МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Харківський національний університет радіоелектроніки
Факультет Комп'ютерних наук
Кафедра Програмної Інженерії

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

другий (магістерський)
(рівень вищої освіти)

Дослідження методів прогнозування показників електронної комерції під час соціальних катастроф

Виконав:
студент 2 курсу, групи ІПЗм-21-2
Ховрат А.В.
(прізвище, ініціали)

Спеціальність 121 – Інженерія програмного забезпечення
Тип програми освітньо-наукова
Керівник доц. Назаров О.С.
(посада, прізвище, ініціали)

Допускається до захисту
Зав. кафедри _____ З.В. Дудар
(підпис) (прізвище, ініціали)

2023 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Програмної інженерії

Рівень вищої освіти другий (магістерський)

Спеціальність 121 – Інженерія програмного забезпечення
(код і повна назва)

Тип програми освітньо-наукова програма

Освітня програма Інженерія програмного забезпечення
(повна назва)

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

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студента Ховрату Артему Вячеславовичу
(прізвище, ім'я, по батькові)

1. Тема роботи: Дослідження методів прогнозування показників електронної комерції під час соціальних катастроф

затверджена наказом університету від «29» березня 2023 р. № 302 Ст

3. Термін подання студентом до екзаменаційної комісії «12» травня 2023 р.

4. Вихідні дані до роботи *встановлений календарний план роботи, методичні вказівки до оформлення пояснювальної записки, методи прогнозування засновані на векторній авторегресії.*

5. Перелік питань, що потрібно опрацювати в роботі *аналіз предметної галузі, огляд наявних математичних моделей, модифікація базових алгоритмів, дослідження можливості прискорення базових моделей, створення плану експерименту для дослідження, опис імплементації алгоритмів, дослідження результатів* .

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Аналіз предметної галузі	29.03.2023	виконано
2	Виявлення проблематики галузі	31.03.2023	виконано
3	Здійснення огляду математичних моделей	03.04.2023	виконано
4	Розробка алгоритму передобробки даних	06.04.2023	виконано
5	Дослідження можливості прискорення	08.04.2023	виконано
6	Побудова плану експерименту	10.04.2023	виконано
7	Імплементація прототипів обраних моделей	12.04.2023	виконано
8	Аналіз результатів експерименту	13.04.2023	виконано
9	Написання пояснювальної записки	18.04.2023	виконано
10	Підготовка доповіді та презентації	26.04.2023	виконано
11	Підготовка роботи для перевірки на антиплагіат та проходження нормоконтролю	30.04.2023	виконано
12	Оцінка роботи рецензентом	05.05.2023	виконано
13	Отримання відзиву від керівника роботи	06.05.2023	виконано
12	Попередній захист кваліфікаційної роботи	09.05.2023	виконано
13	Здача роботи у електронний архів	09.05.2023	виконано
14	Отримання допуску до захисту	10.05.2023	виконано
15	Захист кваліфікаційної роботи	12.05.2023	виконано

Дата видачі завдання 29 березня 2023 р.

Студент _____
(підпис)

Керівник роботи _____ доц. Назаров О.С.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка до кваліфікаційної роботи, 120 сторінок, 63 рисунки, 7 таблиць, 4 додатки, 43 джерела.

АНАЛІЗ ДАНИХ, ВЕКТОРНА АВТОРЕГРЕСІЯ, ЕКОНОМІКА, ЕЛЕКТРОННА КОМЕРЦІЯ, НЕВИЗНАЧЕНІСТЬ, ПРОГНОЗУВАННЯ, ПРИЙНЯТТЯ РІШЕНЬ, РИЗИК, СОЦІАЛЬНА КАТАСТРОФА.

Об’єктом дослідження є методи прогнозування показників електронної комерції під час соціальних катастроф.

Метою роботи є визначення ефективності використання модифікованих авторегресійних моделей для прогнозування показників ділової активності ринку електронної комерції під час соціальних катастроф.

У результаті роботи було досліджено ефективність використання авторегресійних моделей для прогнозування показників ділової активності ринку електронної комерції під час соціальних катастроф; розроблено алгоритм передобробки поліморфних даних про соціальну катастрофу, цільову аудиторію та ринкову кон’юнктуру; визначено можливості паралелізації авторегресійних моделей за допомогою використання технології MapReduce та виграш у швидкості виконання алгоритмів від її використання.

DATA ANALYSIS, VECTOR AUTOREGRESSION, ECONOMICS, E-COMERCE, UNCERTAINTY, FORECASTING, DECISION MAKING, RISK, SOCIAL DISASTER.

The object of the study is methods of forecasting e-commerce indicators during social disasters.

The purpose of the work is to determine the effectiveness of using modified autoregression models for forecasting indicators of business activity of the e-commerce market during social disasters.

As a result of the work, the effectiveness of using autoregression models for forecasting indicators of business activity of the e-commerce market during social

disasters was investigated; an algorithm for refining polymorphic data on social catastrophe, target audience and market conditions was developed; the possibilities of parallelization of autoregressive models using the MapReduce technology and the gain in algorithm execution speed from its use are determined.

Умови публікації пояснювальної записки

Я, Ховрат Артем Вячеславович
(прізвище, ім'я, по батькові)
студент групи ІПЗм-21-2 здобувач вищої освіти на другому (магістерському) рівні
кафедра програмної інженерії,
(повна назва кафедри)
заявляю: моя кваліфікаційна робота на тему
Дослідження методів прогнозування показників електронної комерції під час соціальних катастроф
(назва роботи)
що буде представлена до ЕК для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElArKhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлений (а) з діючим положенням «Про протидію академічному плагіату в ХНУРЕ», згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

ЗМІСТ

ВСТУП

За останні п’ять років світова економіка пережила декілька значних спадів, спочатку викликаних заворушеннями в наслідок поширення коронавірусу, а згодом і війною в Україні. За даними Міжнародної організації праці збитки лише від пандемії перевищили 3 трильйони доларів США [1], а вторгнення Російської Федерації на територію України вже спровокувало енергетичну, продовольчу та міграційну кризи у Європі та Африці [2]. Ефект від зазначених явищ і подібних до них можна розглядати на різних рівнях, однак в рамках чинної роботи зосереджуватимемося на їх опосередкованому впливі на економіку шляхом соціальних зсувів.

За даними дослідників системних ризиків з Університету Штутгарта, соціальний зсув – ситуація за якої поведінка людини змінюється в наслідок відчуття загрози певного рівня до себе чи до свого найближчого оточення [3]. Дослідження проведені Університетом Кембриджа у 2014 році показали важливість врахування цього аспекту при веденні бізнесу. Однак у цій праці в більшості розглядалися локальні зсуви, які могли викликати динаміку базових ринкових показників на рівні декількох відсотків [4]. Данна робота зосереджує увагу на більш рідкісних явищах, які здатні змінити кон’юнктуру ринку в цілому, а в якості прикладу такого ринку, було вирішено обрати ринок електронної комерції.

У подальшому суттєві соціальні зсуви називатимемо «соціальними катастрофами». Їх тривалість визначається без прив’язки до причин, що викликали девіантний зсув. Подібна віддаленість від першоджерела, відповідно до матеріалів Світового банку, пояснюється пост-травматичним стресовим розладом, симптоми якого, навіть на рівні соціальних груп, можуть проявлятися впродовж декількох десятиліть. Наприклад, теракт 11 вересня 2001 року у США досі має певний вплив на процес прийняття рішень населення.

Загалом процес урахування ірраціональних факторів при аналізі ризиків чи прогнозуванні своєї ринкової діяльності є доволі комплексним завданням. Це

одна з причин того, чому в неокласичній економічній теорії подібні складові ігнорувалися. Однак соціальні катастрофи можуть викликати занадто суттєві зміни в наслідок яких такі підходи надаватимуть істотно некоректні результати (з точністю меншою за 70%) [5]. Однією з можливостей вирішення цієї проблеми є застосування комп'ютерних технологій, які, на відміну від людини, здатні швидко імплементувати набір певних математичних операцій для великих обсягів даних. Наразі зазначений підхід є основним при створенні систем підтримки прийняття рішень різного типу та систем аналізу ризиків. В основі вказаного рішення можуть лежати різні технології, однак в рамках даного дослідження було вирішено сконцентруватися на векторних авторегресійних моделях. Подібна фіксація пояснюється розповсюдженістю цього сімейства алгоритмів, суттєвою точністю його застосування для числових даних та простотою врахування великої кількості показників [6]. Окрім цього гнучкість та простота реалізації чинного підходу, що було виявлена у численних дослідженнях [7, 8], у яких було здійснено порівняння з нейронними мережами [9] чи байєсівськими моделями [10], зумовлює доцільність акцентування даної роботи саме на ньому.

