"
"Percentage of Missing Values:"
"31.7%"
"-------------------------"
"Number of Gaps:"
189
"-------------------------"
"Average Gap Size:"
2.084656
"-------------------------"
"Stats for Bins"
"  Bin 1 (311 values from 1 to 311) :      97 NAs (31.2%)"
"  Bin 2 (311 values from 312 to 622) :      97 NAs (31.2%)"
"  Bin 3 (311 values from 623 to 933) :      97 NAs (31.2%)"
"  Bin 4 (309 values from 934 to 1242) :      103 NAs (33.3%)"
"-------------------------"
"Longest NA gap (series of consecutive NAs)"
"3 in a row"
"-------------------------"
"Most frequent gap size (series of consecutive NA series)"
"2 NA in a row (occurring 151 times)"
"-------------------------"
"Gap size accounting for most NAs"
"2 NA in a row (occurring 151 times, making up for overall 302 NAs)"
"-------------------------"
"Overview NA series"
"  1 NA in a row: 11 times"
"  2 NA in a row: 151 times"
"  3 NA in a row: 27 times"
```

Рисунок 3.2 – Статистика пропусків

Статистика показує, що 31,7% вибірки – пусті значення, що виливається у 189 пропусків, з середнім розміром кожного у 2,084 значення. Найбільша кількість пустих значень у четвертій чверті, найдовша послідовність пустих значень – 3, найчастіший розмір пропусків – 2 з появою у 151 раз. До цих пропусків застосовуємо функцію заповнення.

Застосована функція використовує згладжування Калмана на моделях структурних часових рядів (або на представленні простору станів моделі ARIMA) для імпутації. Після виконання функції з заповнення пропусків потрібно знову візуально оцінити вибірку (див. рис. 3.3):

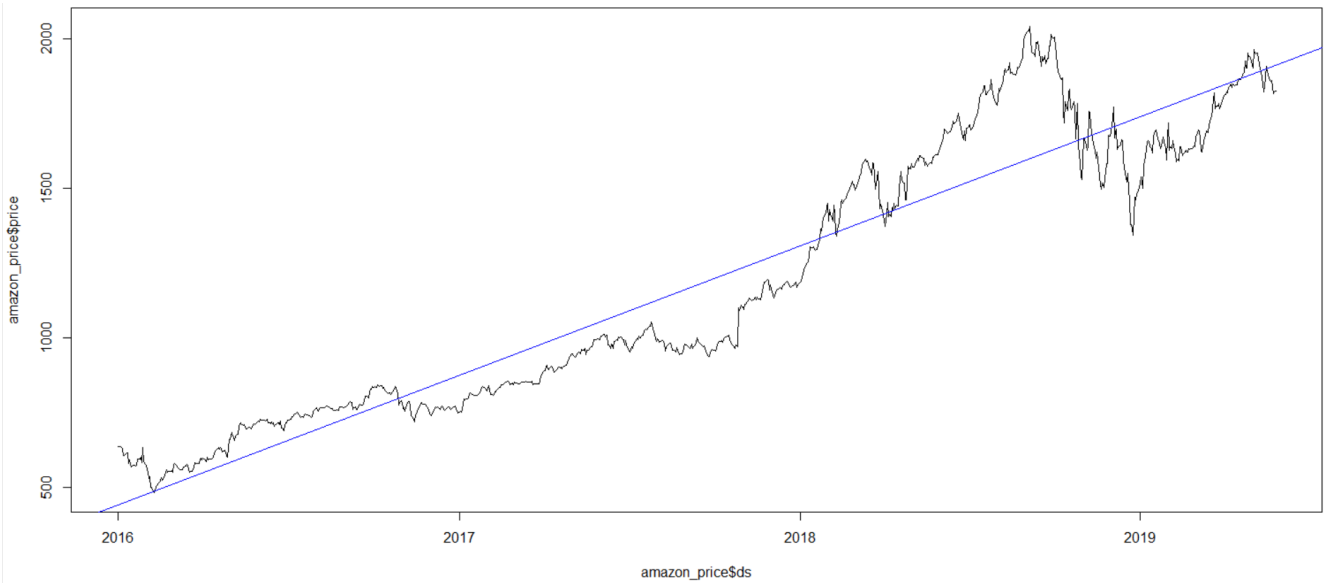

Рисунок 3.3 – Візуалізація вибірки після заповнення пропусків

Графік показує, що описані вище характеристики притаманні початковому набору даних зберігаються і для вибірки з заповненими пропусками. Пропуски заповнені без викидів, у межах можливих відхилень через використання лінійного регресійного відновлення даних закладених у функцію. Аналіз пропусків показав, що найбільшими є пропуски у 3 дні, тому використання лінійної регресії для відновлення є доцільним, оскільки досить точно заповнюються дані з використанням меншої кількості ресурсів. Регресійна лінія (синім кольором на графіку) показує загальний напрямок тренду та дозволяє припустити, що дані є нелінійними.

Отримана вибірка вже готова до роботи і тому її можна ділити на навчальну та тестову підвибірки. Продовжуємо роботу з навчальною. Одразу виконаємо перевірку на автокореляцію, виконавши візуалізацію ACF (див. рис. 3.4). Графік ACF також корисний для визначення нестаціонарних часових рядів. Для стаціонарного часового ряду ACF відносно швидко впаде до нуля, тоді як ACF нестаціонарних даних зменшується повільно у міру збільшення лагів. Якщо дані є сезонними, автокореляції будуть більшими для сезонних лагів (кратних сезонному періоду), ніж для інших лагів.

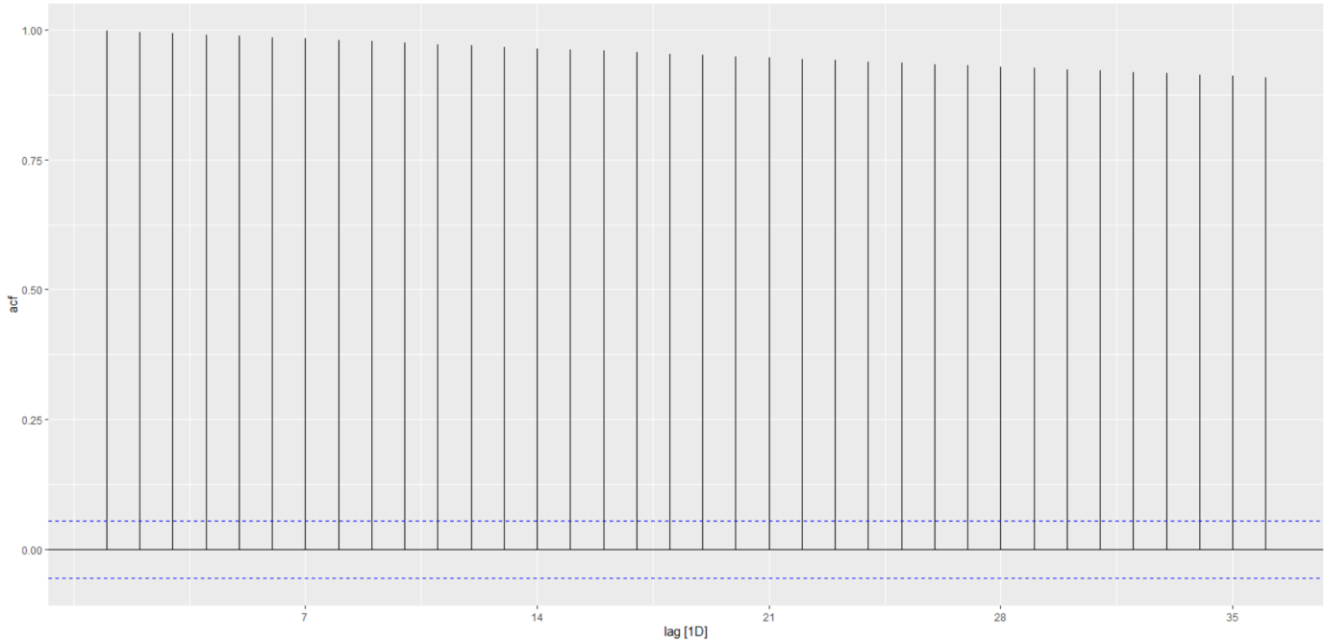

Рисунок 3.4 – Графік ACF для навчальної вибірки

Наведений графік показує вплив тренду з відсутньою сезонністю. Ми також можемо використовувати часткові автокореляції (див. рис. 3.5). Вони вимірюють зв’язок між y_t і y_{t-k} після усунення ефектів лагів 1, 2, 3, …, k−1. Таким чином, перша часткова автокореляція ідентична першій автокореляції, оскільки між ними немає нічого, що потрібно видалити. Кожну часткову автокореляцію можна оцінити як останній коефіцієнт авторегресійної моделі.

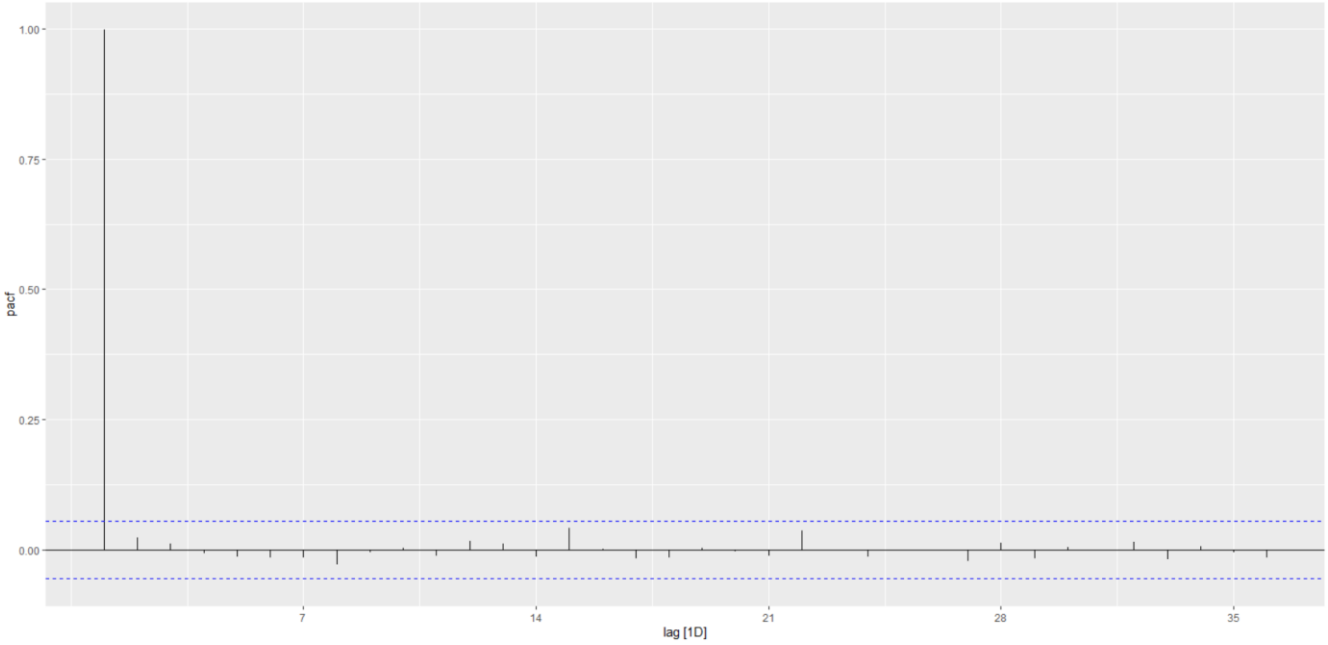

Рисунок 3.5 – Графік PACF для навчальної вибірки

Також необхідно провести тести на нелінійність на нестаціонарність. Набір тестів на нелінійність представлений функцією nonlinearityTest(). Функція виконує тестування за такими показниками:

а) Тест нейронної мережі Teraesvirta на нелінійність.

б) Тест нейронної мережі White на нелінійність.

в) Одноградусний тест Кінана на нелінійність.

г) Тест Маклеода-Лі для умовної гетероскедасності (ARCH).

г) Тест Цая на квадратичну нелінійність у часовому ряду.

д) Тест відношення правдоподібності для порогової нелінійності.

Результати тестування показали, що набір даних є нелінійним. Тест Маклеода-Лі показав 0 підтвердження гомоскедастичності в результаті тесту. Інші тести показали результати p-value: 0,0258; 0,0212; 0,0325; 0,032; 0,0429 відповідно (див. рис. 3.6).

```
** Teraesvirta's neural network test  **
Null hypothesis: Linearity in "mean"
X-squared =  7.310948  df =  2  p-value =  0.02584924

** White neural network test  **
Null hypothesis: Linearity in "mean"
X-squared =  7.705244  df =  2  p-value =  0.02122402

** Keenan's one-degree test for nonlinearity  **
Null hypothesis: The time series follows some AR process
F-stat =  4.579629  p-value =  0.03255081

** McLeod-Li test  **
Null hypothesis: The time series follows some ARIMA process
Maximum p-value =  0

** Tsay's Test for nonlinearity **
Null hypothesis: The time series follows some AR process
F-stat =  4.607871  p-value =  0.03202094

** Likelihood ratio test for threshold nonlinearity **
Null hypothesis: The time series follows some AR process
Alternativce hypothesis: The time series follows some TAR process
X-squared =  11.58424  p-value =  0.04292345
```

Рисунок 3.6 – Результати тестів на нелінійність

Перевірка на нестаціонарність виконується трьома тестами: Розширений тест Дікі-Фуллера, Тест KPSS на рівень стаціонарність, Тест Філіпса-Перрона на одиничний корінь (див. рис. 3.7).