За вказаних позитивних рис у векторної авторегресії наявний суттєвий недолік – швидкість обробки інформації. При розгляді економічних даних, що стосуються ділової активності в умовах соціальних катастроф час прийняття рішень відіграє важливу роль. Задля їх пришвидшення варто застосувати методи паралелізації алгоритмів, наприклад, виконання обчислень у різних потоках обробки інформації. З огляду на сучасні дослідження, класичні методи розпаралелювання не дають істотного результату і в якості найбільш популярної альтернативи наразі розглядається технологія MapReduce [11].

Окрім вказаного варто наголосити, що авторегресія дозволяє спрогнозувати деякий числовий результат, враховуючи навчальну вибірку. Тож, не враховуючи безпосереднього вибору технології для аналізу, необхідно розробити алгоритм, що дозволить врахувати вплив соціальних катастроф на цільову галузь економіки, створивши відповідні числові показники

Таким чином можемо сконстатувати, що у рамках чинної роботи буде створено алгоритм передобробки інформації, що б дозволив врахувати вплив соціальних катастроф на ринок електронної комерції; реалізовано комбіноване застосування цього алгоритму з авторегресійними моделями прогнозування та досліджена можливості розпаралелювання зазначених алгоритмів за допомогою технології MapReduce.

Метою кваліфікаційної роботи є визначення ефективності використання модифікованих авторегресійних моделей для прогнозування показників ділової активності ринку електронної комерції під час соціальних катастроф.

Предметною галуззю чинного дослідження, з огляду на окреслену проблематику, є галузь аналізу та управління ризиками. Це у свою чергу дозволяє акцентувати увагу на доцільність використання результатів проведеної роботи для прийняття управлінських рішень.

1 ОПИС ПРОБЛЕМНОЇ ГАЛУЗІ

1.1 Аналіз предметної області

Поняття соціальної вразливості найчастіше розглядається у розрізі аналізу ризиків [12], який має три ключових етапи: визначення, оцінка та управління. У 20 століття для оптимізації цього процесу були розроблені системи підтримки прийняття рішень (СППР) [13]. Станом на сьогодні ці системи активно імплементуються у різноманітні бізнес-середовища, де їх поділяють за наступними критеріями [14, 15]:

– спосіб підтримки: орієнтовані на знання, на документи, на дані, на комунікації та на моделі;
– взаємодія з користувачем: кооперативна, активна, пасивна.

З огляду на те, що було вирішено розглядати лише авторегресійні моделі, надалі у чинній роботі братимуться до уваги лише орієнтовані на дані системи як пасивні, так і активні. Нейтральна конотація по відношенню до взаємодії з користувачем, пояснюються тим, що активна частина системи, яка надає певні поради, фактично є окремим алгоритмом.

З огляді на профільні рейтингові платформи, найбільш популярними у 2022 році були наступні СППР [16]:

– Hyperproof;
– Soterion;
– Whistic.

Тут варто зауважити, що деякі з них окрім авторегресійного підходу імплементують байєсів підхід та/або методи штучного інтелекту [17, 18]. Аби додатково обґрунтувати вибір сімейства алгоритмів здійснимо поступовий огляд кожної зазначеної системи, аби з’ясувати на яких показниках вони концентруються та які алгоритми використовують. Окрім цього скористаємось відгуками про вказані системи для формування більш об’єктивної оцінки.

Необхідно наголосити, що опорний алгоритм, закладений в середину СППР може бути частиною комерційної таємниці компанії, тому ця інформація не завжди є у відкритих джерелах.

Почнемо з «Hyperproof» (див. рис. 1.1).

Рисунок 1.1 – Домашня сторінка «Hyperproof» [19]

Ця система є активною СППР, що ґрунтується на хмарних технологіях для прискорення обчислень. З огляду на наявну інформацію вона використовує різні засоби у тому числі можливо і авторегресію. Система є пропрієтарною та працює за принципом підписки, однак аналіз відгуків показує на те, що задля повноцінного прогонозування за класичним сценарієм і отримання порад, необхідно надати велику кількість даних, які при цьому довго оброблюються. Водночас процес прогнозування відбувається швидко. Враховуючи специфіку трьох звичних алгоритмічних підходів, можна стверджувати, що «Hyperproof» реалізований або на базі штучного інтелекту, або авторегресії.

Окремо варто зауважити, що ризики на яких зосереджується система пов’язані безпосередньо з виробництвом та кібербезпекою.

З тої інформації, яка є у відкритому доступі, можна зробити висновок, що система не розглядає соціальні катастрофи чи будь-які подібні явища окремо. На підтвердження цієї думки вказує інформація з рейтингової платформи G2 [16]. Під час найбільшої хвилі пандемії коронавірусу кількість користувачів системи

впала, отже, прогнозування швидше за все було не вдалим, тож зсуви викликані надзвичайними явищами, скоріше за все, враховуються як шум чи погрішність.

Перейдемо до системи «Soterion» (див. рис. 1.2). На відміну від попереднього рішення, вона не концентрує свою увагу на ризиках а надає компанії-користувачу повноцінну SAP-систему. У відкритих джерелах було знайдено те, що вона здатна до аналізу корпоративних ризиків, однак їх перелік обмежений і навіть не враховує вже названі ризики кібербезпеки. У цьому випадку, відомо, що задля прогнозування окремих показників ділової активності система використовує авторегресію. Однак, яка з модифікацій, не згадується.

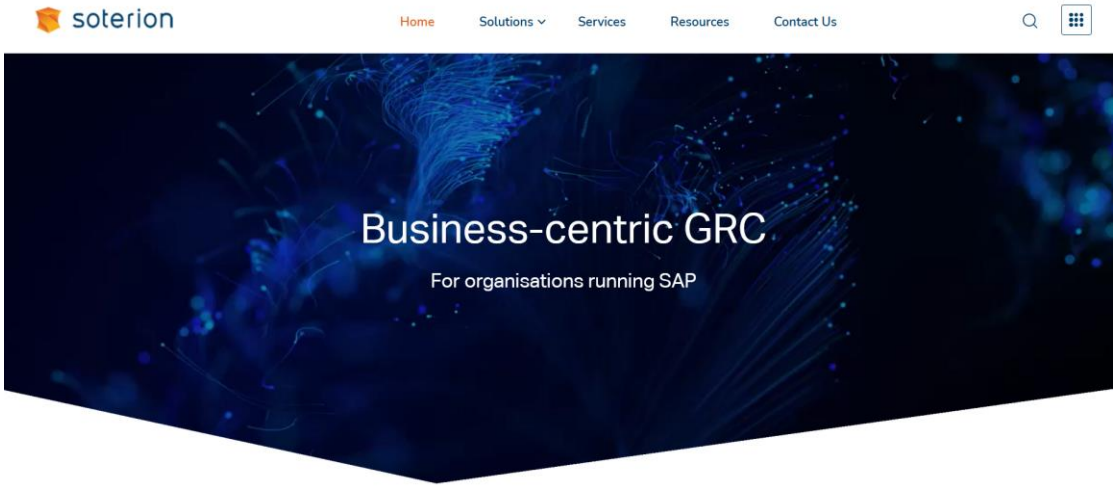


Рисунок 1.2 – Домашня сторінка «Soterion» [20]

З початку пандемії популярність системи зростала, проте після протестів у Сполучених Штатах та вже згодом за повномасштабного вторгнення Російської Федерації в Україну, почала спадати. Таким чином можна стверджувати, що наявна певна нестабільність у врахуванні тих чи інших показників. Можна висунити припущення: як і «Hyperproof», так і «Soterion» не враховують соціальні зсуви як один із ключових факторів впливу під час надзвичайних ситуацій, а сплеск популярності після початку пандемії COVID-19, скоріше за все, пояснюються загальним ростом популярності до подібних систем.

Перейдемо до заключної системи – «Whistic» (див. рис. 1.3) – що також є активною СППР.