```
> adf.test(amazon_price$y)

        Augmented Dickey-Fuller Test

data:  amazon_price$y
Dickey-Fuller = -2.4372, Lag order = 10, p-value = 0.3932
alternative hypothesis: stationary

> kpss.test(amazon_price$y)

        KPSS Test for Level Stationarity

data:  amazon_price$y
KPSS Level = 14.524, Truncation lag parameter = 7, p-value = 0.01

Предупреждение:
В kpss.test(amazon_price$y) : p-value smaller than printed p-value
> pp.test(amazon_price$y)

        Phillips-Perron Unit Root Test

data:  amazon_price$y
Dickey-Fuller Z(alpha) = -11.854, Truncation lag parameter = 7, p-value = 0.4485
alternative hypothesis: stationary
```

Рисунок 3.7 – Тести на не стаціонарність

Результати тестів відповідають сподіванням – два з трьох (тест Дікі-Фулера та KPSS) тестів показали нестаціонарність ряду і тест Філіпса-Перрона показав низький рівень довіри до стаціонарності.

Перейдемо до декомпозиції часового ряду. Описуючи ці часові ряди, ми використовували такі слова, як «тренд» та «сезонність», які потребують більш ретельного визначення.

Тренд

Тренд існує, коли спостерігається тривале збільшення або зменшення даних. Він не повинен бути лінійним. Іноді ми будемо називати тренд «змінним напрямком», коли він може переходити від тенденції зростання до тенденції до зниження.

Сезонність

Сезонна закономірність виникає, коли на часовий ряд впливають сезонні фактори, наприклад пора року або день тижня. Сезонність завжди має фіксований і відомий період.

Циклічність

Цикл виникає, коли дані демонструють зростання та спад, які не мають фіксованої частоти. Ці коливання зазвичай зумовлені економічними умовами та часто пов’язані з «діловим циклом». Тривалість цих коливань зазвичай становить не менше 2 років.

Багато людей плутають циклічну поведінку з сезонною поведінкою, але насправді вони зовсім різні. Якщо коливання не мають фіксованої частоти, то вони є циклічними; якщо частота незмінна і пов’язана з деяким аспектом календаря, то модель є сезонною. Загалом, середня тривалість циклів більша, ніж тривалість сезонної моделі, а величини циклів мають тенденцію бути більш мінливими, ніж величини сезонних моделей.

Багато часових рядів включають тренд, цикли та сезонність. Вибираючи метод прогнозування, нам спочатку потрібно буде визначити шаблони часових

рядів у даних, а потім вибрати метод, який здатний належним чином зафіксувати шаблони.

Дані часових рядів можуть демонструвати різноманітні моделі, і часто корисно розділити часовий ряд на кілька компонентів, кожен з яких представляє основну категорію шаблону.

Раніше обговорювалося три типи складових часових рядів: тренд, сезонність і цикли. Коли ми розкладаємо часовий ряд на компоненти, ми зазвичай об’єднуємо тренд і цикл в один компонент тренд-цикл (часто для спрощення його називають трендом). Таким чином, ми можемо розглядати часовий ряд як такий, що складається з трьох компонентів: компонент трендового циклу, сезонний компонент і компонент залишку (що містить будь-що інше в часовому ряді). Для деяких часових рядів (наприклад, тих, які спостерігаються принаймні щодня), може існувати більше одного сезонного компонента, що відповідає різним сезонним періодам.

Розглянемо один із методів виділення цих компонентів із часового ряду. Часто це робиться для покращення розуміння часових рядів, але це також можна використовувати для підвищення точності прогнозу.

Під час декомпозиції часового ряду іноді корисно спочатку трансформувати або скоригувати ряд, щоб зробити декомпозицію (і подальший аналіз) якомога простішим.

STL – універсальний і надійний метод декомпозиції часових рядів. STL – це абревіатура від «Seasonal and Trend decomposition using Loess», тоді як loess – це метод для оцінки нелінійних залежностей. STL має кілька переваг перед класичною декомпозицією та методами SEATS і X-11:

а) На відміну від SEATS і X-11, STL оброблятиме будь-який тип сезонності, а не лише місячні та квартальні дані.

б) Сезонна складова може змінюватися з часом, і швидкість зміни може контролювати користувач.

в) Плавність тренд-циклу також може контролювати користувач.

г) Він може бути стійким до викидів (тобто користувач може вказати надійне розкладання), так що випадкові незвичні спостереження не впливатимуть на оцінки циклу тренду та сезонних компонентів. Однак вони вплинуть на компонент залишку.

З іншого боку, STL має деякі недоліки. Зокрема, він не обробляє зміну торгового дня або календаря автоматично, і надає лише можливості для адитивної декомпозиції.

Декомпозицію часових рядів можна використовувати для вимірювання сили тенденції та сезонності в часових рядах. Вона записується як

$$y_t=T_t+S_t+R_t, \quad 3.1$$

де T_t – компонент згладженого тренду, S_t – сезонний компонент, а R_t – компонент залишку. Для даних із сильним трендом сезонно скориговані дані повинні мати набагато більше варіацій, ніж компонент залишку. Тому змінна $Var(R_t)/Var(T_t+R_t)$ має бути відносно малим. Але для даних із незначною тенденцією або без неї дві дисперсії мають бути приблизно однаковими. Отже, ми визначаємо силу тенденції як:

$$F_T = \max(0,1 - \frac{Var(R_t)}{Var(T_t+R_t)}) \quad 3.2$$

Це дасть міру сили тренду між 0 і 1. Оскільки дисперсія залишку іноді може бути навіть більшою, ніж дисперсія сезонно скоригованих даних, ми встановлюємо мінімально можливе значення F_T рівним нулю.

Сила сезонності визначається подібним чином, але стосовно даних з виключеним трендом, а не сезонно скоригованих даних:

$$F_S = \max(0,1 - \frac{Var(R_t)}{Var(S_t+R_t)}) \quad 3.3$$

Серія з сезонною силою F_S близькою до 0, майже не виявляє сезонності, тоді як серія з сильною сезонністю матиме F_S близьке до 1, оскільки $Var(R_t)$ буде набагато меншим за $Var(S_t+R_t)$.

Для обраного часового ряду декомпозиція виглядає наступним чином (див. рис. 3.8):

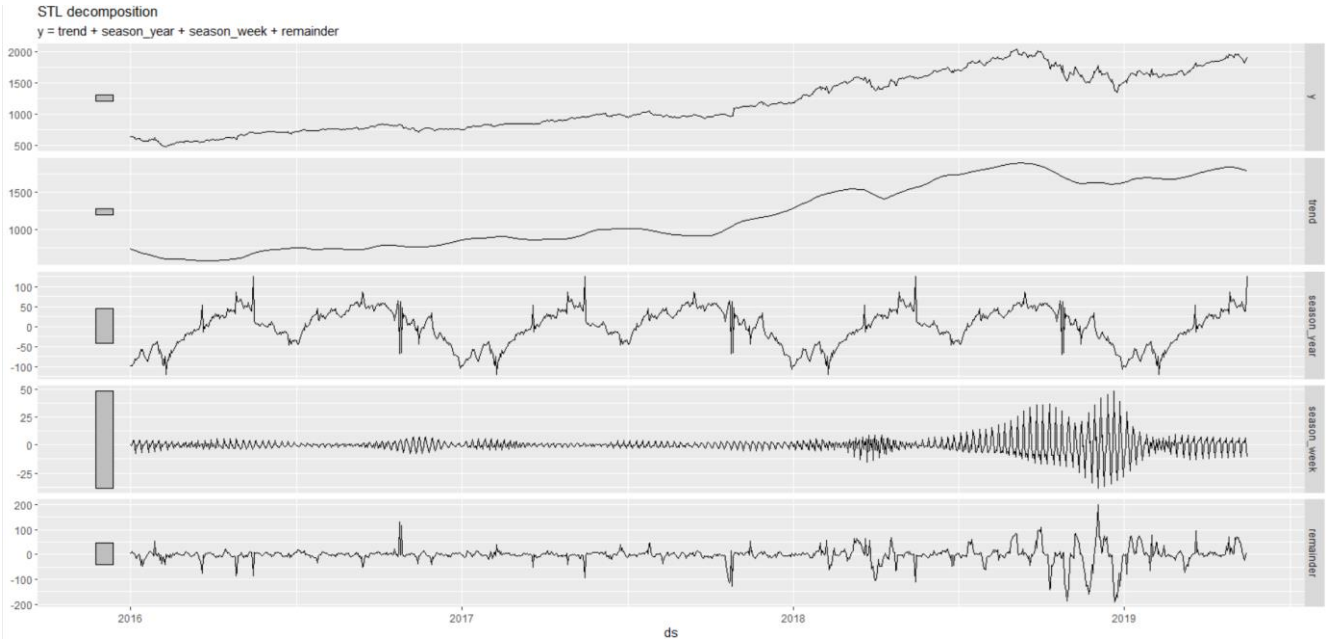

Рисунок 3.8 – Декомпозиція часового ряду навчальної вибірки

На рисунку видно, що дані мають сильний вплив сезонної компоненти, що представлена у двох виглядах: річній та тижневій. Декомпозиція добре передає тренд, що досить точно повторює зміни часового ряду.

3.2 Моделювання на основі модифікованого методу ARIMA

Моделювання ARIMA починається з визначення стаціонарності даних. В наведеному вище пункті вказано, що дані не є стаціонарними, тому було прийнято рішення провести диференціацію даних. Але перед цим потрібно визначити кількість таких диференціацій. Іноді диференційовані дані не виглядатимуть стаціонарними, і може знадобитися диференціювати їх вдруге, щоб отримати стаціонарний ряд. Тобто змоделювати «зміни в змінах» вихідних даних. Іноді звичайної диференціації може бути недостатньо, якщо у даних є сильний вплив сезонної компоненти, тому проводять сезонну диференціацію. Сезонна диференціація – це різниця між спостереженням і попереднім спостереженням того самого сезону. Для визначення кількості та типів диференціацій проводимо тест одиничного кореня. У статистиці перевірка

одиничного кореня перевіряє, чи змінна часового ряду є нестаціонарною та чи має одиничний корінь. Нульова гіпотеза зазвичай визначається як наявність одиничного кореня, а альтернативною гіпотезою є стаціонарність, стаціонарність тенденції або вибуховий корінь залежно від використовуваного тесту. Для виконання такого тесту виконується у два кроки: визначення кількості необхідних первинних диференціацій; визначення кількості сезонних диференціацій. Це процес використання послідовності тестів KPSS для визначення відповідної кількості перших диференціацій виконується за допомогою функції unitroot_ndiffs(). Подібною функцією для визначення того, чи потрібна сезонна диференціація, є unitroot_nsdiffs(), яка використовує міру сезонної сили для визначення необхідної кількості сезонних диференціацій.

```
> unitroot_ndiffs(amazon_price$y)
ndiffs
     1
> unitroot_nsdiffs(amazon_price$y)
nsdiffs
      0
```

Рисунок 3.9 – Проведення тестування одиничного кореня

Отримані результати показують, що потрібно провести одну первинну диференціацію (див. рис. 3.10).

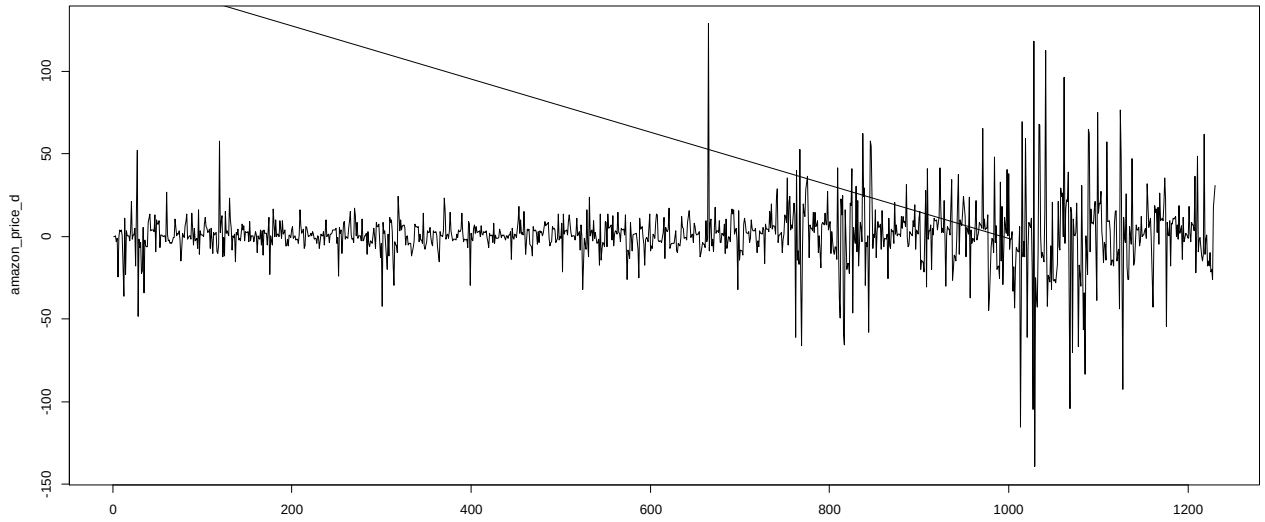

Рисунок 3.10 – Диференційовані дані

Отриманий результат нагадує стаціонарні дані, тому можна переходити до перевірок (див. рис. 3.11-3.12).

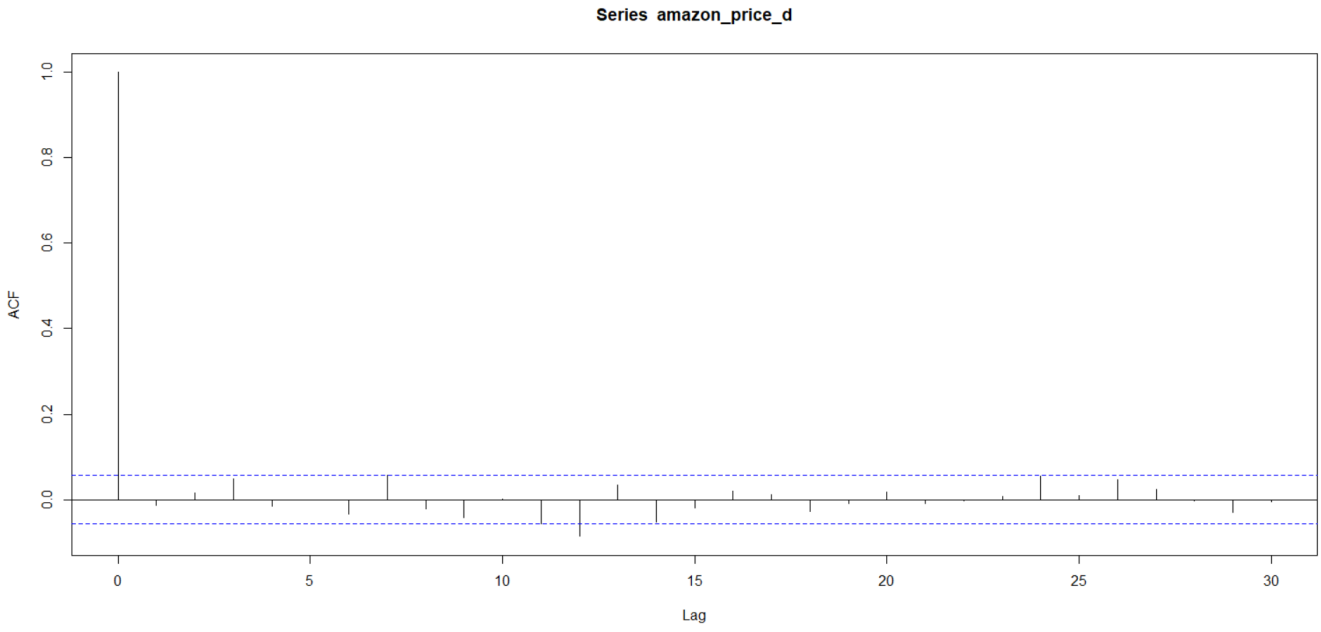

Рисунок 3.11 – ACF диференційованих даних

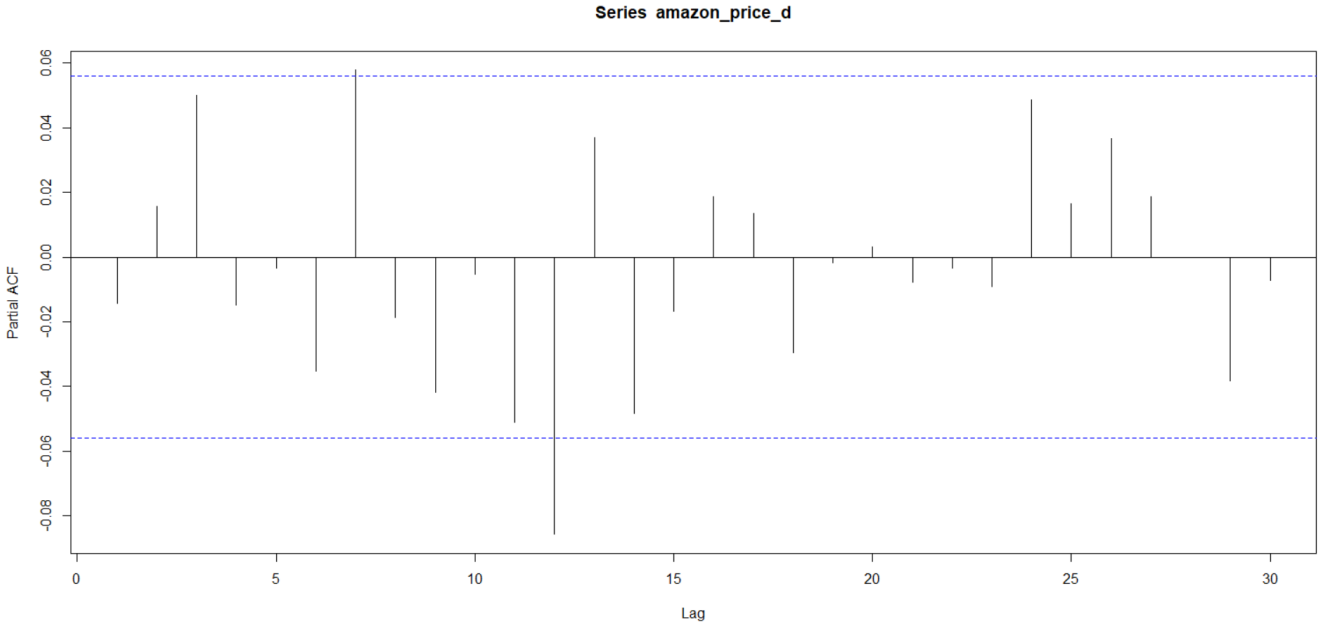

Рисунок 3.12 – PACF диференційованих даних

Дані ACF показують стаціонарність, а PACF майже повністю відповідають вигляду білого шуму. Тепер можна підбирати моделі. Одразу створимо моделі автоматичним пошуком і, за результатами побудованих ACF та PACF, відберемо моделі-кандидати вручну. Після визначення відповідної моделі ми далі навчаємо її на обраних даних. Одну або кілька специфікацій моделі можна оцінити за допомогою функції model().

Автоматичний підбір, що був заснований на методах: повного перебору; швидкого перебору; перебору із згладжуванням вхідного набору, показав такі моделі:

Таблиця 3.1 – Моделі ARIMA підібрані автоматично

Несезонна складова (p, d, q)	Сезонна складова (P, D, Q)	AIC	BIC
c(0, 1, 0)	c(2, 0, 0)	10642,26	10662,72
c(0, 1, 0)	c(0, 0, 2)	10642,32	10662,78
c(0, 1, 2)	c(2, 0, 2)	-7016,39	-6975,46

Ручний підбір, за результатами побудованих ACF та PACF, показав такі моделі з результатами тестування за інформаційними критеріями Акаіке та Байєса:

Таблиця 3.2 – Результат ручного підбору моделей-кандидатів для модифікованого методу ARIMA

Несезонна складова (p, d, q)	Сезонна складова (P, D, Q)	AIC	BIC
c(0, 1, 0)	c(0, 0, 2)	10642,32	10662,78
c(0, 1, 1)	c(0, 0, 1)	10645,48	10665,94
c(0, 1, 2)	c(0, 0, 1)	10647	10672,58
c(0, 1, 3)	c(0, 0, 1)	10645,69	10676,38
c(0, 1, 1)	c(1, 0, 1)	10645,11	10670,68
c(0, 1, 2)	c(1, 0, 1)	10646,69	10677,38
c(0, 1, 3)	c(1, 0, 1)	10645,76	10681,57

За результатами отриманими з автоматичного підбору моделей можна заключити, що додаткове згладжування не дало очікуваних результатів, а повний та швидкий перебір параметрів моделей показав хоча і різні результати, але за інформаційними критеріями вони мають невелику різницю, тому будемо вважати, що обидві моделі підходять до наступного кроку. Третю модель залишаємо для більшої вибірки порівнянь кінцевих результатів та для надання системності для наборів даних, де згладжування є необхідним кроком.

Щодо ручного перебору, найкращий результат показала модель, параметри якої співпадають із отриманими у результаті автоматичного повного пошуку. Для

випадку, коли ручний підбір не співпадає з автоматичним, візьмемо ще ыншу модель, яка і за показниками виявилась другою. Після створення моделі важливо перевірити, наскільки добре вона працює з даними. Існує кілька діагностичних інструментів, доступних для перевірки поведінки моделі, а також показників точності, які дозволяють порівнювати одну модель з іншою [45]. Серед них є інформаційні критерії Акаіке та Байєса, показані вище; R^2 (коефіцієнт детермінації); критерій Дюрбіна-Ватсона. Також потрібно перевірити залишки на нормальність розподілення та провести тест портманто. Якщо в моделі було використано перетворення, тоді часто корисно подивитися на залишки на перетвореній шкалі. Це називається «інноваційними залишками». Залишки корисні для перевірки того, чи модель адекватно зафіксувала інформацію в даних. Для цього ми використовуємо інноваційні залишки. Таким чином, ми перевірятимемо чотири моделі ARIMA.

Так як інформаційні критерії Акаіке та Байєса вже наведені вище, одразу перевіряємо коефіцієнт детермінації та критерій Дюрбіна-Ватсона. Коефіцієнт детермінації для моделі з константою приймає значення від 0 до 1. Чим ближче значення коефіцієнта до 1, тим сильніша залежність. Оцінюючи регресійних моделей це інтерпретується як відповідність моделі даним. Для прийнятних моделей передбачається, що коефіцієнт детермінації повинен бути хоча б не менше 50% (у цьому випадку коефіцієнт множинної кореляції перевищує за модулем 70%). Моделі з коефіцієнтом детермінації вище 80% можна визнати досить добрими (коефіцієнт кореляції перевищує 90%). Значення коефіцієнта детермінації означає 1 функціональну залежність між змінними. Статистика Дюрбіна-Ватсона є тестовою статистикою для виявлення автокореляції в залишках регресійного аналізу. Він названий на честь професора Джеймса Дюрбіна, британського статистика та економетриста, та Джеффрі Стюарта Ватсона, австралійського статистика. Статистика Дюрбіна-Ватсона завжди прийматиме значення від 0 до 4. Значення DW = 2 вказує на відсутність автокореляції.

Для першої моделі автоматичного підбору значення R^2 склали: 0,998327, а DW: 2,017735. За показниками можна визначити, що модель досить добре себе показала за коефіцієнтом детермінації – майже повна залежність, а тест Дюрбіна-Ватсона показав відсутність автокореляції. Тепер подивимось на залишки (див. рис. 3.13).

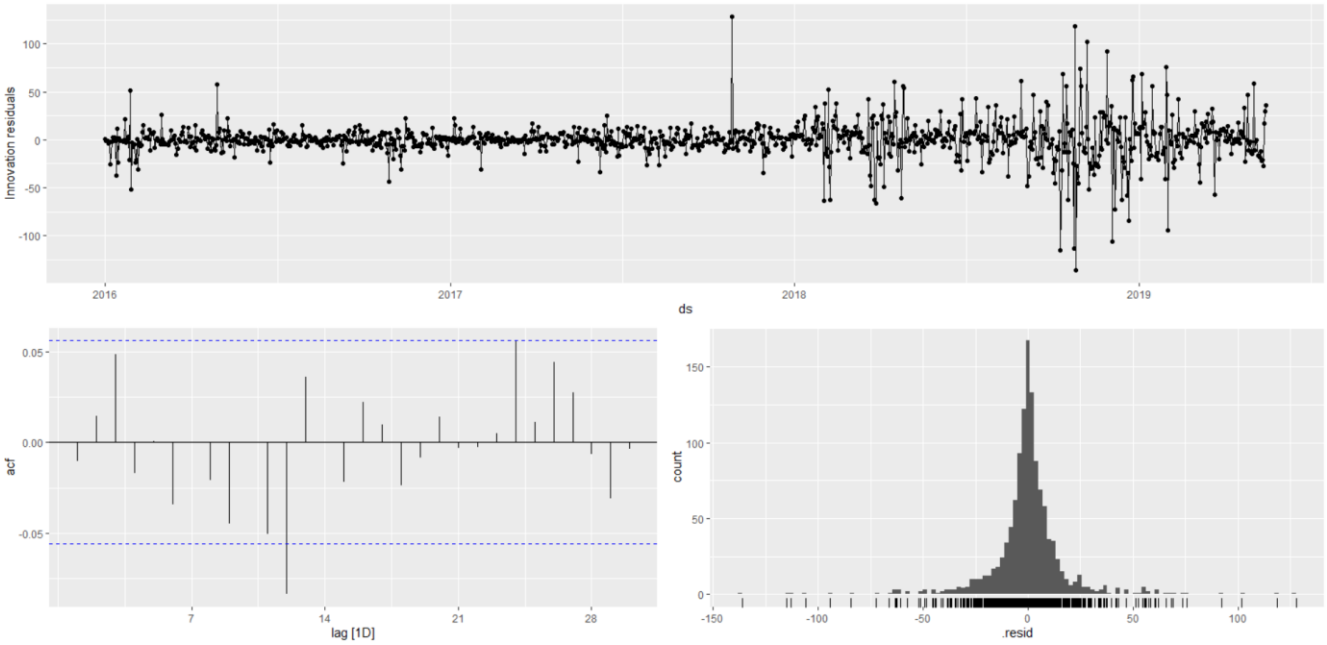

Рисунок 3.13 – Залишки автоматично сформованої моделі методом повного перебору (А – залишки, Б – ACF, В – розподіл залишків)

На залишках видно, що вони нормально розподілені і стаціонарні. Є викид на 12 лазі, але він одиночний і не дуже сильний, тобто все ще відповідає білому шуму. Результат тесту портманто: 0,0678, що означає – залишки для нашої моделі часових рядів є незалежними, що часто є припущенням, яке ми робимо під час створення моделі.

Для другої моделі автоматичного підбору, що була утворена методом швидкого перебору (також ця модель є найкращою з моделей, що були відібрані вручну), значення R^2 склали: 0,998327, а DW: 2,017926. За показниками можна визначити, що модель досить добре себе показала за коефіцієнтом детермінації –

майже повна залежність, а тест Дюрбіна-Ватсона показав відсутність автокореляції. Тепер подивимось на залишки (див. рис. 3.14).

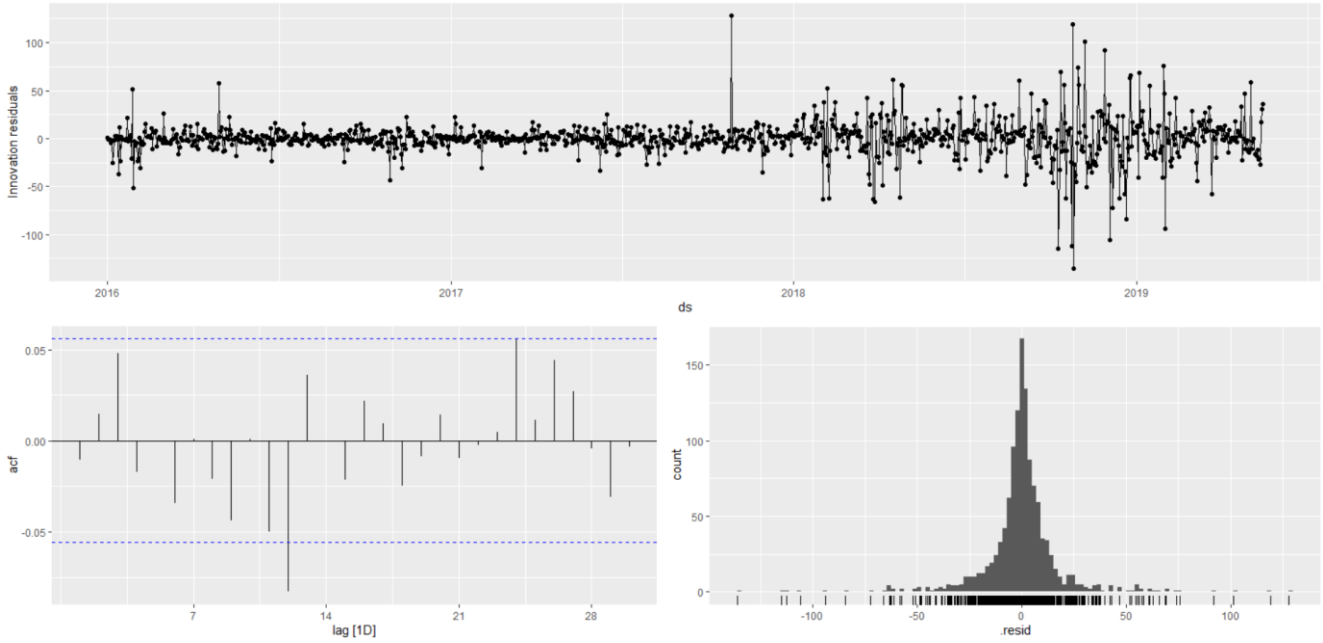

Рисунок 3.14 – Залишки автоматично сформованої моделі методом швидкого перебору (А – залишки, Б – ACF, В – розподіл залишків)

На залишках видно, що вони нормально розподілені і стаціонарні. Є викид на 12 лазі, але він одиночний і не дуже сильний, тобто все ще відповідає білому шуму. Результат тесту портманто: 0,0725, що означає – залишки для нашої моделі часових рядів є незалежними.

Для третьої моделі автоматичного підбору, що була утворена методом автоматичного перебору зі згладжуванням, значення R^2 склали: 0,9983226, а DW: 1,992346. За показниками можна визначити, що модель досить добре себе показала за коефіцієнтом детермінації – майже повна залежність, а тест Дюрбіна-Ватсона майже показав відсутність автокореляції. Тепер подивимось на залишки (див. рис. 3.15).

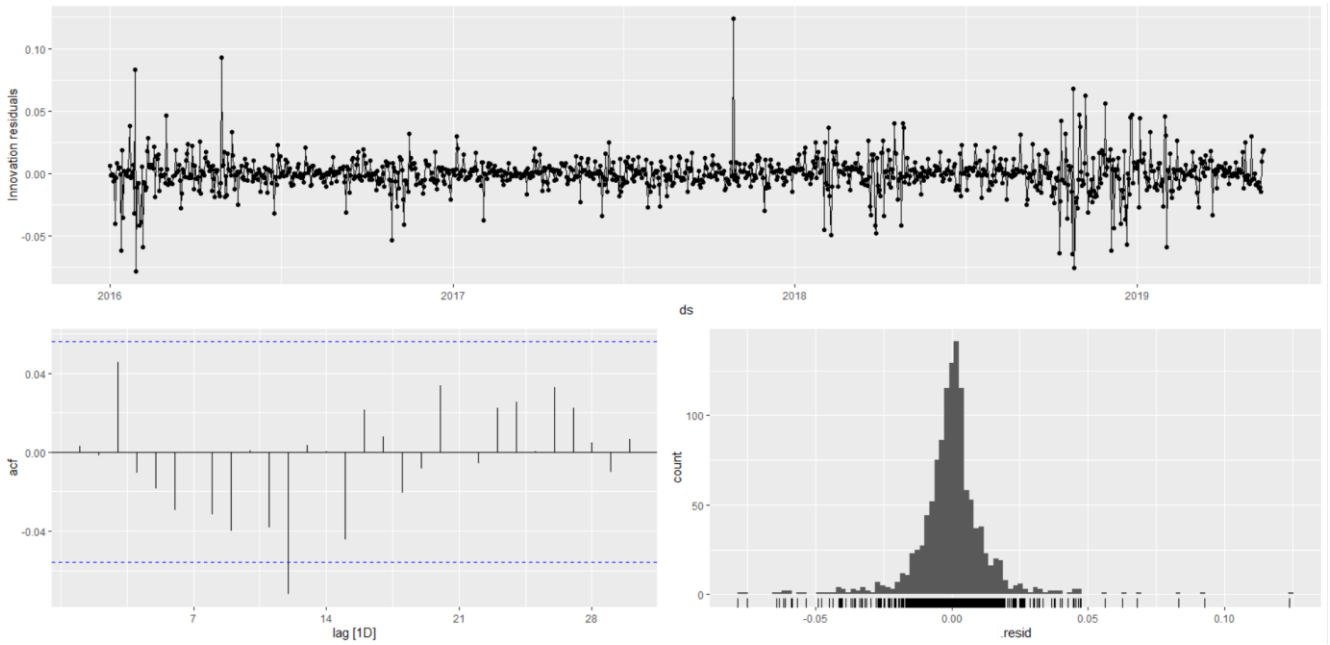

Рисунок 3.15 – Залишки автоматично сформованої моделі методом автоматичного перебору зі згладжуванням (А – залишки, Б – ACF, В – розподіл залишків)

На залишках видно, що вони нормально розподілені і стаціонарні. Є викид на 12 лазі, але він одиночний і не дуже сильний, тобто все ще відповідає білому шуму. Результат тесту портманто: 0,201, що означає – залишки для нашої моделі часових рядів є незалежними.

Для четвертої моделі, що була обрана за модифікованим алгоритмом ARIMA, значення R^2 склали: 0,9983226, а DW: 1,997119. За показниками можна визначити, що модель досить добре себе показала за коефіцієнтом детермінації – майже повна залежність, а тест Дюрбіна-Ватсона майже показав відсутність автокореляції. Тепер подивимось на залишки (див. рис. 3.16).

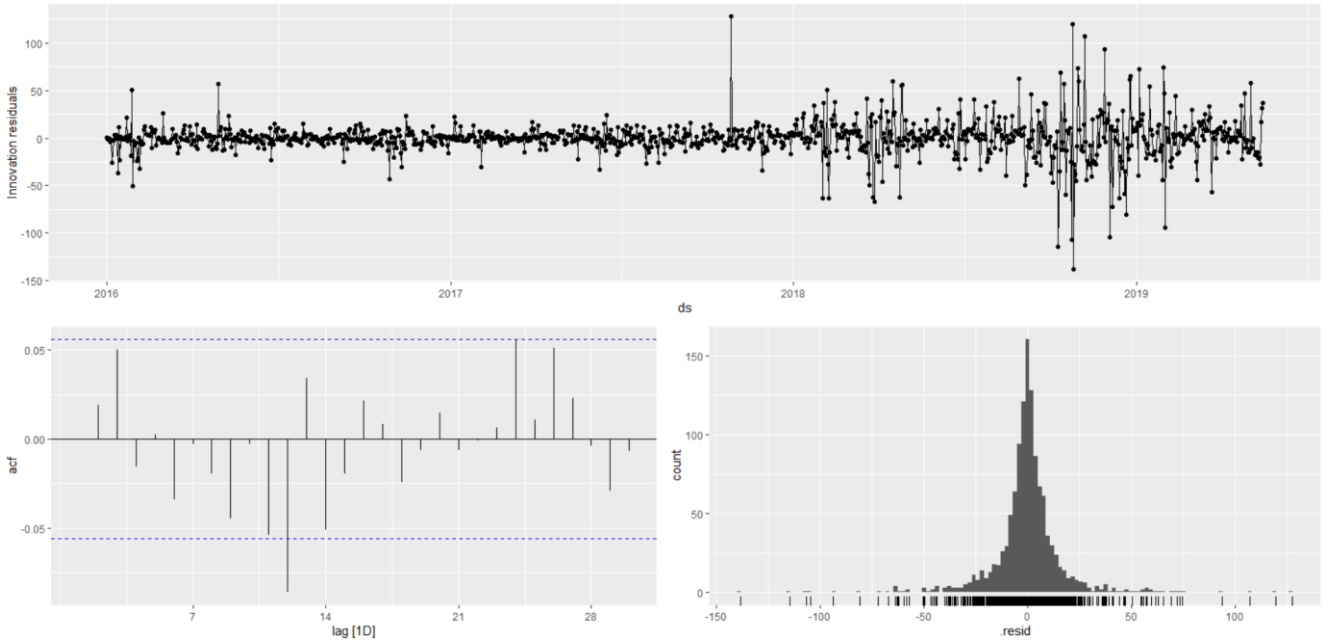

Рисунок 3.16 – Залишки моделі, що була відібрана другою за модифікованим алгоритмом ARIMA (А – залишки, Б – ACF, В – розподіл залишків)