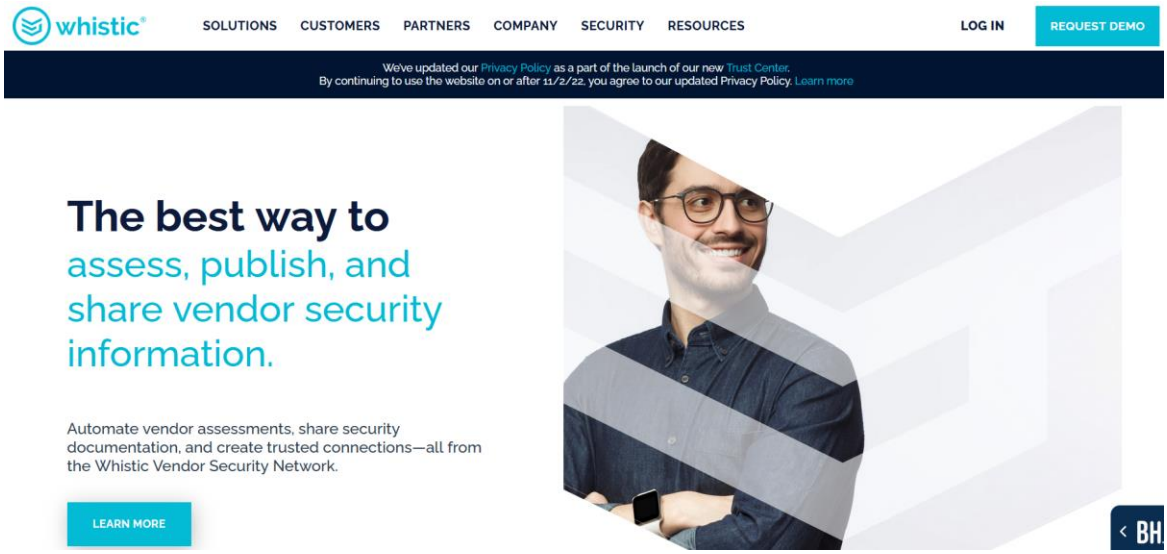


Рисунок 1.3 – Домашня сторінка «Whistic» [21]

Зазначена система, на відміну від попередніх, концентрується суто на безпеці компанії. Модуль прогнозування в ній присутній і, з огляду на наявну інформацію, він теж може використовувати авторегресію. Виходячи з описів, що були знайдені у відкритих джерелах, можна зробити висновок що система враховує значну волатильність показників компанії під час кризи, однак сказати напевно, що саме береться до уваги, не є можливим.

Окрім наведених трьох систем було вирішено переглянути ще 10 з рейтингу G2 та ознайомитися з відгуками про них. З огляду на те, що найголовнішими показниками під час соціальних катастроф можна вважати точність прогнозу та час його отримання (бо від цього залежить швидкість реакції управління), а при цьому до всіх систем з рейтингу є критика, щодо цих критеріїв, можна стверджувати, що навіть якщо авторегресія застосовується, то вона не враховує соціальні заворушення. Це в свою чергу підтверджує актуальність чинного дослідження. Аналіз відгуків, також показав, що системи, які відкрито використовують імовірнісний підхід працюють значно повільніше, а ті, що сконцентровані на штучному інтелекті, часто ігнорують будь-який з видів значної волатильності (тут під значною мається на увазі миттєва зміна індексу S&P більше ніж на 3%). Таким чином можемо стверджувати про доцільність вибору саме авторегресії в рамках обраної задачі.

Оскільки аналіз СППР вказав лише на те, що використовується авторегерсія, при цьому не зазначаючи відповідних типів, було вирішено провести додаткове дослідження наукових джерел. З огляду на його результати можна зробити висновок, що найбільш вживаними алгоритмами є: класична векторна авторегресія, сезонна векторна авторегресія, векторна авторегресія розподіленого лагу, векторна авторегресія рухомого середнього та векторна авторегресія інтегрованого рухомого середнього.

Варто зауважити, що аби врахувати соціальну катастрофу як зовнішній вплив необхідно розглядати модифікації зазначених алгоритмів з урахуванням екзогенних змінних.

Таким чином можемо сформувати наступний набір альтернатив для багатокритеріальної задачі прийняття рішень:

- EC-VAR: модель векторної авторегресії з екзогенними регресорами;
- EC-VARS: модель векторної сезонної авторегресії з екзогенними регресорами;
- EC-VARL: модель векторної авторегресії розподіленого лагу з екзогенними регресорами;
- EC-VARMA: модель векторної авторегресії рухомого середнього з екзогенними регресорами;
- EC-VARIMA: модель векторної авторегресії інтегрованого рухомого середнього з екзогенними регресорами.

1.2 Аналіз ринку електронної комерції

Ринок електронної комерції у цілому можна вважати відносно стабільним у середньостроковій перспективі, це пояснюється наступними особливостями:

- на ньому функціонує певний набір великих міжнародних гравців (наприклад, Amazon чи Alibaba), декілька великих гравців національного рівня (для України можна розглядати компанію Rozetka),

та значна кількість малих гравців (особливо, якщо враховувати бізнес у соціальних мережах), тобто він є монополістичною конкуренцією;

– кількість товарів та їх тип для великих компаній є сильно диверсифікованим: від продуктів та друкованої продукції до побутової техніки і специфічних знарядь праці. Вказаний перелік охоплює товари першої необхідності, тому у разі надзвичайних ситуацій загальний попит на продукції певного магазину може не змінитися.
– суттєва залежність від наявності інтернету та вразливість до проблем пов’язаних з кібербезпекою створює підґрунтя для сучасних «надзвичайних ситуацій», наприклад, витік особистих даних із сайту магазину.

Загалом компанії, які взаємодіють на ринку електронної комерції, можна поділяти за різними критеріями. У рамках чинного дослідження при визначенні факторів, що необхідно врахувати у алгоритмі передобробки, братимуться до уваги наступні:

– розмір: малі, середні, великі (за кількістю покупців);
– наявність у компанії торговельної платформи: якщо платформа є, то необхідно враховувати сильний зв’язок із доходом компанії та малим бізнесом, що функціонує на платформі;
– наявність у компанії не онлайн відділень: можна зауважити, що в залежності від характеру надзвичайної ситуації, бізнес, який здатний торгувати за межами інтернету буде або отримувати надприбутки (наприклад, під час війни, в регіонах-реципієнтах біженців), або нести збитки (наприклад, під час пандемії).

Стосовно цільових показників, то ними можуть бути наступні дані (перелік не є обмеженим):

– дохід/прибуток на конкретні товари;
– дохід/прибуток від магазинів (у випадку, якщо у компанії декілька інтернет-магазинів чи компанія має реальні відділення);

– біржові показники (ціни акцій на момент закриття або відкриття біржі, аукціонні ціни, обсяги торгівлі тощо).

1.3 Постановка задачі

З огляду на інформацію наведену у попередньому пункті можна сформувати задачу чинної роботи – проведення дослідження, яке порівнює ефективність застосування алгоритмів сімейства векторної авторегресії для прогнозування показників ділової активності на ринку електронної комерції.

Тут під словом ефективність мається на увазі наступний набір показників:

– час прогнозування;
– точність прогнозу;
– якість врахування екзогенних показників;
– час на підготовку даних;
– можливість врахування шумів.

Аби виконати поставлену задачу розділимо її на набір підзадач:

– здійснити ознайомлення з математичним представленням сімейства алгоритмів векторної авторегресії;
– визначити умови за яких є можливість розглянути соціальну катастрофу як фактор зовнішнього впливу;
– побудувати алгоритм який би дозволив розглядати соціальні катастрофи та побіжні фактори як кількісні змінні;
– перевірити можливість для подальшої оптимізації моделі;
– побудувати план експерименту дослідження;
– проведення експерименту;
– формалізація результатів експерименту;
– формування переліку можливих завдань для подальшого дослідження.

2 МАТЕМАТИЧНЕ ПРЕДСТАВЛЕННЯ

2.1 Огляд математичних моделей

Перед тим як перейти до модифікації базових алгоритмів сімейства VAR розглянемо їх математичне представлення [22]:

$$\Phi_0 y_t = \Phi_1 y_{t-1} + \cdots + \Phi_p y_{t-p} + \Theta_0 u_t + \Theta_1 u_{t-1} + \cdots + \Theta_q u_{t-q}, \quad (2.1)$$

де y_t – K-мірний часовий ряд;

Φ_i, Θ_j – матриці розмірності $K \times K$, $i = \overline{1,p}$, $j = \overline{1,q}$;

u_t – K-мірний вектор білого шуму з нульовим середнім та наступною невиродженою коваріаційною матрицею $\Sigma = \mathbb{E}(u_t, u_t')$.

Варто зауважити, що Φ_0 та Θ_0 можуть бути нормалізовані до 1.

Із наведеної формули (2.1) випливає те, що класичне сімейство VAR-моделей прогнозує лише статичні змінні. Аби врахувати екзогенні показники було вирішено використати модифікацію коригування помилок (скорочено EC). Подібне коригування необхідне за умови, що декілька ендогенних змінних мають спільний стохастичний тренд, що характерно для показників ринкової діяльності підприємств. Загальна формула для модифікованого EC-VAR сімейства алгоритмів матиме наступних вигляд:

$$\Phi_0 \Delta y_t = \Pi y_{t-1} + \Psi_1 \Delta y_{t-1} + \cdots + \Psi_{p-1} y_{t-p+1} + \Theta_0 u_t + \cdots + \Theta_q u_{t-q}, \quad (2.2)$$

де $\Pi = -(\Phi_0 - \Phi_1 - \cdots - \Phi_p)$;

$\Psi_i = -(\Phi_{i+1} + \cdots + \Phi_p)$, $i = \overline{1, p-1}$.

У випадку вказаному у (2.2) важливе місце посідають екзогенні змінні, власне вони і слугуватимуть базисом для врахування соціальної катастрофи у

кількісному представлені. Однак тут варто зауважити, що вони не є передумовами, не додають додаткових обмежень щодо типів і не додають економічного підґрунтя моделям.