На залишках видно, що вони нормально розподілені і стаціонарні. Є викид на 12 лазі, він повторюється на 24 лазі, тому ми не можемо сказати, що дані відповідають вимогам білого шуму. Результат тесту портманто: 0,0464, що означає – залишки для нашої моделі часових рядів є залежними.

Таблиця 3.3 – Порівняння обраних моделей ARIMA за вказаними метриками

Модель	AIC	BIC	R^2	DW	Portmanteau test
Автоматично згенерована	10642,26	10662,72	0,998327	2,017735	0,0678
Шивдко автоматично згенерована	10642,32	10662,78	0.998327	2,017926	0,0725
Автоматично згенерована зі згладжуванням	-7016,39	-6975,46	0,9983226	1,992346	0,201
Модифікований алгоритм ARIMA	10642,32	10662,78	0,9983226	1,997119	0,0464

Результати порівняння показали, що автоматично згенерована модель за показниками AIC та BIC є найкращою, а от за іншими показниками, особливо за тестом портманто швидко автоматично згенерована ARIMA показала кращі результати, тому очікується, що результати прогнозування для неї теж будуть кращими. Також цікавим є той факт, що друга за показниками AIC та BIC модель модифікованого алгоритму ARIMA показала не набагато гірший результат, ніж перша за цими показниками модель.

3.3 Моделювання GAM

Як було зазначено вище, очікується, що сезонна компонента обраного часового ряду складається з двох частин: тижневої та річної. Але задля системного підходу, GAM використовуватиме і місячну компоненту, тому моделі будуть представлені таким набором: місячна адитивного типу; річна мультиплікативного; річна мультиплікативного і тижнева адитивного. Для цього набору моделей також проведемо порівняння за метриками, визначимо кращу і перейдемо до прогнозування з використанням вже всіх моделей.

Для першої моделі побудуємо декомпозицію засобами Prophet:

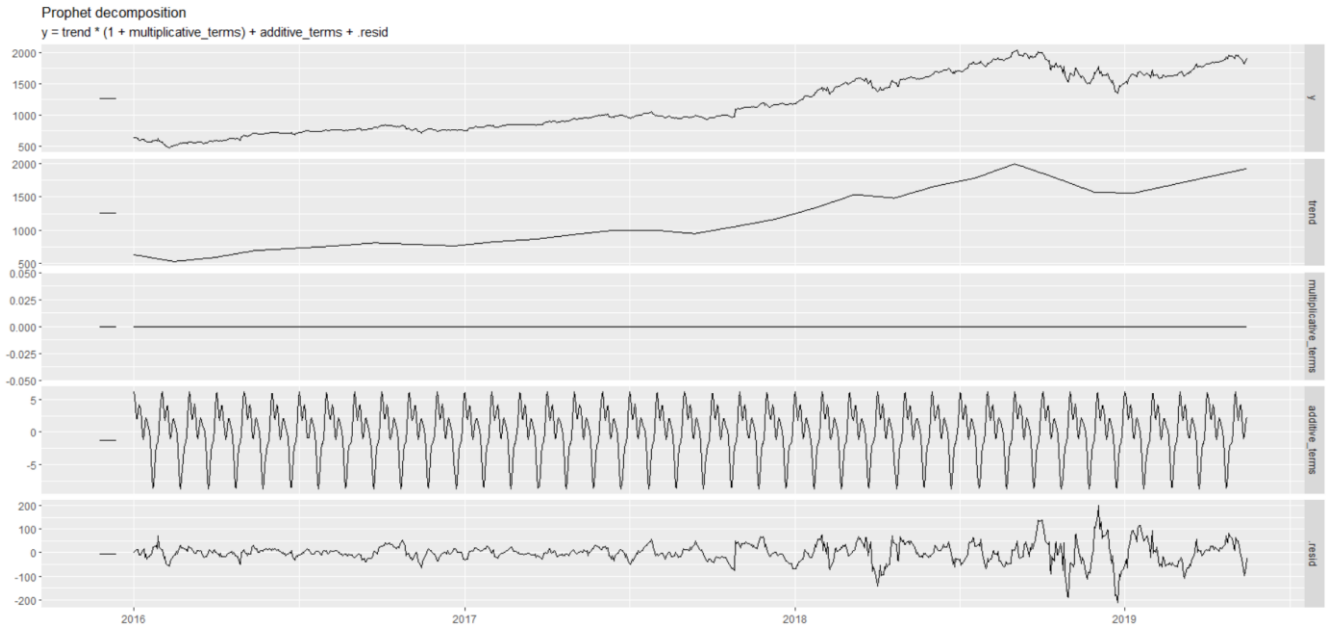

Рисунок 3.17 – Декомпозиція першої GAM з місячною сезонною компонентою адитивного типу засобами Prophet

Як видно з рисунку, засоби Prophet показують лише одну сезонну компоненту для даної моделі – місячну, яка, як видно з тренду, погано вписується в загальну картину відомостей про набір даних. Сам тренд не достатньо згладжений і лише місцями правильно показує коливання даних. Тепер поглянемо детальніше на залишки:

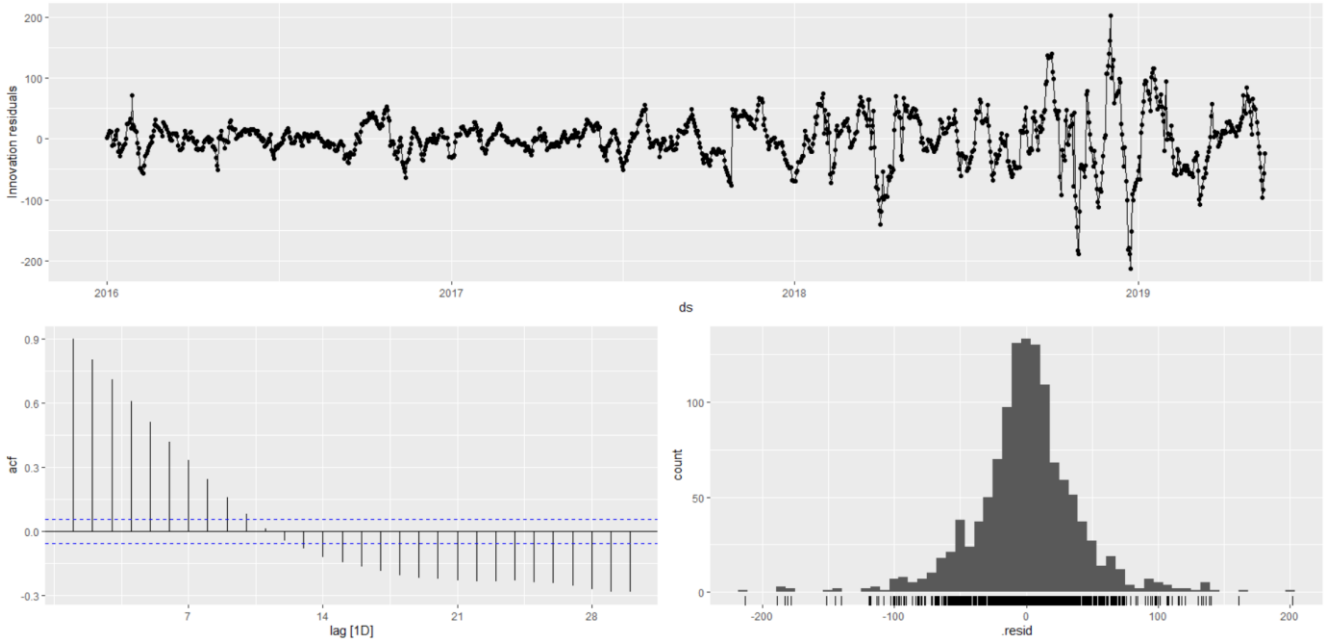

Рисунок 3.18 – Залишки першої GAM (А – залишки, Б – ACF, В – розподіл залишків)

Залишки мають нормальний розподіл. Коливання ACF за межами норми, мають спадний характер і переходять у від'ємну частину після 12 лагу. Загалом це вказує на наявність автокореляції між даними. Значення коефіцієнту детермінації: 0,9918001, а Дюрбіна-Ватсона: 0,2011187. За показниками можна визначити, що модель досить добре себе показала за коефіцієнтом детермінації – майже повна залежність, а тест Дюрбіна-Ватсона показав наявність автокореляції. Такий результат отримується тому, що модель будується зважаючи на місячну сезонну компоненту, яка не представлена явно у наборі, тому модель недонавчається на дійсно важливих показниках.

Побудуємо для другої моделі декомпозицію засобами Prophet:

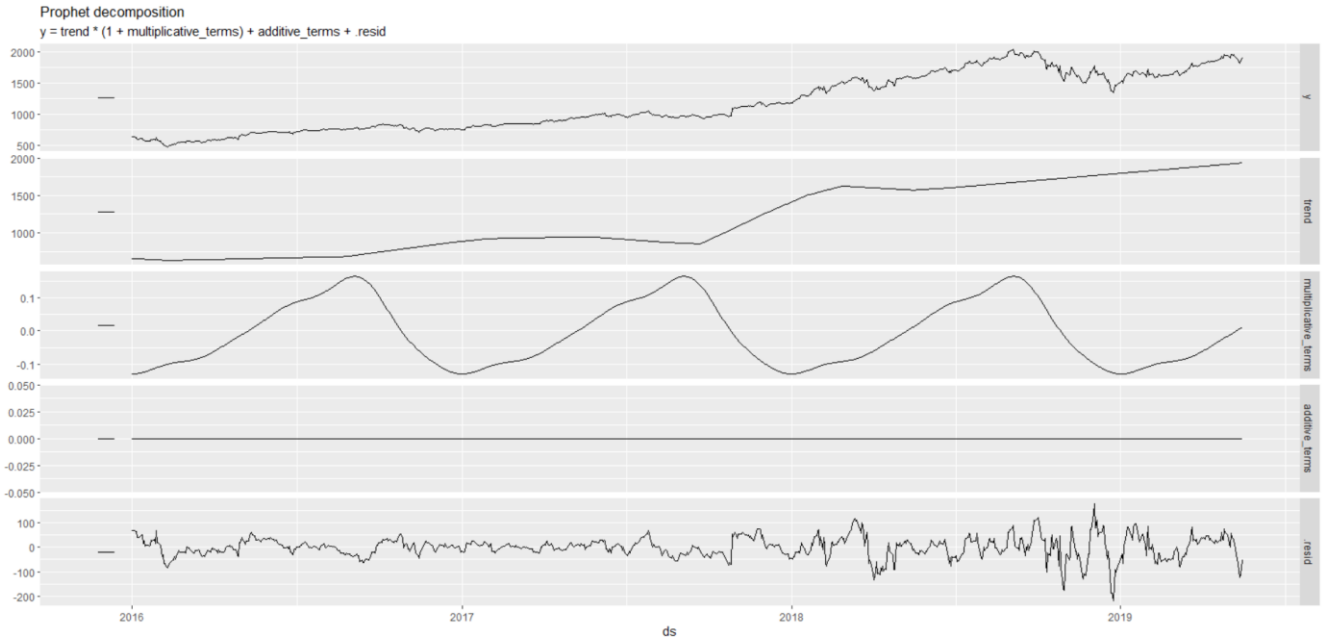

Рисунок 3.19 – Декомпозиція другої GAM з річною сезонною компонентою мультиплікативного типу засобами Prophet

Як видно з рисунку, засоби Prophet показують лише одну річну сезонну компоненту мультиплікативного типу, яка набагато краще відображає набір даних. Тренд все ще неповністю відповідає основним коливанням даних і місцями має досить різкі зміни, але результат значно краще ніж у попередньої моделі . Тепер поглянемо детальніше на залишки:

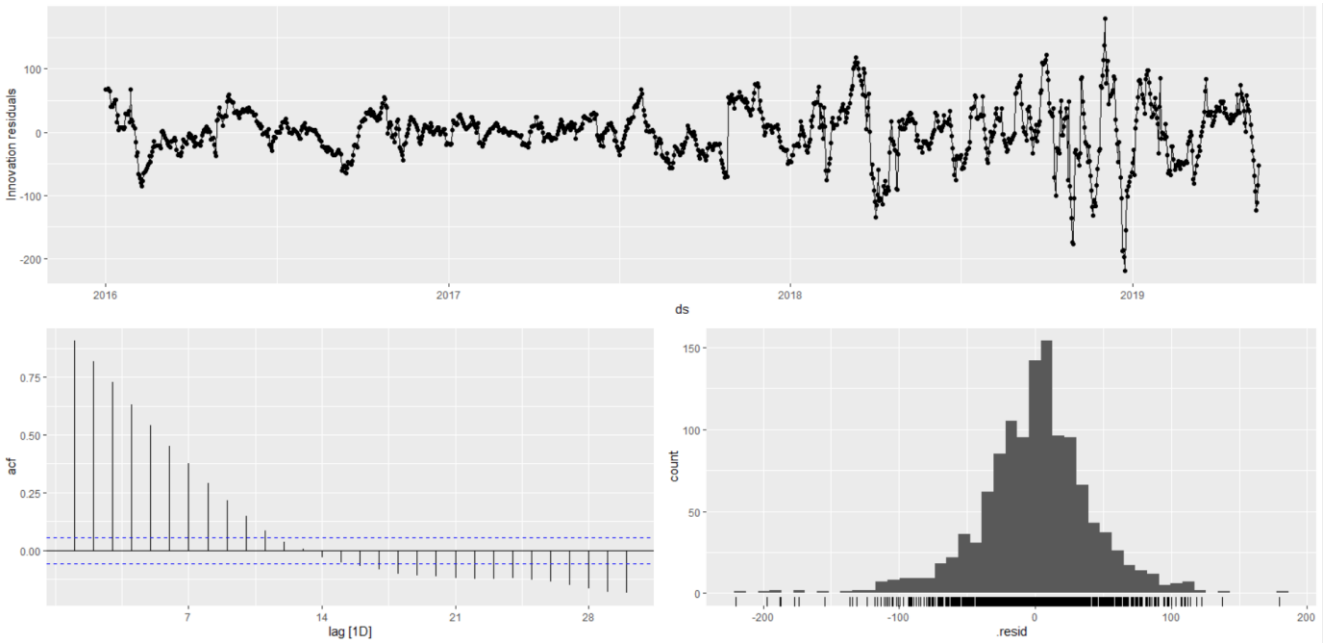

Рисунок 3.20 – Залишки другої GAM (А – залишки, Б – ACF, В – розподіл залишків)

Залишки мають нормальний розподіл, але з певними відхиленнями у від’ємній частині. Коливання ACF за межами норми, мають спадний характер і переходять у від'ємну частину після 12 лагу. Загалом це вказує на наявність автокореляції між даними. Значення коефіцієнту детермінації: 0,9909922, а Дюрбіна-Ватсона: 0,1840706. За показниками можна визначити, що модель досить добре себе показала за коефіцієнтом детермінації – майже повна залежність, а тест Дюрбіна-Ватсона показав наявність автокореляції. Такий результат отримується тому, що модель будується зважаючи на річну сезонну компоненту, яка мало представлена у наборі даних, оскільки він відображає події тільки трьох повних років.

Побудуємо на останок декомпозицію для третьої моделі засобами Prophet:

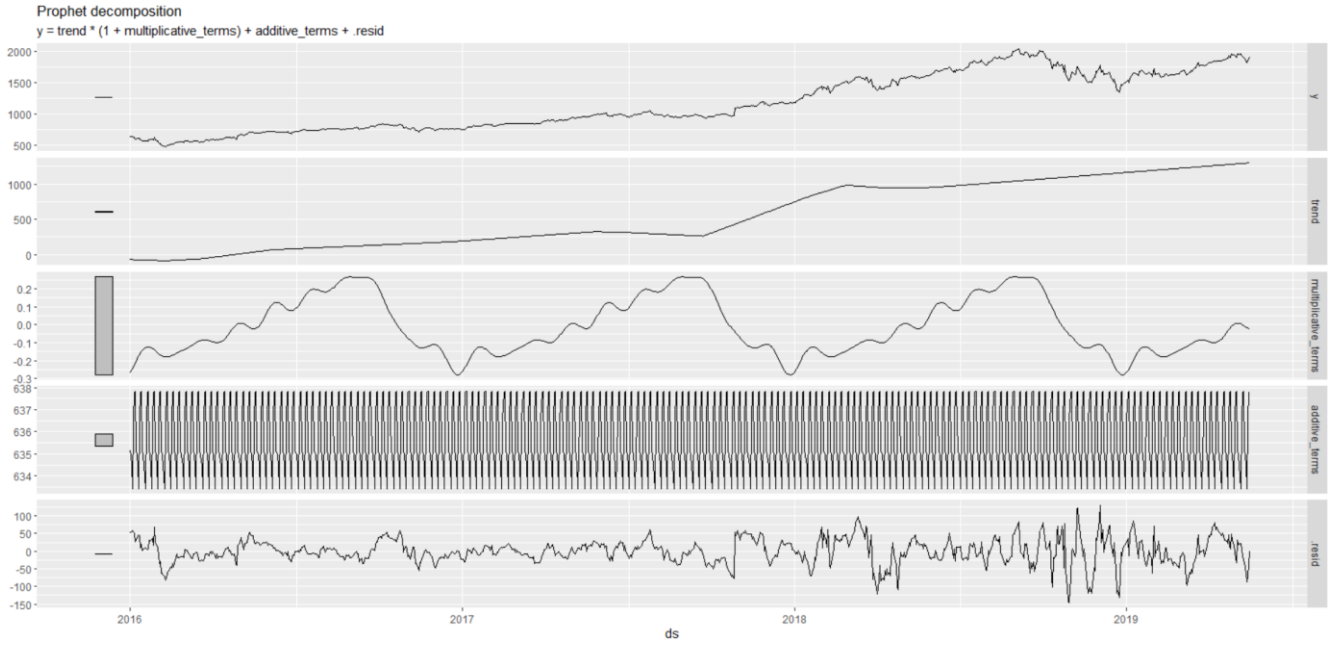

Рисунок 3.21 – Декомпозиція третьої GAM з річною сезонною компонентою мультиплікативного типу і тижневою адитивного засобами Prophet

Як видно з рисунку, засоби Prophet показують обидві річну і тижневу сезонні компоненти. Тренд все ще неповністю відповідає основним коливанням даних і місцями має досить різкі зміни, але результат значно краще ніж у попередньої моделі . Тепер поглянемо детальніше на залишки:

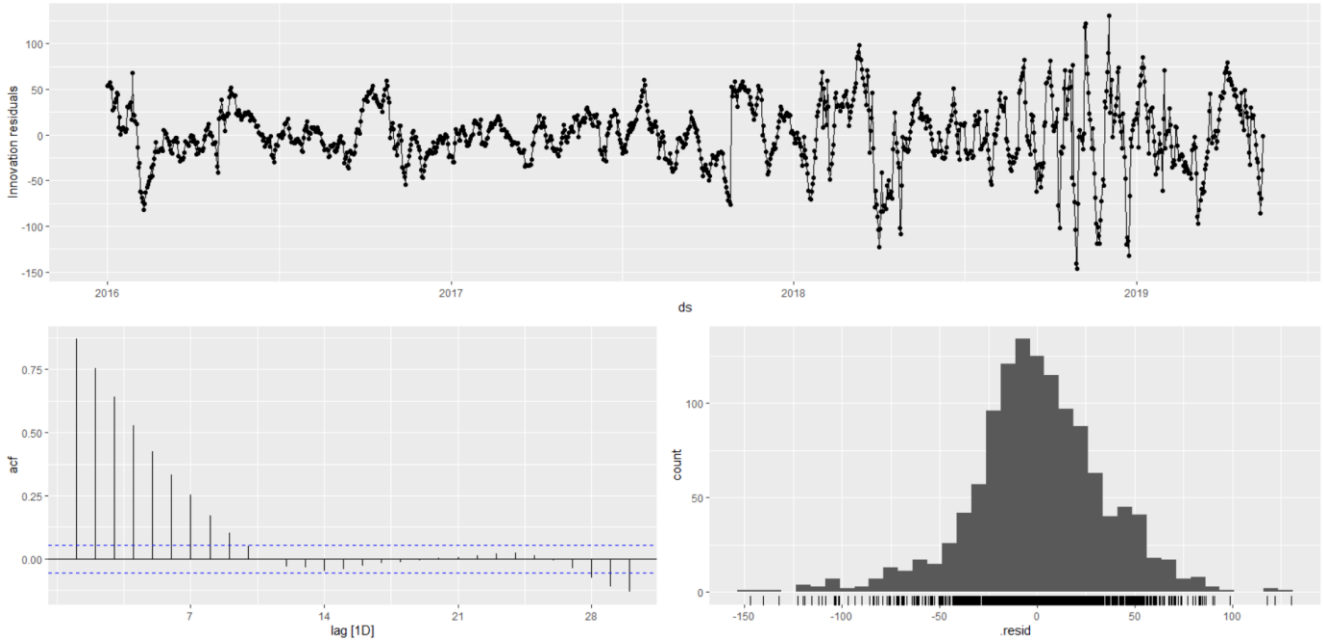

Рисунок 3.22 – Залишки третьої GAM (А – залишки, Б – ACF, В – розподіл залишків)

Залишки мають нормальний розподіл, але з певними відхиленнями у додатній частині. Коливання ACF за межами норми, мають спадний характер і переходять у норму частину після 10 лагу. Загалом це вказує на наявність автокореляції між даними, але набагато слабшу ніж у попередніх моделях. Значення коефіцієнту детермінації: 0,993826, а Дюрбіна-Ватсона: 0,2595217. За показниками можна визначити, що модель досить добре себе показала за коефіцієнтом детермінації – майже повна залежність, а тест Дюрбіна-Ватсона показав наявність автокореляції. Такий результат отримується тому, що модель будується зважаючи на річну сезонну компоненту, яка мало представлена у наборі даних, оскільки він відображає події тільки трьох повних років. В результаті 3 модель показала себе краще інших, тому від неї очікується і кращий прогноз.

Таблиця 3.4 – Порівняння обраних GAM за вказаними метриками

Модель	R^2	DW
Місячна сезонна компонента	0,9918001	0,2011187
Річна сезонна компонента	0,9909922	0,1840706
Річна і тижнева сезонні компоненти	0,993826	0,2595217

3.4 Прогнозування

Для пронозування будуть використані усі моделі описані вище, як ARIMA, так і GAM. Коли відповідну модель визначено, оцінено та перевірено, настав час створювати прогнози за допомогою forecast(). Найпростіший спосіб використання цієї функції – вказати кількість майбутніх спостережень для прогнозування. Наприклад, прогнози для наступних 10 спостережень можуть бути створені за допомогою h = 10. Ми також можемо використовувати природну мову; наприклад, h = "2 years" можна використовувати для прогнозування на два роки в майбутньому.

В інших ситуаціях може бути зручніше надати набір даних майбутніх періодів часу для прогнозування. Це зазвичай потрібно, коли ваша модель використовує додаткову інформацію з даних, наприклад екзогенні регресори. Додаткові дані, необхідні для моделі, можна включити в набір даних спостережень для прогнозування.

Для подальшого порівняння дамо спрощені назви всім моделям: автоматично згенерована ARIMA – ARIMA 1; шивдко автоматично згенерована ARIMA – ARIMA 2; автоматично згенерована зі згладжуванням – ARIMA 3; модифікований алгоритм ARIMA (друга за якістю модель) – ARIMA 4; місячна сезонна компонента – GAM 1; річна сезонна компонента – GAM 2; річна і тижнева сезонні компоненти – GAM 3.

Для оцінки якості прогнозування використовуватимуться різні метрики, опис яких додається нижче:

Таблиця 3.5 – Метрики, що використовуються для оцінки якості прогнозування та їх властивості

Метрика	Опис
ME: Mean Error	Середня помилка — це неофіційний термін, який зазвичай стосується середнього значення всіх помилок у наборі. «Помилка» в цьому контексті — це невизначеність у вимірюванні, або різниця між виміряним значенням і істинним/правильним значенням
RMSE: Root Mean Squared Error	RMSE завжди буде більше або дорівнює MAE; чим більша різниця між ними, тим більша дисперсія в окремих помилках у вибірці. Якщо RMSE=MAE, то всі помилки мають однакову величину
MAE: Mean Absolute Error	MAE вимірює середню величину помилок у наборі прогнозів, не враховуючи їхнього напрямку. Він вимірює точність безперервних змінних. І MAE, і RMSE можуть коливатися від 0 до ∞. Це негативно орієнтовані оцінки: нижчі значення кращі.
MPE: Mean Percentage Error	Це обчислене середнє відсоткове значення помилок, на які прогнози моделі відрізняються від фактичних значень прогнозованої кількості.
MAPE: Mean Absolute Percentage Error	MAPE у відсотках має сенс лише для значень, де мають сенс ділення та співвідношення. Немає сенсу розраховувати відсоток температур, які можуть виникнути MAPE вище 100%. Тоді це може призвести до негативної точності, яку людям може бути важко зрозуміти. Помилка, близька до 0%

Метрика	Опис
	=> підвищення точності прогнозу. Приблизно 2,2% MAPE означає, що модель є приблизно 97,8% точною для прогнозування наступних спостережень.
MASE: Mean Absolute Scaled Error	Інваріантність масштабу: не залежить від масштабу даних, тому її можна використовувати для порівняння прогнозів у наборах даних із різними масштабами. Добре для шкал, які не мають значущого 0, однаково штрафує позитивні та негативні помилки прогнозу. Значення більше одиниці вказують на те, що однокрокові прогнози у вибірці за наївним методом працюють краще, ніж прогнозні значення, які розглядаються. При порівнянні методів прогнозування перевагу надається методу з найнижчим MASE.
RMSSE: Root Mean Squared Scaled Error	Це квадратний корінь варіанту метрики втрат MASE. Ця метрика також підходить для періодичних серій попиту, оскільки вона не надаватиме нескінченних або невизначених значень, якщо навчальні дані не є плоскими часовими рядами. У цьому випадку функція повертає велике значення замість inf.
ACF1: Autocorrelation of errors at lag 1	Це міра того, наскільки на поточне значення впливають попередні значення в часовому ряду. Як правило, можна очікувати, що функція автокореляції впаде до 0, коли точки стають більш відокремленими

Почнемо з візуалізації прогнозу для ARIMA 1, після чого оцінемо якість прогнозу за різними метриками:

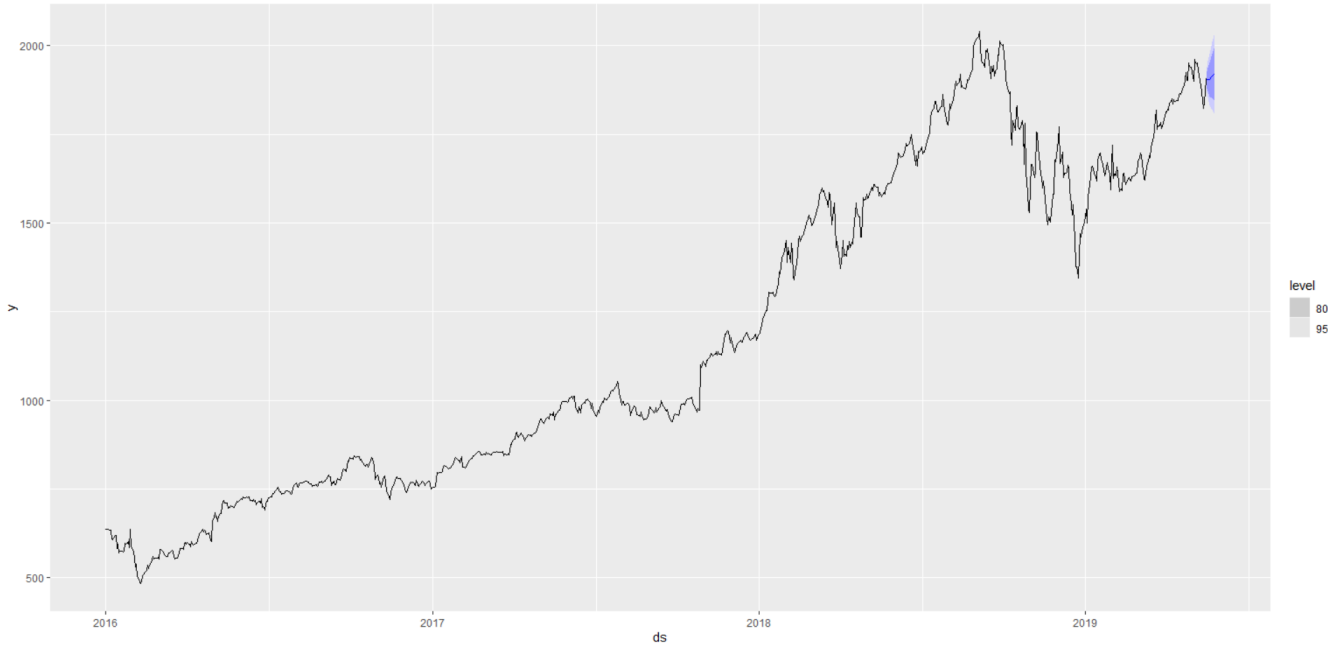

Рисунок 3.23 – Прогнози з використанням моделі ARIMA(0, 1, 0)$(2, 0, 0)_7$. Показано 80% і 95% інтервалів передбачення

Таблиця 3.6 – Оцінка точності моделі ARIMA 1 за різними метриками

ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
-60,4	67,9	60,4	-3,29	3,29	1,79	1,37	0,681

Так як прогнозування відбувається для часових рядів, то визначатимемо точність прогнозу за MAPE, адже ця метрика найпростіша для розуміння. Так, для ARIMA 1 похибка прогнозування становить 3,29%. Інші метрики використовуватимемо для порівняння усіх прогнозів.

Переходимо до візуалізації та оцінки прогнозування ARIMA 2:

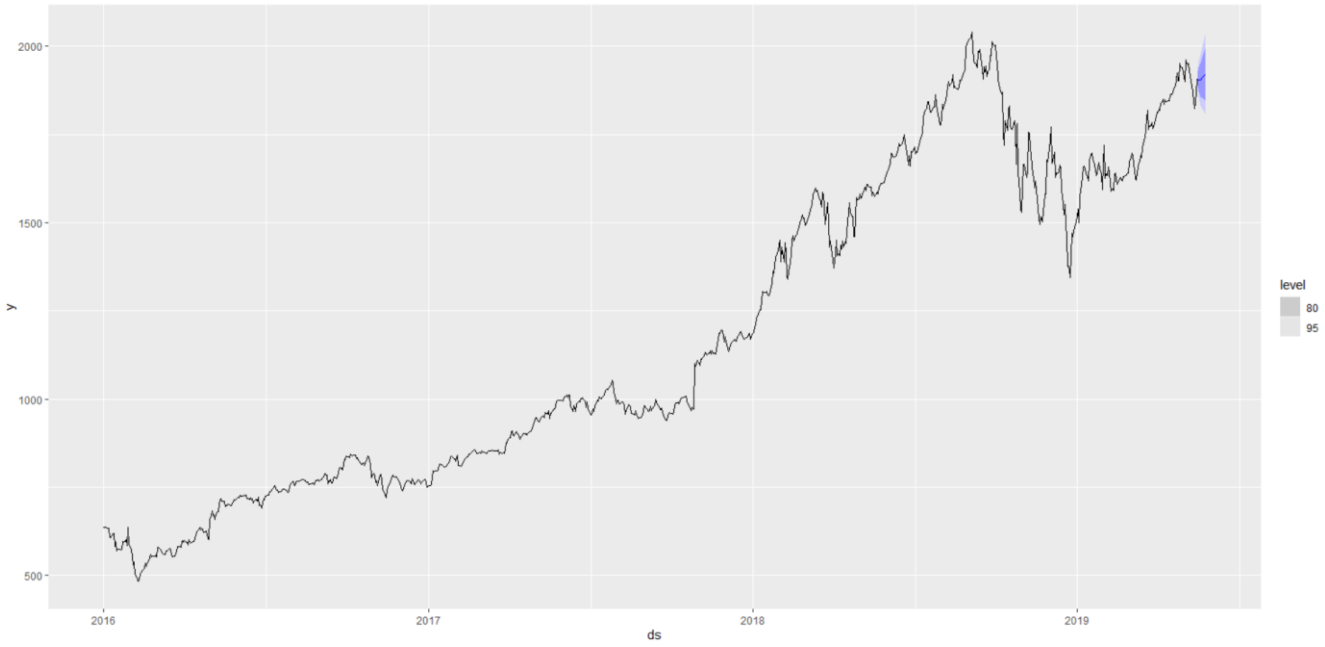

Рисунок 3.24 – Прогнози з використанням моделі ARIMA(0, 1, 0)$(0, 0, 2)_7$. Показано 80% і 95% інтервалів передбачення

Таблиця 3.7 – Оцінка точності моделі ARIMA 2 за різними метриками

ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
-60,5	67,9	60,5	-3,29	3,29	1,79	1,37	0,681

Для ARIMA 2 похибка прогнозування становить 3,29%. Результати за іншими метриками дуже схожі на ARIMA 1.

Візуалізація та оцінки прогнозування ARIMA 3:

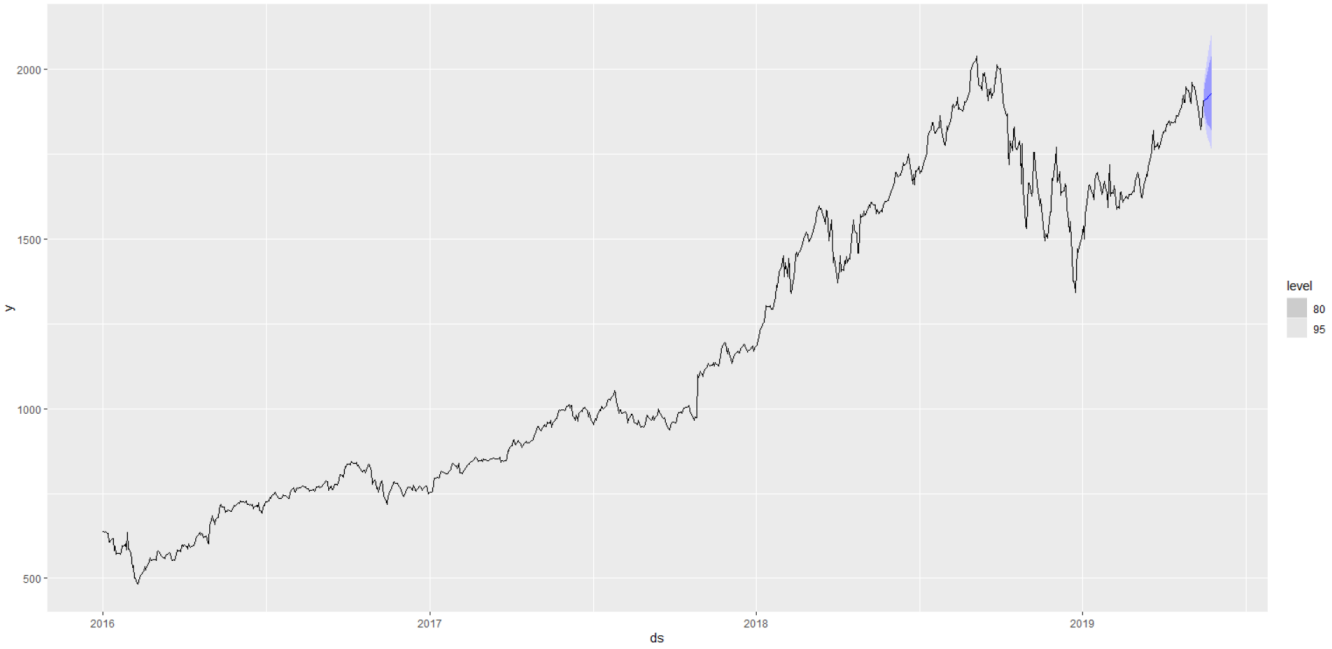

Рисунок 3.25 – Прогнози з використанням моделі ARIMA(0, 1, 2)(2, 0, 2)$_7$. Показано 80% і 95% інтервалів передбачення

Таблиця 3.8 – Оцінка точності моделі ARIMA 3 за різними метриками

ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
-68,7	76,0	-68,7	-3,74	3,74	2,04	1,54	0,677

Для ARIMA 3 похибка прогнозування становить 3,74%. Хоча похибка виросла не сильно, ми бачимо сильно збільшені довірчі межі на рисунку.

Візуалізація та оцінки прогнозування ARIMA 4:

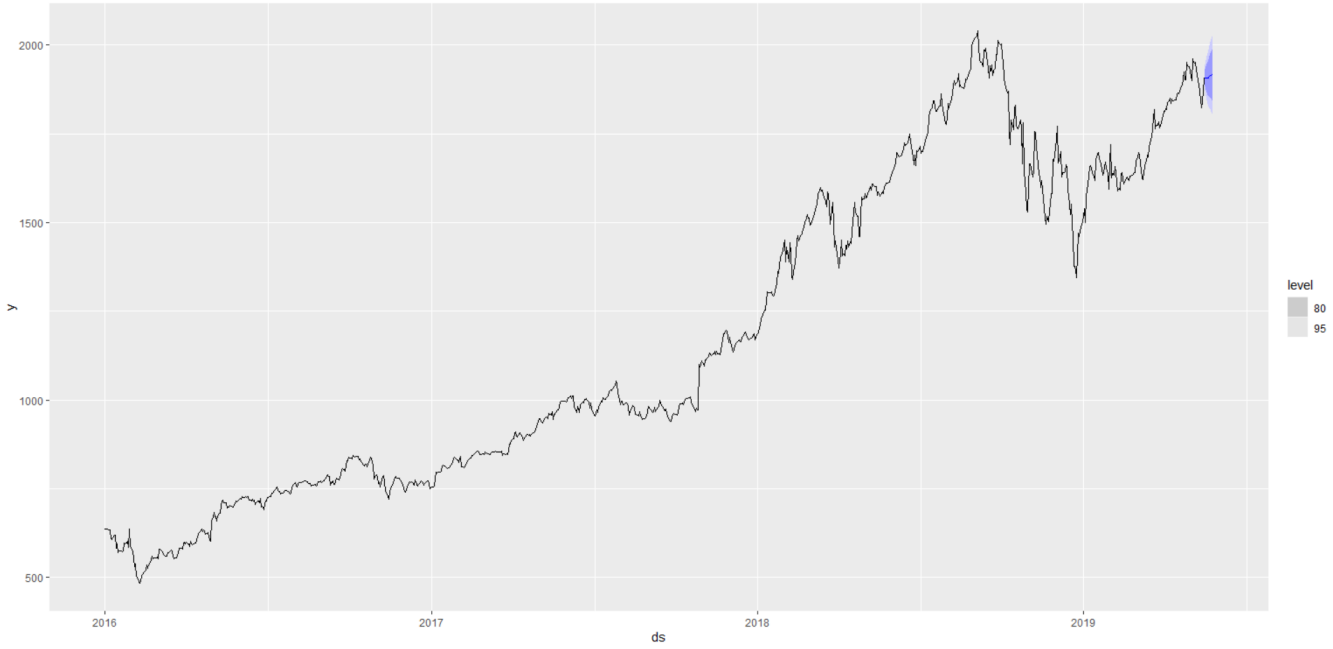

Рисунок 3.26 – Прогнози з використанням моделі ARIMA(0, 1, 1)(0, 0, 1)$_7$. Показано 80% і 95% інтервалів передбачення

Таблиця 3.9 – Оцінка точності моделі ARIMA 4 за різними метриками

ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
-60,5	67,2	60,5	-3,29	3,29	1,80	1,36	0,675

Для ARIMA 4 похибка прогнозування становить 3,29%.

Переходимо до GAM. Візуалізація та оцінки прогнозування GAM 1:

Рисунок 3.27 – Прогнози з використанням моделі GAM 1. Показано 80% і 95% інтервалів передбачення

Таблиця 3.10 – Оцінка точності моделі GAM 1 за різними метриками

ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
-91,6	96,4	91,6	-4,97	4,97	2,72	1,95	0,657

Для GAM 1 похибка прогнозування становить 4,97%.

Візуалізація та оцінки прогнозування GAM 2:

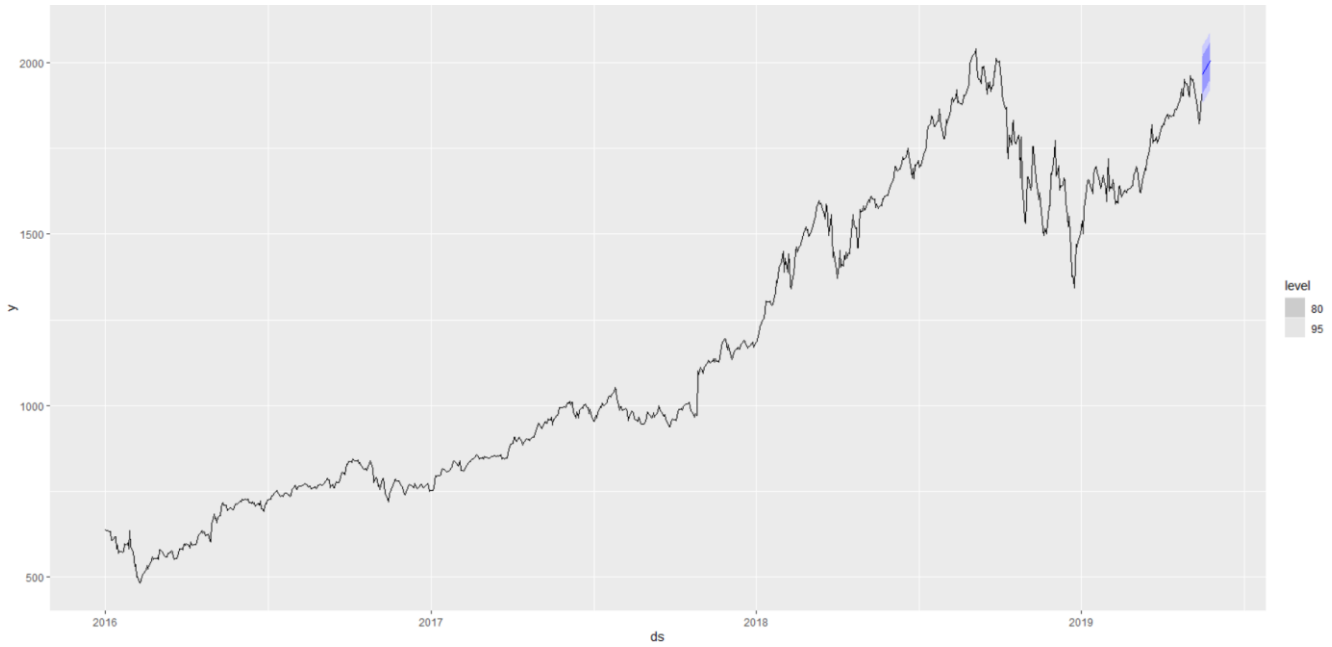

Рисунок 3.28 – Прогнози з використанням моделі GAM 2. Показано 80% і 95% інтервалів передбачення

Таблиця 3.11 – Оцінка точності моделі GAM 2 за різними метриками

ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
-134	139	134	-7,28	7,28	3,98	2,82	0,682

Для GAM 2 похибка прогнозування становить 7,28%.

Візуалізація та оцінки прогнозування GAM 3:

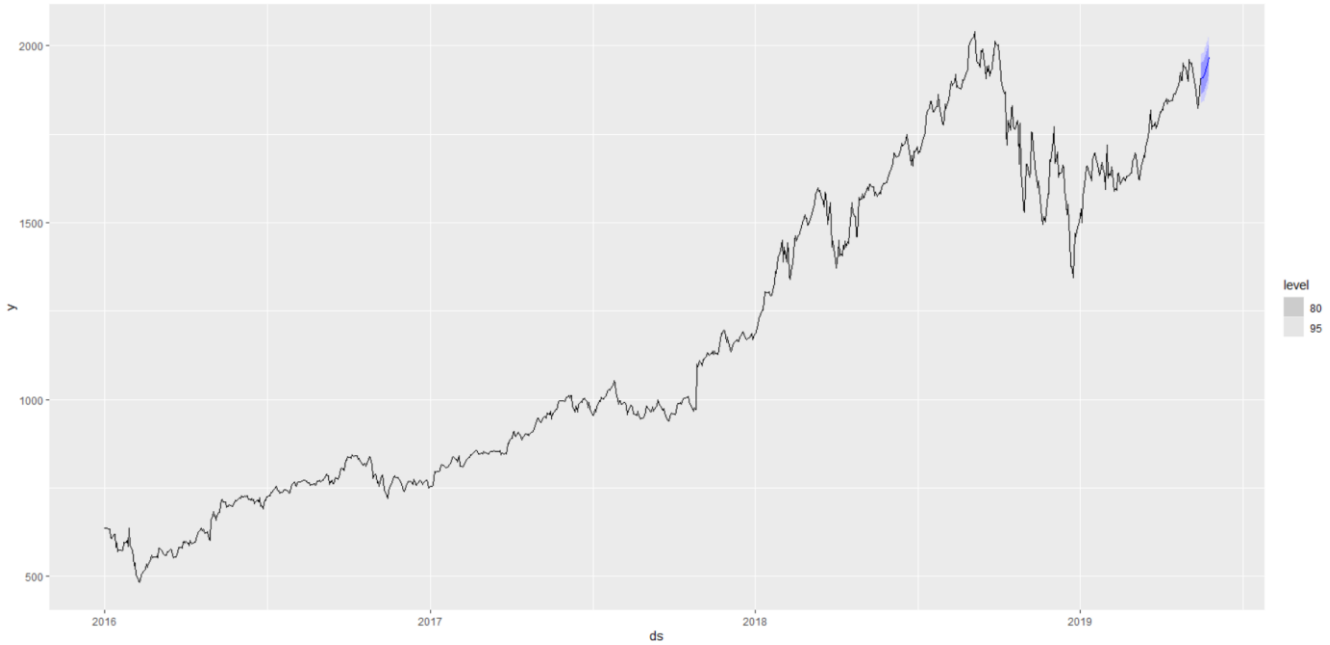

Рисунок 3.29 – Прогнози з використанням моделі GAM 3. Показано 80% і 95% інтервалів передбачення

Таблиця 3.12 – Оцінка точності моделі GAM 3 за різними метриками

ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
-79,4	90,7	79,4	-4,32	4,32	2,36	1,83	0,706

Для GAM 3 похибка прогнозування становить 4,32%.

Таблиця 3.13 – Порівняльна таблиця результатів прогнозування за різними метриками для всіх моделей

Модель	ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
ARIMA 1	-60,4	67,9	60,4	-3,29	3,29	1,79	1,37	0,681
ARIMA 2	-60,5	67,9	60,5	-3,29	3,29	1,79	1,37	0,681
ARIMA 3	-68,7	76,0	-68,7	-3,74	3,74	2,04	1,54	0,677
ARIMA 4	-60,5	67,2	60,5	-3,29	3,29	1,80	1,36	0,675
GAM 1	-91,6	96,4	91,6	-4,97	4,97	2,72	1,95	0,657
GAM 2	-134	139	134	-7,28	7,28	3,98	2,82	0,682

Модель	ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
GAM 3	-79,4	90,7	79,4	-4,32	4,32	2,36	1,83	0,706

Досить часто корисним буде оцінити також наскільки широкими виходять довірчі межі і для цього можна застосувати оцінку Вінклера. Для спостережень, які потрапляють в інтервал, оцінка Вінклера є просто довжиною інтервалу. Отже, низькі бали пов’язані з вузькими інтервалами. Проте, якщо спостереження виходить за межі інтервалу, застосовується штраф, пропорційний тому, наскільки далеко спостереження знаходиться за межами інтервалу. Для даного набору даних оцінку проведемо на 80% довірчому інтервалі. Часто нас цікавить весь розподіл прогнозу, а не окремі квантилі чи інтервали прогнозу. У цьому випадку ми можемо усереднити квантильні оцінки за всіма значеннями p, щоб отримати безперервну рейтингову оцінку ймовірності або CRPS.

Таблиця 3.14 – Порівняльна таблиця результатів оцінки Вінклера та CRPS для всіх моделей

Модель	Оцінка Вінклера	CRPS
ARIMA 1	215	40,6
ARIMA 2	215	40,6
ARIMA 3	178	42,4
ARIMA 4	209	40,6
GAM 1	512	70,1
GAM 2	904	111
GAM 3	486	62,8

За результатами порівняння прогнозів за різними метриками модель ARIMA 1 виявилась найкращою. З отриманих даних видно, що при порівнянні оцінок точності для інтервалів довіри модель ARIMA 4 виявилась найкращою.

Спробуємо перевірити, чи допоможе комбінація двох моделей, що найкраще описують вхідні дані у покращенні прогнозу, а саме: ARIMA 2 та GAM 3. Хоча ці моделі не дають найкращі результати, вони мають певні переваги у використанні більш точних показників тренду та сезону, а ARIMA 2 взагалі має деякі однакові результати за метриками MPE, MAPE, MASE з моделлю ARIMA 1.

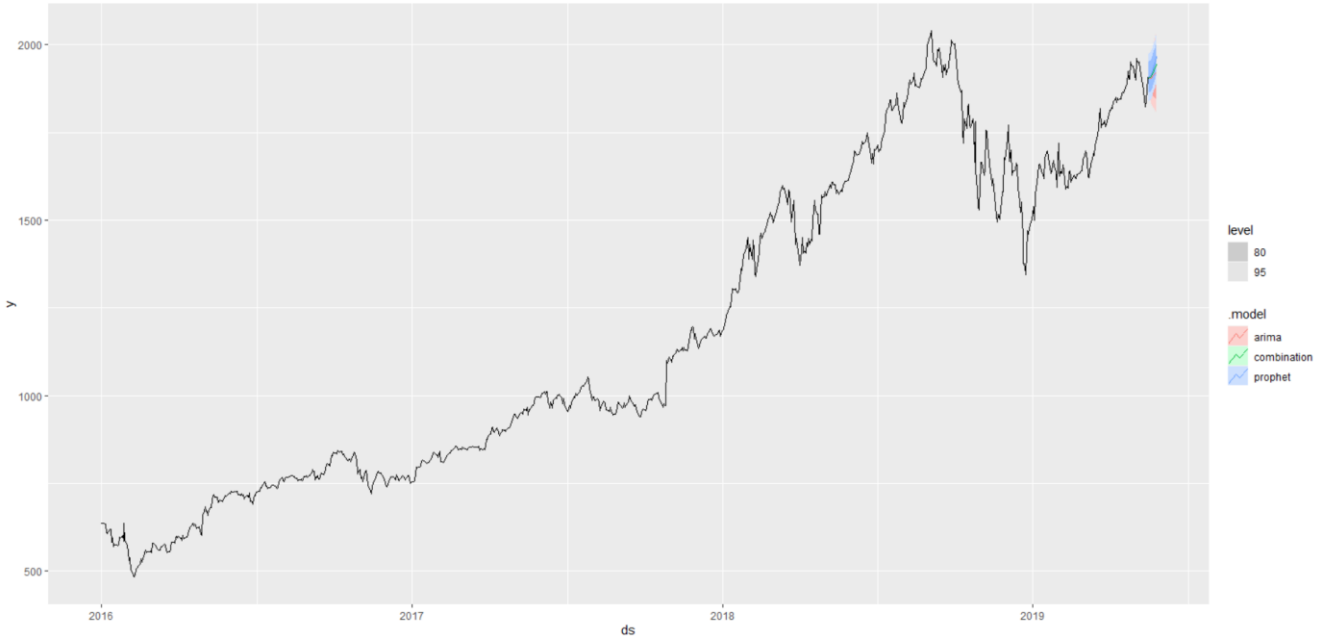

Рисунок 3.30 – Прогнози з використанням комбінованої моделі. Показано 80% і 95% інтервалів передбачення для кожної складової та їх комбінації

На рисунку погано видно різницю між прогнозами через його короткостроковість та невеликий об’єм у порівнянні з вхідним набором даних, тому поглянемо ближче:

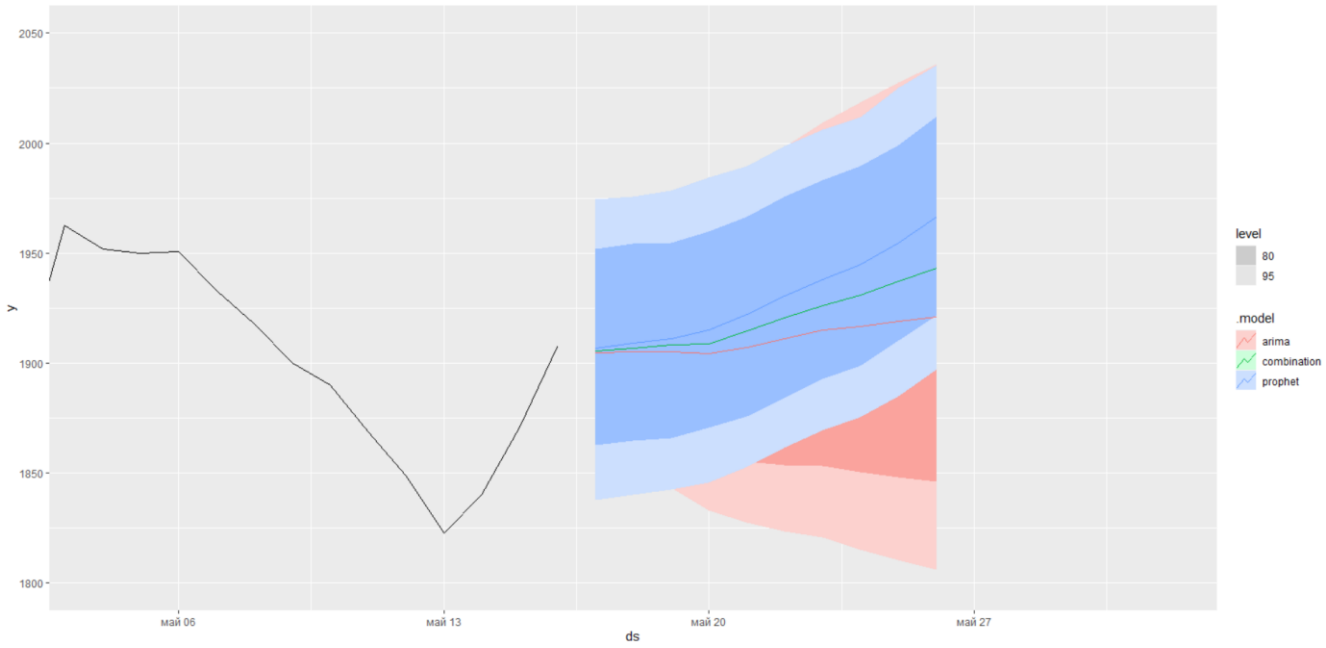

Рисунок 3.31 – Прогнози з використанням комбінованої моделі. Показано 80% і 95% інтервалів передбачення для кожної складової та їх комбінації. Тільки прогнозна частина