Необхідно зазначити, що приймаючи рішення щодо змінних і лагів, які слід включити в модель треба бути обережним щоб VAR-модель не стала занадто параметризованою. Це може призвести до проблем, коли є забагато параметрів відносно спостережень у даних (тобто проблема ступенів свободи).

Наприклад, після включення 6 змінних у модель із 8 лагами (тобто дворічними лагами для квартальних даних), кожне рівняння у VAR міститиме: $1 + (6 \times 8) = 49$ коефіцієнтів (включно зі сталою).

Необхідно зауважити, що прогноз на декілька діб наперед, який здійснюється на основі екзогенних змінних потребує даних, що фактично теж є спрогнозованими і містять певну похибку.

У рамках цієї роботи було вирішено обмежтися денним прогнозом на поточний день, бо, виходячи зі специфіки галузі, таким чином можна гарантувати відсутність невизначеності у екзогенних змінних.

У макроекономічних чи фінансових дослідженнях, до яких відноситься і чинна робота, виникає ситуація обмеженості даних і неточності оцінки параметрів. За таких обставин перенасиченість параметрів буде накладатися на похибку даних.

Тож, оскільки неможливо включити всі змінні, що можуть впливати на ринкові процеси під час соціальних катастроф, було вирішено провести експертне оцінювання серед економістів, ризик-менеджерів та соціологів з метою виявлення найбільш впливових зовнішніх показників для функціонування будь-якого ринку та ринку електронної комерції зокрема.

Це у свою зменшить вплив суб’єктивності при виборі відповідних цільових факторів.

Після визначення моделі необхідно також обрати відповідну довжину затримки моделі VAR. Для цього можна використати інформаційні критерії [23].

Одним із таких є AIC (Akaike information criterion). У спрощеному вигляді його можна представити наступним чином:

$$AIC = 2k - 2ln(\hat{L}), \tag{2.3}$$

де k – кількість параметрів у моделі;

$\hat{L}$ – максимальна точність моделі.

Модель можна вважати якісною у тому випадку, якщо показник наведний у (2.3) відносно малий.

Іншим показником можна вважати BIC (Bayesian information criterion), який можна подати наступним чином:

$$BIC = kln(n) - 2ln(\hat{L}), \tag{2.4}$$

де k – кількість параметрів у моделі;

n – розмір набору даних;

$\hat{L}$ – максимальна точність моделі.

Правило щодо мінімізації критерію для (2.4) залишається.

Перед тим як перейти безпосередньо до розгляду модифікації базових алгоритмів варто зауважити, яким чином можна визначити те, що параметр є зовнішнім. Зазвичай для цього використовується тест причинності Грейнджера.

Вказаний тест передбачає кореляцію між поточним значенням однієї змінної та минулими значеннями інших, це не означає, що зміни в одній змінній викликають зміни в іншій.

У нашому випадку стан соціальна катастрофа не залежить від стану ринку електронної комерції, тому ми можемо розглядати поняття пов’язані з нею як зовнішні змінні.

2.2 Модифікація базових алгоритмів

Загалом прийнято вважати, що соціальна катастрофа впливає на всіх ринкових суб’єктів від фізичної особи до держави. Теорія бізнес-управління вказує на те, що характер впливу на фізичних осіб відображається у поведінці цільової аудиторії [24], зміни в роботі компанії та регуляції ринку в цілому (що є представленням держави) об’єднується у мікроекономічний профіль кон’юнктури ринку. З огляду на те, що для кожного показника необхідно врахувати саму катастрофу, почнемо розгляд модифікації з аналізу можливостей конвертації її опису у кількісний вигляд.

Задля того, аби якісні показники відображали об’єктивний погляд на катастрофу, було вирішено провести експертне оцінювання серед 10 ризик-менеджерів різних компаній у Харкові та Новомосковську та 10 соціологів з Харкова та Дніпра. Найбільшу кількість разів згадувались наступні чотири показники:

- загальний текстовий опис соціальної катастрофи;
- новини пов’язані з каталізуючим явищем (тут під каталізатором мається на увазі подія, яка викликала соціальні зсуви);
- суб’єктивна оцінка працівників компанії щодо готовності до непердбачуваних обставин;
- термін дії катастрофи (з моменту початку дії каталізатору).

Останній індикатор має кількісне представлення (однак, показник корегуватимемо відносно календарного року), а суб’єктивна оцінка працівників легко конвертується у кількісний вигляд, якщо проводити опитування маючи шкалу від 0 до 100, де 0 – “компанія зовсім не готова до непередбачуваних обставин”, а 100 – “компанія цілком готова до непередбачуваних обставин”. Після опитування результат нормуватиметься до 1.

Для конвертації у кількісний вигляд використаємо принципи контент-аналізу [25]. У цілому алгоритм має бути наступним:

1) видаляємо з тексту небуквені вирази;
2) розбиваємо на речення та на слова очищений текст;
3) застосовуємо операцію стемінгу – скорочення слова до його основи [26];
4) задля унеможливлення похибок стемінгу проводимо лематизацію – приведення словоформи до леми [27];
5) здійснюємо уніфікацію утвореного словника;
6) прибираємо слова з мінімальним лінгвістичним навантаженням, наприклад, сполучники;
7) знаходимо показник частотної характеристики TF-IDF [28];
8) знаходимо показник полярності кожного слова з використанням наявних корпусів слів;
9) перемножуємо TF-IDF та полярність задля знаходження частотно-полярного показника (надалі ЧПП);
10) знаходимо суму отриманого ЧПП для кожного отриманого тексту (або його частини);
11) нормалізуємо значення суми у межах від 0 до 1, де 0 – відсутність катастрофи, а 1 – суттєвий негативний вплив.

Формалізовано можемо подати характеристику соціальної катастрофи наступним чином:

$$SSD = \frac{SDO \times SDS \times t}{R}, \tag{2.5}$$

де *SDS* – результат контент-аналізу опису наданого компанією;

SDO – результат контент-аналізу опису, отриманого з новин;

t – термін дії катастрофи;

R – готовність до непередбачуваних обставин.

Тепер перейдемо до основних показників профілю кон’юнктури та цільової аудиторії. Для цільової аудиторії, по аналогії з показником соціальної катастрофи,

було вирішено провести експертне оцінювання, в ході якого виявлено, що найбільш вагомими для цього індикатору є наступні чинники:

– загальний текстовий опис цільової аудиторії;

– чисельність аудиторії [29];

– обсяг доходу у відношенні до загального доходу.

Останні два індикатори є кількісними, тож, для них достатньо здійснити нормалізацію від 0 до 1. Задля конвертації текстового опису цільової аудиторії скористаємось принципом кластерного-аналізу. Перші 5 кроків алгоритму подібні тим, що зазначалися для опису катастрофи. Заключні кроки є дещо видозміненими:

6) узагальнюємо опис аудиторії за допомогою набору характеристик запропонованим Робертом Плутчіком: очікування, злість, огида, смуток, страх, подив, надія, довіра [30, 31];
7) формуємо емоційне забарвлення кожного слова з опису аудиторії, надаючи значення кожної характеристики у межах від 0 до 100;
8) підсумовуємо отримані значення, врахувавши знак емоції;
9) нормуємо значення у межах від 0 до 1;
10) враховуємо загальні ринкові парадоксальні ситуації, які згадувались раніше (ефекти Веблена, Гіффена, сноба) шляхом введення відповідного показника, який обчислюється як модуль різниці нормованого індикатору чисельності та індикатору частки доходу.

Формалізовано можемо подати соціальну катастрофу наступним чином:

$$\begin{cases} TASD = \mu TAD|CD - RD|, (RD \geq 0.5 \wedge CD \geq 0.5) \vee (RD < 0.5 \wedge CD < 0.5) \\ TASD = TAD, (RD > 0.5 \wedge CD < 0.5) \vee (RD < 0.5 \wedge CD > 0.5) \end{cases}, \quad (2.6)$$

де μ – коефіцієнт нормалізації ринкових парадоксів;

TAD – результат контент-аналізу опису цільової аудиторії;

CD – нормований індикатор чисельності;

RD – індикатор частки доходу.

Варто зауважити, що формула (2.6) не враховує вплив соціальної катастрофи вказаний у (2.5), тож для чинної задачі введемо наступну формулу:

$$TAOI = TASD^{SSD}. \tag{2.7}$$

Усі алгоритми модифікованого сімейства векторної авторегресії розглядають зовнішні змінні як числові ряди, а не скаляри, тож отриманий у результаті обчислення за формулою (2.7) показник необхідно векторизувати. Аби досягти цього використаємо отриманий результат у якості центру нормального розподілу, який слугуватиме екзогенною змінною для обраних алгоритмів модифікованого сімейства EC-VAR.