Як бачимо з рисунку, модель ARIMA має найбільший інтервал довіри, а от для комбінації взагалі не відображається. Результат прогнозу вираховувався по середньому, що чітко видно на рисунку. Перейдемо до числових показників: метрик та оцінок.

Таблиця 3.15 – Порівняльна таблиця результатів прогнозування за різними метриками для комбінованої моделі

Модель	ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1	Winkler	CRPS
ARIMA	-60,5	67,9	60,5	-3,29	3,29	1,79	1,37	0,681	215	40,6
Combination	-69,7	79,0	69,7	-3,80	3,80	2,07	1,60	0,699	697	69,7
GAM	-79,4	90,7	79,4	-4,32	4,32	2,36	1,83	0,706	480	62,7

За результатами порівняння отримуємо висновок, що використання комбінованого підходу не дало покращення загального результату – модель

ARIMA 1 так і залишилася найкращою, але комбінована модель показала себе краще, ніж GAM 3, водночас збільшивши показники статистики Вінклера та CRPS, що є поганою ознакою. Обрані моделі за емпіричними ознаками погано впоралися з комбінацією. Використання комбінованого підходу не дало очікуваних покращень тому, що модель GAM 3 дає дуже оптимістичні результати прогнозування, в той час як прогноз за ARIMA 1 є близьким до реальних даних а підхід по середньому не дає можливість зробити прогноз ще більш помірним. Можна сказати, що для набору даних, де одна модель дає занадто оптимістичний, а інша занадто песимістичний прогнози, підхід по середньому покаже себе набагато краще.

Висновки до розділу 3

Набір даних представлений вартістю акцій компанії Amazon за 2016-2019 роки. Завантаживши дані, можна побачити, що вони представляють собою часовий ряд, визначений 3 змінними: дата, назва компанії, ціна.

Загалом, етапи прогнозування можна описати таким чином: обробка вхідних даних, створення моделей, власне прогнозування, висновки по прогнозам.

Багато моделей мають різні вимоги до даних; деякі вимагають, щоб ряди були в порядку часу, інші вимагають відсутності пропущених значень. Перевірка ваших даних є важливим кроком для розуміння їх характеристик, і її слід завжди робити перед оцінкою моделей.

Дані не мають чітко вираженої річної сезонності, тренд змінюється з плином часу рухаючись здебільшого вгору до третьої чверті 2018 року, має різкий спадний характер до початку 2019 і знову має направлення вгору до березня 2019 року. Остання частина представлена травнем, що має направлення тренду вниз.

Багато часових рядів включають тренд, цикли та сезонність. Вибираючи метод прогнозування, нам спочатку потрібно буде визначити шаблони часових рядів у даних, а потім вибрати метод, який здатний належним чином зафіксувати шаблони.

Моделювання ARIMA починається з визначення стаціонарності даних. Але перед цим потрібно визначити кількість таких диференціацій. Іноді диференційовані дані не виглядатимуть стаціонарними, і може знадобитися диференціювати їх вдруге, щоб отримати стаціонарний ряд. Іноді звичайної диференціації може бути недостатньо, якщо у даних є сильний вплив сезонної компоненти, тому проводять сезонну диференціацію.

Моделі GAM будуть представлені таким набором: місячна адитивного типу; річна мультиплікативного; річна мультиплікативного і тижнева адитивного.

Коли відповідну модель визначено, оцінено та перевірено, настав час створювати прогнози за допомогою forecast(). Найпростіший спосіб використання цієї функції – вказати кількість майбутніх спостережень для прогнозування.

Досить часто корисним буде оцінити також наскільки широкими виходять довірчі межі і для цього можна застосувати оцінку Вінклера.

За результатами порівняння прогнозів за різними метриками модель ARIMA 1 виявилась найкращою. З отриманих даних видно, що при порівнянні оцінок точності для інтервалів довіри модель ARIMA 4 виявилась найкращою.

За результатами порівняння для комбінованого прогнозування отримуємо висновок, що використання комбінованого підходу не дало покращення загального результату – модель ARIMA 1 так і залишилася найкращою.

ВИСНОВКИ

В результаті виконання роботи було проаналізовано попередні дослідження на схожі теми, зокрема обрано моделі, що гідно показали себе у цих роботах, методи машинного навчання для обробки даних та методи покращення результатів прогнозування. Також було проаналізовано деякі економічні аспекти акцій комерційних компаній, основні характеристики та етапи прогнозування.

Для вирішення поставленої задачі було обрано середовище розробки RGui. RGui базується на R, де R – мова програмування і програмне середовище для статистичних обчислень, аналізу та зображення даних в графічному вигляді.

За основу прогнозування було обрано модифікований алгоритм ARIMA. У статистиці та економетриці, зокрема в аналізі часових рядів, модель авторегресійної інтегрованої ковзної середньої (ARIMA) є узагальненням моделі авторегресійної ковзної середньої (ARMA). Обидві ці моделі адаптуються до даних часових рядів або для кращого розуміння даних, або для прогнозування майбутніх точок у ряді (прогнозування). На противагу їй використано GAM. GAM забезпечує гнучку специфікацію відповіді, визначаючи модель у термінах гладкої функції як заміну детальних параметричних зв’язків на коваріатах. Ця гнучкість і доцільність досягаються ціною представлення гладких функцій у подібному шаблоні та вибору рівня гладкості.

Побудувавши відповідні моделі та прогнози по ним, було також використано підхід комбінування моделей для покращення результату. За результатами порівняння отримуємо висновок, що використання комбінованого підходу не дало покращення загального результату на обраному наборі даних, але також визначено, що для набору даних, де одна модель дає занадто оптимістичний, а інша занадто песимістичний прогнози, підхід по середньому покаже себе набагато краще.

За більшістю метрик модель ARIMA(0, 1, 0)$(2, 0, 0)_7$ показала найкращий результат прогнозування з мінімальним рівнем помилок. Побудована

інтелектуальна система прогнозування вартості комерційних компаній на основі методів машинного навчання допомогла досягти поставленої мети, а саме – підвищення якості прогнозування вартості комерційних компаній за рахунок виконання прогнозного оцінювання послідовності спостережень та динаміки змін їх в майбутньому. Порівняння відбувалося з базовою моделлю ARIMA, що в автоматичному режимі пропонується середовищем розробки. За метрикою MAPE ця модель показала 5,27% похибки, а отримана в результаті модифікованого алгоритму модель 3,29%.

В рамках магістерської роботи було розроблено лабораторну роботу для п’ятого курсу на тему «Аналіз часових рядів і прогнозування за моделлю ARIMA».

В ході дослідження умов роботи за комп’ютером, в рамках розділу з охорони праці, було визначено, що законодавство України чітко регламентує правила та вимоги щодо використання комп’ютерної техніки на підприємстві, безпосередньо та охорони праці при роботі з комп’ютером.

Магістерська робота пройшла апробацію на XXV Всеукраїнській науково-практичній конференції «МОГЛИЛЯНСЬКІ ЧИТАННЯ – 2022: Досвід та тенденції розвитку суспільства в Україні: глобальний, національний та регіональний аспекти».

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. What is Shares? *The Economic Times* : веб-сайт. URL: https://economictimes.indiatimes.com/definition/shares (дата звернення: 10.02.2023).

2. What can be forecast? *Otexts* : веб-сайт. URL: https://otexts.com/fpp3/what-can-be-forecast.html (дата звернення: 10.02.2023).

3. Forecasting data and methods. *Otexts* : веб-сайт. URL: https://otexts.com/fpp3/data-methods.html (дата звернення: 10.02.2023).

4. The basic steps in a forecasting task. *Otexts* : веб-сайт. URL: https://otexts.com/fpp3/basic-steps.html (дата звернення: 10.02.2023).

5. Othman M. S., Ghadeer J. M. M. A modified ARIMA model for forecasting chemical sales in the USA. *Journal of Physics: Conference Series.* 2021. doi:10.1088/1742-6596/1879/3/032008.

6. Alsharif M. H., Younes M. K., Kim J. Time Series ARIMA Model for Prediction of Daily and Monthly Average Global Solar Radiation. *Symmetry*. 2019. 11(2):240. https://doi.org/10.3390/sym11020240.

7. Nashirah A. B., Sofian R. Autoregressive Integrated Moving Average (ARIMA) Model for Forecasting Cryptocurrency Exchange Rate in High Volatility Environment: A New Insight of Bitcoin Transaction. *International Journal of Advanced Engineering Research and Science (IJAERS)*. 2017. Vol. 4. DOI:10.22161/ijaers.4.11.20.

8. Bata, M., Carriveau, R., Ting, D.SK. Short-term water demand forecasting using hybrid supervised and unsupervised machine learning model. *Smart Water 5.* (2020). https://doi.org/10.1186/s40713-020-00020-y.

9. Durdu Ö. F. A hybrid neural network and ARIMA model for water quality time series prediction. *Engineering Applications of Artificial Intelligence*. 2010. Vol. 23. P. 586-594. https://doi.org/10.1016/j.engappai.2009.09.015.

10. Sun J. Forecasting COVID-19 pandemic in Alberta, Canada using modified ARIMA models. *Computer Methods and Programs in Biomedicine Update*. 2021. Vol. 1. https://doi.org/10.1016/j.cmpbup.2021.100029.

11. Zhu N., Zhang W., Wang W., Li X. A Novel Coronavirus from Patients with Pneumonia in China. *New England Journal of Medicine*. 2019. https://doi.org/10.1056/NEJMoa2001017.

12. David L. H. Data sharing and outbreaks: best practice exemplified. *The Lancet*. 2020. P. 469-470. https://doi.org/10.1016/S0140-6736(20)30184-7.

13. Wang Y., Shen Z., Jiang Y. Comparison of ARIMA and GM(1,1) models for prediction of hepatitis B in China. *PLoS ONE*. 2018. https://doi.org/10.1371/journal.pone.0201987.

14. Liu Q., Li Z., Ji Y., Martinez L., Zia U. H., Javaid A., Lu W., Wang J. Forecasting the seasonality and trend of pulmonary tuberculosis in Jiangsu Province of China using advanced statistical time-series analyses. *Infect Drug Resist*. 2019. doi: 10.2147/IDR.S207809.

15. Rubio L., Alba K. Forecasting Selected Colombian Shares Using a Hybrid ARIMA-SVR Model. *Mathematics*. 2022. https://doi.org/10.3390/math10132181.

16. Jung G., Choi S. Forecasting Foreign Exchange Volatility Using Deep Learning Autoencoder-LSTM Techniques. *Machine Learning Applications in Complex Economics and Financial Networks*. 2021. https://doi.org/10.1155/2021/6647534.

17. Tripathi M., Kumar S., Inani S. Exchange Rate Forecasting Using Ensemble Modeling for Better Policy Implications. *Journal of Time Series Econometrics*. 2021. P. 43-71. https://doi.org/10.1515/jtse-2020-0013.

18. Wei H., Yoshiteru N., Shou-Yang W. Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*. 2005. Vol. 32. P. 2513-2522. https://doi.org/10.1016/j.cor.2004.03.016.

19. Wint N. C. Time Series Data Mining: Comparative Study of ARIMA and Prophet Methods for Forecasting Closing Prices of Myanmar Stock Exchange. *Journal of Computer Applications and Research*. 2020. Vol.1.

20. Package ‘fable.prophet’. *cran.r-project.org* : веб-сайт. URL: http://lib.znau.edu.ua/jirbis2/images/phocagallery/2017/Pryklady_DSTU_8302_2015.pdf (дата звернення: 10.02.2023).

21. Janiesch C., Zschech P., Heinrich K. Machine learning and deep learning. *Electron Markets.* 2021. Vol. 31, P. 685–695. https://doi.org/10.1007/s12525-021-00475-2.

22. Young T., Hazarika D., Poria S., Cambria E. Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational Intelligence Magazine*. 2018. Vol. 13. no. 3. P. 55-75. doi: 10.1109/MCI.2018.2840738.

23. Goodfellow I., Bengio Y., Courville A. Deep Learning. *MIT Press*. 2016. URL: https://www.deeplearningbook.org (дата звернення: 10.02. 2023).

24. Russell S. J., Norvig P. Artificial Intelligence: A Modern Approach : навч. посіб. Вид. 3-е. 2010. 1151 с. URL: https://zoo.cs.yale.edu/classes/cs470/materials/aima2010.pdf (дата звернення: 10.02. 2023).

25. Serena H. C., Anthony J. J., John P. N. Artificial Intelligence techniques: An introduction to their use for modelling environmental systems. *Mathematics and Computers in Simulation*. 2008. Vol. 78. P. 379-400. https://doi.org/10.1016/j.matcom.2008.01.028.

26. Brynjolfsson E., McAfee A. The Business of Artificial Intelligence. *Harvard Business Review*. 2017. URL: https://hbr.org/2017/07/the-business-of-artificial-intelligence (дата звернення: 10.02.2023).

27. Jordan M. I., Mitchel T. M. Machine learning: Trends, perspectives, and prospects. *Science*. 2015. Vol. 349. P. 255-260. DOI: 10.1126/science.aaa8415.

28. Hastie T., Tibshirani R., Wainwrith M. Statistical Learning with Sparsity : монографія. 2015.

29. Muthukrishnan R., Rohini R. LASSO: A feature selection technique in predictive modeling for machine learning. *IEEE International Conference on Advances in Computer Applications (ICACA)*. 2016. P. 18-20, doi: 10.1109/ICACA.2016.7887916.

30. Bishop C. M. Pattern recognition and machine learning : підручник. Information Science and Statistics. 2006. 758 с.

31. What is statistical analysis? *Whatis* : веб-сайт. URL: https://www.techtarget.com/whatis/definition/statistical-analysis (дата звернення: 10.02.2023).

32. What is Statistical Analysis? Types, Methods and Examples. *Simplilearn* : веб-сайт. URL: https://www.simplilearn.com/what-is-statistical-analysis-article (дата звернення: 10.02.2023).

33. What is R? *Webarchive* : веб-сайт. URL: https://web.archive.org/web/20080724195808/http://wiki.r-project.org/rwiki/doku.php?id=getting-started:what-is-r:what-is-r (дата звернення: 10.02.2023).

34. Autoregressive integrated moving average. *Wikipedia* : веб-сайт. URL: https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average (дата звернення: 10.02.2023).

35. Time Series Forecasting With ARIMA In R. *Levelup* : веб-сайт. URL: https://levelup.gitconnected.com/time-series-forecasting-with-arima-in-r-a0d1f8f8b92f (дата звернення: 10.02.2023).

36. ARIMA models. *Otexts* : веб-сайт. URL: https://otexts.com/fpp3/arima.html (дата звернення: 10.02.2023).

37. Hocking R. R. The Analysis of Linear Models : навч. посіб. Вид. 1-е. 1985. 400 с.

38. Ravindra K., Rattan P., Mor S., Aggarwal A. N. Generalized additive models: Building evidence of air pollution, climate change and human health. *Environment International.* 2019. Vol. 132.

39. Dehghan, A., Khanjani, N., Bahrampour, A. The relation between air pollution and respiratory deaths in Tehran, Iran- using generalized additive models. *BMC Pulm Med.* Vol. 18**.** 2018. https://doi.org/10.1186/s12890-018-0613-9

40. Ravindra K. Emission of black carbon from rural households kitchens and assessment of lifetime excess cancer risk in villages of North India. *Environment International*. 2019. Vol. 122. P. 201-212. https://doi.org/10.1016/j.envint.2018.11.008.

41. Hastie T. J. Generalized Additive Models : навч. посіб. 1992. 59 с.

42. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning. *Springer New York*. 2021. 607 с. https://doi.org/10.1007/978-1-0716-1418-1.

43. Wood S.N. Generalized Additive Models: An Introduction with R. *Chapman and Hall/CRC*. 2006. https://doi.org/10.1201/9781420010404.

44. Moritz S., Bartz-Beielstein T. ImputeTS: Time Series Missing Value Imputation in R. URL: https://cran.r-project.org/web/packages/imputeTS/vignettes/imputeTS-Time-Series-Missing-Value-Imputation-in-R.pdf.

45. Nguyen H.V., Naeem M. A., Wichitaksorn N., Pears R. A smart system for short-term price prediction using time series models. *Computers & Electrical Engineering*. 2019. vol. 76. P. 339-352.

46. Охорона праці при роботі з комп’ютером. *Довідник* : веб-сайт. URL: https://cutt.ly/t2jJxve (дата звернення: 10.02.2023).

47. Державні санітарні правила і норми роботи з візуальними дисплейними терміналами електронно-обчислювальних машин. *Закон* : веб-сайт. URL: https://zakon.rada.gov.ua/rada/show/v0007282-98#Text (дата звернення: 10.02.2023).

ДОДАТОК А

Фрагмент коду створення комбінованої моделі на основі модифікованого методу ARIMA та налаштованої GAM