Перейдемо до наступного показника, що необхідно врахувати – профіль ринкової кон'юнктури. Відповідно до мікроекономічної теорії він може включати велику кількість різни показників, однак після проведення опитування серед експертів, було виявлено наступні характеристики, що враховуватимуться надалі:

– обсяги інноваційної діяльності;
– фінансова стабільність;
– монополізація ринку;
– стан цільової галузі та світової економіки в цілому;
– соціальна катастрофа, що виникла.

Аби врахувати монополізацію ринку будемо користуватися індексом Ґерфіндаля-Ґіршмана [32]:

$$HHI = \sum_{j=1}^{N} s_j^2, \tag{2.8}$$

де N – кількість компаній на обраному ринку;

s – частка ринку, що належить компанії.

З огляду на те, що ринок електронної комерції має характеристики монополістичної конкуренції з окремими регіональними проявами олігополії, для знаходження приблизного значення індексу, достатньо знати значення часток ринку для найбільших компаній на ньому.

Інноваційна діяльність виражена відповідним індикатором є знаходиться наступним чином:

$$IAI = \frac{IR}{MT}, \tag{2.9}$$

де *IR* – частка доходу від інновацій;

MT – загальний грошова маса компанії.

Соціальна катастрофа, що виникла характеризується раніше наведеним показником *SSD*.

Фінансова стабільність компанії є класичним економічним показником (скорочено *FSI*), який відображає можливості компанії відповідати зобов'язанням у довго- та середньострокові перспективах [33].

Загалом можемо записати наступне формалізоване представлення наведених вище показників діяльності компанії, врахувавши (2.8) та (2.9):

$$MRI = (\frac{s_t^2 \times IAI \times FSI}{HHI})^{SSD}, \tag{2.10}$$

де *N* – кількість компаній на обраному ринку;

s_t – частка ринку, що належить компанії *t*;

FSI – показник фінансової стабільності компанії;

IAI – показник інноваційної діяльності компанії;

HHI – індекс Ґерфіндаля-Ґіршмана;

SSD – показник соціальної катастрофи.

Як і у випадку показника *TAOI* в ході обчислення за формулою (2.10) буде отримане скалярне значення, аби векторизувати його, використаємо результат як центр нормального розподілу.

В отриманому числовому ряді не враховується стан цільової галузі та світової економіки.

Для врахування стану економіки в цілому оберемо ряд класичних макроекономічних показників:

– рівень світового ВВП;

– рівень цін на енергоресурси [34];

– індекс S&P 500.

Вказані дані вже є числовими рядами, тож для того, аби мати змогу використати їх у якості екзогенних змінних, достатньо їх нормалізувати.

Рівень світового валового внутрішнього продукт є показником, що відображає стан економіки світу на поточний момент. В макроекономічній теорії запропонованій Фрідманом при прогнозуванні власної бізнес-діяльності необхідно розглядати ВВП держави в якій ця діяльність здійснюється. Оскільки ринок електронної комерції часто не обмежується однією державою, було вирішено розглядати показник світового ВВП.

Показник рівня на енергоресурси прямо впливає на логістику компаній і відповідно на ціни їх продукції. Оскільки дані щодо продукції чи індексу споживчих цін не обов’язково будуть цільовими при прогнозуванні, необхідно врахувати вплив реакцій компанії на зростання вартості контрактів з контрагентами, для цього і розглядається вказаний показник.

Індекс S&P за своєю сутністю є внутрішньодержавним індексом, однак враховуючи події 2007 року, коли цей показник суттєво впав, та реакцію бізнесу у всьому світі в цілому (через прив’язку його до долара), було вирішено додати його як окрему екзогенну змінну.

Стан галузі визначатимемо за портфелем акцій п’яти найбільших компаній на ринку електронної комерції. За останніми даними це [35]:

– Jingdong Mall (тікер JD);

– Amazon (тікер AMZN);

– Meituan (тікер 3690.HK);

– Alibaba (тікер BABA);

– Pinduoduo (тікер PDD).

Необхідно зазначити, що під портфелем мається на увазі тип Unit, а не ETF, відповідно кожна компанія у цьому цільовому активі матиме однакову частку. Під ціною розглядатимемо показник Adjusted Close – середню вартість ціни акцій при закритті основної торгової сесії за попередні 20 робочих днів.

Як у попередньому випадку, дані вже є числовими рядами, однак їх необхідно нормалізувати для коректної роботи алгоритмів.

Врешті, у якості зовнішніх змінних для сімейства модифікованих моделей EC-VAR розглядатимуться наступні ряди:

– характеристика цільової аудиторії;

– характеристика діяльності компанії на ринку;

– стан світової економіки виражений трьома факторами;

– стан галузі виражений портфелем акцій.

Вказані вище часові ряди з екзогенними змінними необхідно буде змістити на один день назад, вважаючи, що поведінка економічних показників станом на певний цільовий день у більшості випадків є реакцією га на попередній день. Водночас за допомогою цієї операції, як уже зазначалося раніше, буде нівельовано проблему неточності зовнішніх змінних при прогнозуванні.

2.3 Можливості прискорення

За наявних великих обсягів даних, швидкодія алгоритмів сімейства VAR доволі обмежена, хоча виграш у порівнянні з іншими технологіями спостерігається.

Задля вирішення вказаної проблеми можна використовувати принципи паралелізму, однак вони не гарантують значний виграш у швидкості [36].

У якості базової альтернативи цим принципам найчастіше використовують технологію MapReduce, тому було вирішено імплементувати принципи цієї технології задля визначення того який виграш у швидкості вона дає і таким чином перевірити доцільність її використання.

Сутність підходу, який закладений у технологій MapReduce полягає у розподіленні загального набору даних на окремі вузли. Ця процедура виконується за допомогою функцій мапінгу, з подальшим застосуванням обраних алгоритмів та редуктора, який збирає дані зі всіх вузлів та уніфікує їх [37].

Загалом ця технологія знаходить свою реалізацію у декількох фреймворках, однак було вирішено використати MapReduce, що пропонується Hadoop [38].

Архітектуру цього підходу можна зобразити так, як показано на рисунку нижче (див. рис. 2.1).

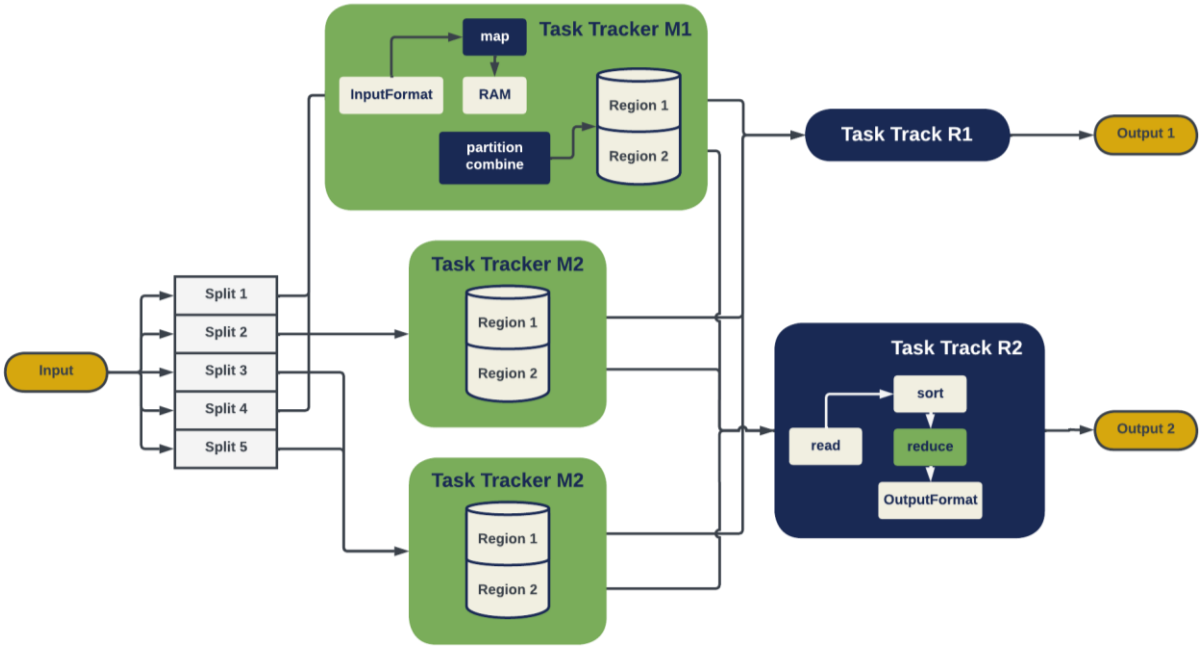


Рисунок 2.1 – Архітектура реалізації MapReduce у Hadoop

Як бачимо у цьому випадку окрім власне функцій мапінгу та редукції наявні функції розподілення та комбінування, вони необхідні для того, аби відповідь з кожного вузла надходила в об’єднаному вигляді.

Також варто звернути увагу, що перед редукцією виконується сортування результатів.

Серед переваг обраного підходу можна виділити:

– масштабованість;

– відносну дешевизну;

– простоту використання (хоча технологія ускладнена необхідністю у налаштувань конфігурацій);

– можливість контролю виконання.

Серед недоліків варто виділити необхідність у створенні великої кількості коду та прихованість обробки.

Аби краще зрозуміти те, яким чином можна імплементувати сімейство алгоритмів VAR за допомогою MepReduce, наведемо лістинг з псевдокодом для функції мапінгу:

```
Input: max window size & record
Output: Addends of predicted value
for (i = 2 to w) do
    count = 0
    diff_i = i*(i+1)/2
    while(i >= count) do
        k′ = j + '.' + (i-(count-1))
        v′ = data8count/diff_i
        context.write(k′,v′)
        count++
    end while
end for #map
```

При цьому редуктор можна подати наступним чином:

```
Input: output of Combiners k′′′, list <v′′>
Output: Optimal window size % minMSE
A shared variable MSE is initialized to 0 for each
window size from (2 to w)
predicted = 0
while (v′′.hasNext()) do
    if (tag = 1) do
     actual = v′′.value
    else
     predicted = predicted + v′′.value
    end if
end while
Split k′′ indo window size & addend number
```

```
w = window size
MSEW = MSEW + (actual-predicted)^2
minMSE = minimum(MSE)
keyFinal = IndexOf(minMSE)
valueFinal = minMSE
output.collect(keyFinal, valueFinal) # reduce
```

Стосовно функцій розподілу та комбінування, які характерні безпосередньо для технології Hadoop, то вони мають доволі схожий вигляд на вищезазначені функції редуктора та мапера, однак оброблюють дані в рамках одного вузла та декількох регіонів пам’яті в цих вузлах.

У свою чергу вказаний підхід дозволяє пришвидшити роботу обраної технології при великих обсяг одноманітних даних.

Зазначене твердження та наявність в алгоритмі передобробки інформації про котирування акцій, зміну індексу S&P 500, цін на енергоресурси, що є показниками, які можна віднести до Big Data і стало підґрунтям для вибору Hadoop реалізації технології MapReduce, з метою отримання найбільшого можливого прискорення роботи.

3 ПЛАН ЕКСПЕРИМЕНТУ

Під планом експерименту в рамках чинної роботи маємо на увазі наступний набір характеристик:

- загальні умови в яких виконується експеримент (враховуючи умови для реалізації MapReduce);
- фактори ефективності – набір характеристик за якими оцінюватимуться моделі;
- принцип порівняння моделей.

Поступово визначимо кожну зазначену частину плану.

3.1 Загальні експериментальні умови

Враховуючи специфіку запропонованого дослідження, було обрано метод контрольованого експерименту.

Базове середовище виконання має наступний набір характеристик для роботи алгоритмів:

- CPU: Intel Core i5-1135G7;
- RAM: 16 Гб;
- VRAM: 4 Гб;
- ОС: Ubuntu 21.04.

Вказані характеристики майже в повному обсязі були продубльовані на віртуальних вузлах, на яких планується здійснюватися процес частково обчислення (було зменшено RAM з 16 до 8). Їх кількість становитиме від 3 до 4.

У якості засобу обчислення часу виконання було обрано бібліотеку datetime для Python 3, що має точність до наносекунди.

Аби базові обчислення не сповільнювали роботу програми вирішено використати бібліотеки numpy та pandas.

Для обробки природніх мов (враховуючи лематизацію, токенізацію та інші необхідні функції) було вирішено обрати python-версію бібліотеки nltk.

3.2 Фактори ефективності

Аби мати змогу порівняти зазначені вище алгоритми авторегресії необхідно визначитися з основними критеріями вибору.

З огляду на те, що розглядається задача прогнозування під час надзвичайних ситуацій, найбільш важливими критеріями є час прогнозування та точність прогнозу.

Загалом було обрано наступний перелік факторів:

– час прогнозування;

– точність прогнозу;

– якість врахування екзогенних показників;

– час на підготовку даних;

– можливість врахування шумів.

Визначившись з критеріями здійснимо опис відповідних шкал оцінювання.

Час прогнозування та підготовки даних вимірюватиметься у мілісекундах за допомогою вказаної вище бібліотеки. Сам показник при цьому не обмежуємо. Заміри здійснюватимемо по 5 разів.

Можливість врахування шумів позначимо за допомогою множини $\{0,1\}$. 0 вказуватиме на те, що подібна можливість не передбачена, а 1 – навпаки.

Точність прогнозу (з англійської accuracy) вимірюватимемо у долях, при цьому можливе значення належатиме наступній множині: $\{x \in [0,1], x \in N\}$. Замір точності здійснюватимемо для 4 різних наборів даних.

Варто зауважити, що оскільки розглядається задача регресії, а не класифікації знаходження точності буде здійснюватися за допомогою середньоквадратичного відхилення, так як наведено у наступній формулі:

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}, \quad (3.1)$$

де N – кількість значень які необхідно спрогнозувати;

y_i – реальне значення показника;

$\widehat{y_i}$ – спрогнозоване значення показника.

Аби перетворити наведений у (3.1) показник у коректне відображення здійснюватимемо нормування показника відносно наступного еталонного значення E:

$$E = \frac{\sum_{i=1}^{N} y_i}{N}, \quad (3.2)$$

де N – кількість значень які необхідно спрогнозувати;

y_i – реальне значення показника.

Варто зауважити, що з точки зору задачі регресії загалом для (3.2) може виникнути ситуації, коли спрогнозоване значення у декілька разів більше за реальне, тоді вказана нормалізація буде некоректною [39]. Однак, виходячи з аналізу досліджень проведених в розрізі прогнозування економічних показників і використанні авторегресії для цього, вказана ситуація є малоймовірною, якщо подібна проблема таки виникне, то вирішуватиметься окремо. Вважатимемо вказане твердження обмеженням чинної роботи.

З огляду на те, що екзогенні змінні можуть бути присутні як до безпосередньо обробки цільових показників, так і у ході їх обробки, маємо наступну шкалу оцінювання для якості врахування екзогенних змінних:

– екзогенні змінні враховуються необмежено як на початку обробки, так і під час обробки – 5 балів;

– екзогенні змінні враховуються необмежено на початку обробки та обмежено під час обробки – 4 бали;
– екзогенні змінні враховуються необмежено на початку обробки і не враховуються під час обробки – 3 бали;
– екзогенні змінні обмежено враховуються під час обробки – 2 бали;
– екзогенні змінні не враховуються – 1 бал.

Аби знизити вплив можливих вимірювання, викликаних проблемами з точністю роботи часових модулів чи оточення, було вирішено проводити по 5 замірів для показників часу та перевірити точність прогнозування на двох вибірках даних з 2015 по 2022 роки компаній, які фактично є одними з найбільших гравців ринку електронної комерції, однак на які ринок впливає по різному. Зокрема, Amazon та Alibaba [40, 41] з лише інтернет рітейлінгом та Walmart і Big Mart з реальними магазинами [42, 43].

Фрагмент вибірки для Amazon наведено на рисунку нижче (див. рис. 3.1):

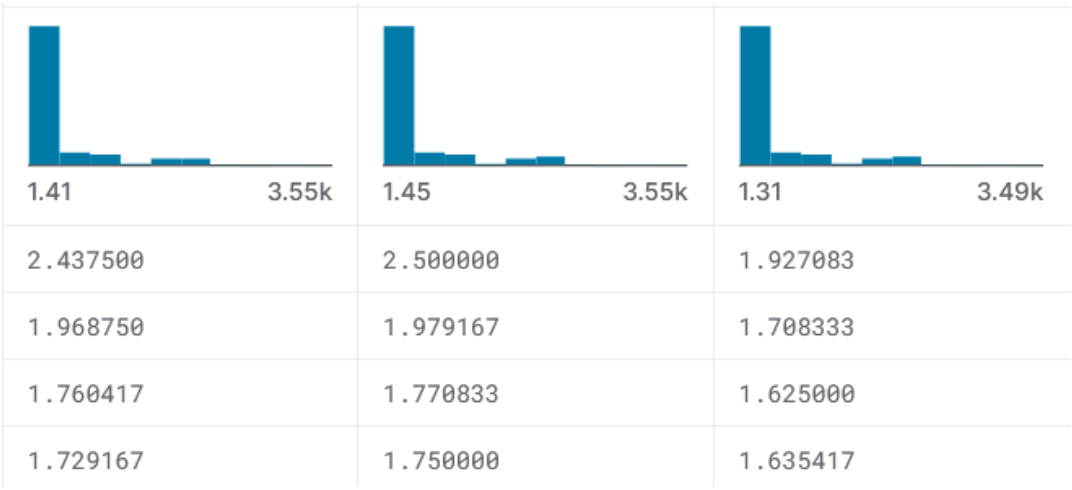


Рисунок 3.1 – Фрагмент набору даних Amazon

Загалом у вказаний набір даних входять наступні поля (наводяться лише ті, що будуть використані при прогнозуванні):

– Date: дата відносно якої наведені дані;
– Open: загальнонаціональна ціна акцій компанії при відкритті основної торговельної сесії;

– High: найвища ціна акцій за вказаний день;
– Low: найнижча ціна акцій за вказаний день;
– Close: загальнонаціональна ціна акцій компанії при закритті основної торговельної сесії;
– Volume: обсяг торгів за день;
– iCBC (Continious Book Clearing Price) – ціна акцій на аукціоні закриття у визначений день. Якщо iCBC рівний 0, то вважається, що аукціон не відбувся;
– Auction Volume – обсяг аукціону у кількостях акцій за певний день;
– isHalt – чи була зупинка торгівлі у визначений день (якщо значення істинне, то ціну закриття необхідно змінювати взявши середню ціну за попередні 20 робочих днів).