```
companies_price <- read_excel("share_price.xlsx")
str(companies_price)
companies_price$ds = as.Date(companies_price$ds)
companies_price$price = as.numeric(companies_price$price)
unique(companies_price$share)
amazon_price <- select(filter(companies_price, share == "amzn"), ds, price)
str(amazon_price)
amazon_price <- as_tsibble(amazon_price, key=NULL, index=ds)
plot(amazon_price$ds, amazon_price$price, type = "l")
statsNA(amazon_price$price)
amazon_price$price <- na_kalman(amazon_price$price)
summary(amazon_price)
plot(amazon_price$ds, amazon_price$price, type = "l")
abline(reg=lm(amazon_price$price~amazon_price$ds, amazon_price), col="blue")
colnames(amazon_price) <- c('ds', 'y')
all_data <- amazon_price
train <- amazon_price %>% filter(ds < as.Date("2019-05-17"))
test <- amazon_price %>% filter(ds >= as.Date("2019-05-17"))
amazon_price <- train
ACF(amazon_price, lag_max = 36) %>% autoplot()
PACF(amazon_price, lag_max = 36) %>% autoplot()
nonlinearityTest(amazon_price$y, TRUE)
adf.test(amazon_price$y)
kpss.test(amazon_price$y)
pp.test(amazon_price$y)
model_quality <- function(model){
r2 <- cor(fitted(model)$.fitted, amazon_price[2])^2
dw <-sum((residuals(model)$.resid - lag(residuals(model)$.resid))^2, na.rm =
TRUE)/sum(residuals(model)$.resid^2, na.rm = TRUE)
return(cat("R^2 = ", r2, "DW = ", dw, "\n"))}
comb_model <- amazon_price %>% model(prophet = prophet(y ~ season("year", 10, 1,
type = "multiplicative") +
season("week", 10, 1, type="additive")), arima = ARIMA(y)) %>% mutate (combination
= (prophet + arima) / 2)
comb_model_forecast <- comb_model %>% forecast(h = 10)
autoplot(comb_model_forecast, amazon_price)
autoplot(comb_model_forecast, amazon_price) + coord_cartesian(
xlim = c(as.Date("2019-05-04"), as.Date("2019-06-01")), ylim = c(1800, 2050))
accuracy(comb_model_forecast, all_data, list(winkler = winkler_score), level = 80)
accuracy(comb_model_forecast, all_data, list(crps = CRPS))
```