Фрагмент вибірки для Alibaba наведений на рисунку нижче (див. рис. 3.2), схожий на попередній, однак сама компанія функціонує на іншому національному ринку.

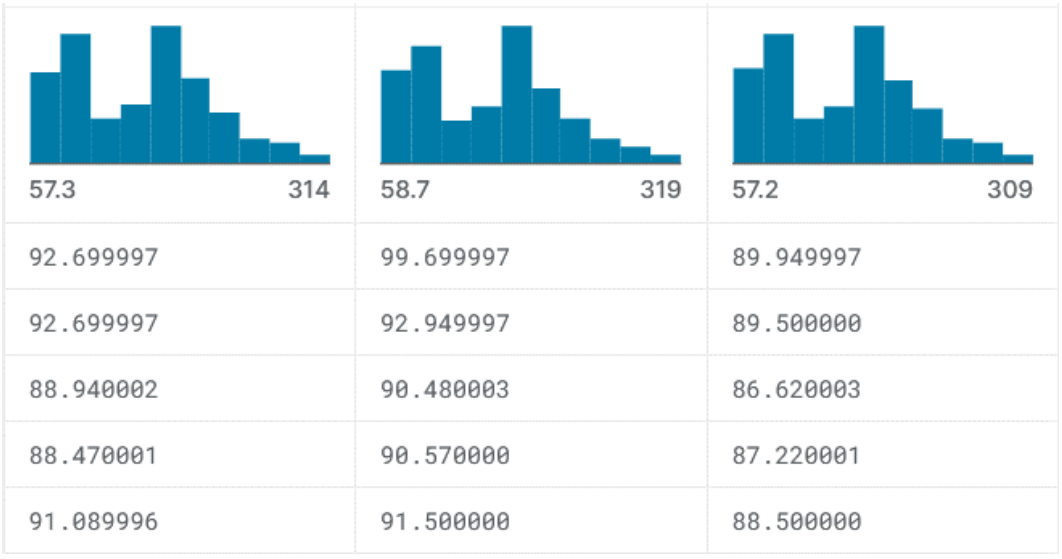


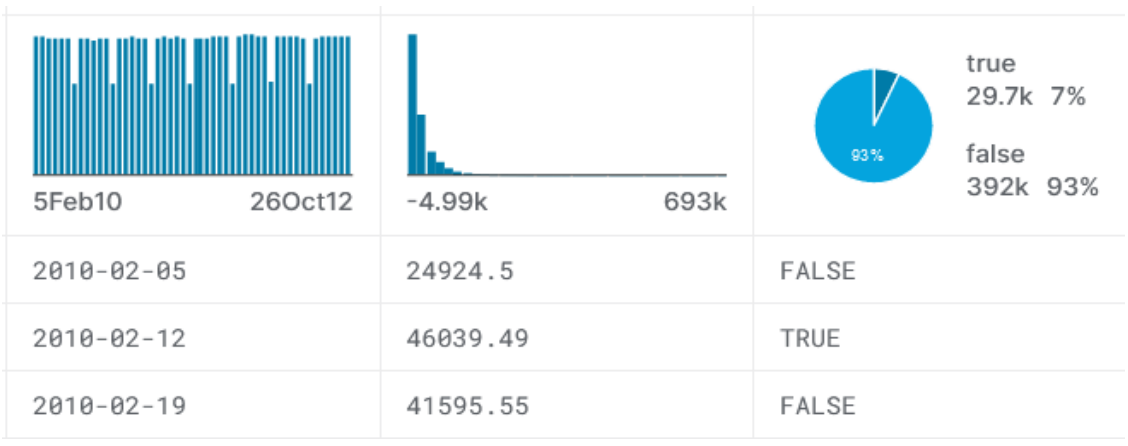
Рисунок 3.2 – Фрагмент набору даних Alibaba

Загалом у вказаний набір даних входять наступні поля (наводяться лише ті, що будуть використані при прогнозуванні):

– Date: дата відносно якої наведені дані;

– Open: загальнонаціональна ціна акцій компанії при відкритті основної торговельної сесії;
– Limit Up: обмеження згори динаміки цін на акції на останній момент основної торговельної сесії;
– Limit Down: обмеження знизу динаміки цін на акції на останній момент основної торговельної сесії;
– Close: загальнонаціональна ціна акцій компанії при закритті основної торговельної сесії;
– Volume: обсяг торгів за день;
– NBB (National Best Bid) – найкраща ціна купівлі цінного паперу на останній момент основної торговельної сесії;
– NBO (National Best Offer) – найкраща ціна продажу цінного паперу на останній момент основної торговельної сесії.

Вибірка для Walmart містить інформацію з 2010 року, фрагмент вибірки наведено нижче (див. рис. 3.3):



2010-02-05	24924.5	FALSE
2010-02-12	46039.49	TRUE
2010-02-19	41595.55	FALSE

Рисунок 3.3 – Фрагмент набору даних Walmart

У вказаний набір даних входять наступні поля (наводяться лише ті, що будуть використані при прогнозуванні):

– Store: номер магазину відносно якого неведені дані (якщо значення рівне 0, то мається на увазі онлайн магазин);
– Date: дата відносно якої наведені дані;

– Fuel Price: ціна пального необхідного для доставки товарів у вказаний магазин за певний день;
– Consumer Price Index: індекс споживчих цін на певний день (параметр є спільним для всіх магазинів);
– Volume: загальна кількість проданих товарів у магазині за певний день;
– Customers: загальна кількість клієнтів у магазині за певний день;
– Revenue: загальний рівень доходу магазину за певний день;
– Profit: загальний рівень прибутку магазину за певний день.

Як бачимо, характер даних відрізнятиметься: для Amazon та Alibaba дані стосуватимуться біржової інформації, а для Walmart інформації про рівні продажу. Стосовно даних Big Mart, вони стосуються рівня продажів на окремі товари групи товарів, а не інформація за магазинами та регіонами діяльності відділень компанії.

Фрагмент вибірки для цієї компанії наведено нижче (див. рис. 3.4):

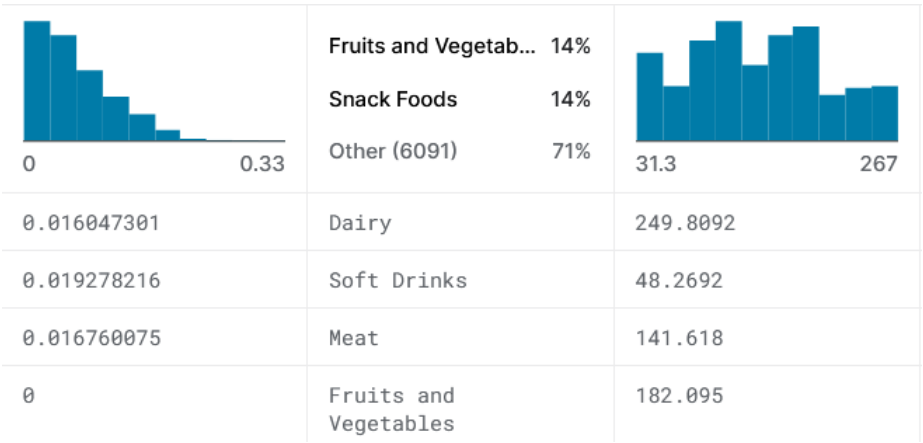


Рисунок 3.4 – Фрагмент набору даних Big Mart

У вказаний набір даних входять наступні поля (наводяться лише ті, що будуть використані при прогнозуванні):

– ProductID: ідентифікатор певного продукту, який одночасно продається як в реальних так і онлайн-магазинах;
– Date: дата відносно якої наведені дані;

- Visibility: наскільки продукт привертає увагу споживачів (для онлайн-магазину вирахований черезметрику сайту, а для реальних магазинів через кількість звернень до продавцв консультантів та обсяг продажів);
- Volume: загальний обсяг проданого товару, за певний день;
- Customers: загальна кількість покупців певного товару за певний день
- Revenue: загальний рівень доходу за певний день від продажів певного товару;
- Profit: загальний рівень прибутку за певний день від продажів певного товару.

Як бачимо чотири обраних датасети не тільки покривають різні національні ринки, а і різні за сутністю дані. Зазначений полморфізм у даниз дозволить краще визначити наскільки авторегресійні алгоритми в цілому показують точний результат.

3.3 Принцип порівняння моделей

Аби визначити яка з моделей є найбільш ефективною за наведеними вище критеріями було вирішено застосувати принцип лінійної адитивної згортки з ваговими коефіцієнтами. Значення згортки визначатиметься наступним чином:

$$L(A_i) = \sum_{j=0}^{n} \omega_{ij} \times m_{ij}, \tag{3.3}$$

де A_i – i-й авторегресійний алгоритм;

n – загальна кількість метрик для порівняння;

m_{ij} – значення j-ої метрики для i-го авторегресійного алгоритму;

ω_{ij} – значення вагового коефіцієнту j-ої метрики для i-го авторегресійного алгоритму.

У формулі (3.3) ваги визначаються за важливістю кожної з метрик.

При прогнозуванні показників економічної діяльності найбільш важливим є точність. На другому місці – врахування зовнішніх показників, на третьому – час роботи алгоритму.

Таким чином можемо призначити можемо призначити кожному з критеріїв певну кількість балів, що дозволить вирахувати ваговий коефіцієнт, визначений у формулі (3.3):

- для точності 16 очок;
- для можливості врахування екзогенних змінних 8 очок;
- для економії часу прогнозування та часу підготовки по 4 очки;
- для врахування шумів 2 очка.

Виходячи з призначених балів, можемо визначити вагові коефіцієнти ω_{ij} для кожної з метрик:

- для точності: 16/30=8/15;
- для можливості врахування екзогенних змінних: 8/30=4/15;
- для економії часу прогнозування та часу підготовки: 4/30=2/15;
- для врахування шумів: 2/30=1/15.

4 ІМПЛЕМЕНТАЦІЯ МОДЕЛЕЙ

Почнемо процес імплементації зі створення алгоритму для передобробки даних, зокрема для проведення контент-аналізу текстової інформації. Як зазначалося вище, для обробки природної мови було використано бібліотеку nltk. Нижче наведено лістинг коду, що здійснює створення словника без дублікатів:

```
def process_text(text: str, language: str) -> str:
      stemmer = PorterStemmer()
      lemmatizer = WordNetLemmatizer()
      stop_words = set(stopwords.words(
            get_language_name(language)
      ))
      sentences = nltk.sent_tokenize(text)
      sentences_processed = []
      for sentence in sentences:
            words_processed = []
            words = nltk.word_tokenize(sentence)
            words = [word for word in words if word not in stop_words]
            for word in words:
                  word = stemmer.stem(word)
                  word = lemmatizer.lemmatize(word)
                  words_processed.append(word)
            words = [word for word in words_processed \
                                    if word not in stop_words]
            sentences_processed.append(" ".join(words))
      text_processed = " ".join(sentences_processed)
      return text_processed
```

У якості базису для стемінгу було використано стеммер Портера, для лематизації – лематизатор WordNET, а для видалення слів, що мають низьке лінгвістичне забарвлення – модуль stopwords. Варто зауважити, що зазначений модуль не підтримує усіх мов, однак, враховуючи те, що Amazon та Walmart є американськими компаніями, це не є проблемою.

Задля знаходження TF-IDF характеристики було вирішено використати модулі бібліотеки sklearn, а характеристику полярності знайти за допомогою модуля vader з nltk.

Функції нормування екзогенних та ендогенних змінних, очищення цільового набору даних від порожніх значень та селекція необхідних для аналізу даних здійснювалась за допомогою pandas.

Портфель акцій та S&P індекс було отримано з використанням Yahoo Finance, а дані про рівень світового ВВП та рівень цін на енергоресурси, отримані з бази даних Світового банку та trading economics.

Нижче наведемо фрагмент коду, що працює з API Yahoo Finance для отримання SPY – портфеля акцій (яким можна торгувати), що прив'язаний до індексу S&P 500:

```
def get_spy() -> pd.DataFrame:
      today = date.today()
      d1 = today.strftime("%Y-%m-%d")
      end_date = d1
      d2 = date.today() - timedelta(days=365)
      d1 = d2.strftime("%Y-%m-%d")
      start_date = d2

      data = yf.download(
        'AMZN',
        start=start_date,
        end=end_date,
        progress=False
      )

      data["Date"] = data.index
      data = data[["Date", "Open", "High", "Low", "Close", "Adj Close",
      "Volume"]]
      data.reset_index(drop=True, inplace=True)

      return data
```

Наступним кроком є написання обраних векторних авторегресійних моделей.

Задля уникнення помилок в реалізації алгоритмів, було вирішено використати бібліотеку statsmodel з можливістю налаштування гіперпараметрів авторегресії. Приклад використання для моделі EC-VARMA наведено нижче:

```
def make_varmax_forecast(values_low, values_high, danger) -> tuple[str]:
    data = list()
    for i, val in enumerate(values_low):
        row = [val, values_high[i]]
        data.append(row)
    model = VARMAX(data, order=(1, 1))
    model_fit = model.fit(disp=False)
    danger_exogenous = []
    for i in range(constants.STEPS):
        danger_exogenous.append([danger])
```

```
    forecast = model_fit.forecast(
           exog=danger_exogenous,
           steps=constants.STEPS
    )
    for item in forecast:
        values_low.append(item[0])
        values_high.append(item[1])
    return values_low, values_high
```

Задля реалізації MapReduce-підходу на основі Hadoop необхідно створити файли з редуктором та мапінгом, псевдокод для яких було наведено у підрозділі 3.3 чинного звіте та модифікувати файли-конфігурацій, вказавши відповідні шляхи та сервер для розгортання. Приклад видозміни конфігурацій наведено нижче:

```
<property>
    <name>yarn.nodemanager.local-dirs</name>
    <value>/var/lib/hadoop-yarn/cache/${user.name}/nm-local-dir</value>
</property>
```

Вище вказана властивість YARN Mode, що визначає шлях для зберігання файлів у локальній файловій системі.

5 АНАЛІЗ РЕЗУЛЬТАТІВ

5.1 Розгляд доцільності використання алгоритму передобробки

Для початку наведемо наступний рисунок (див. рис. 5.1) як підтвердження того, що ЕС-модифікація алгоритмів шляхом додавання обробки наведених раніше даних, була необхідна. Для цього використаємо набір даних про динаміку ціни закриття на американській біржі торговельної компанії Target, отриманої за допомогою Alpha Vantage.

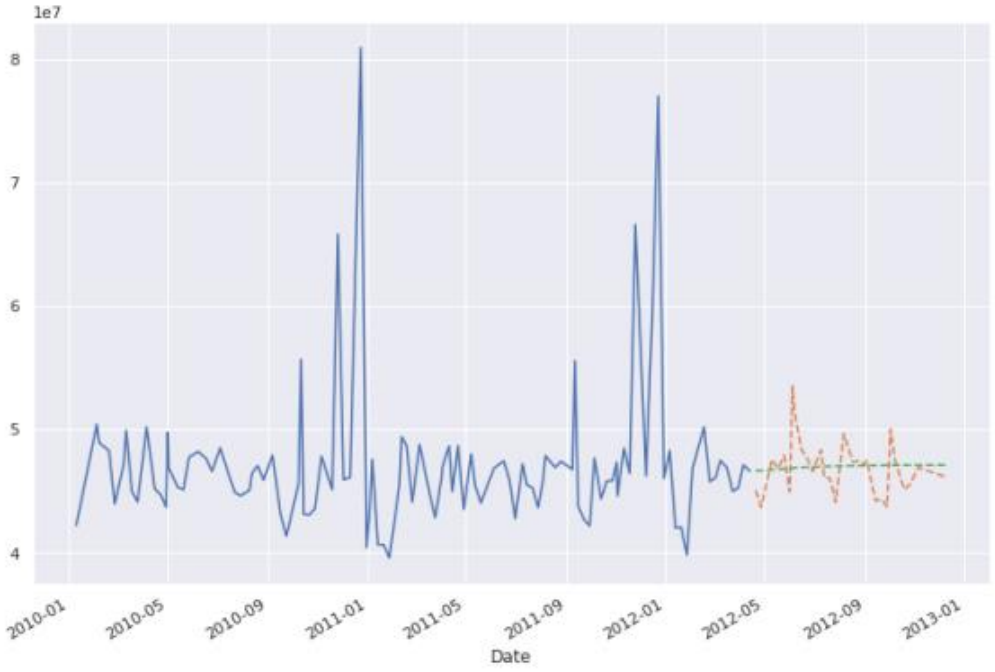


Рисунок 5.1 – Динаміка результатів прогнозування для алгоритму без передобробки та без врахування екзогенних змінних

Зелена лінія на рисунку вище вказує прогнозоване значення для алгоритму авторегресії інтегрованого рухомого середнього без екзогенних змінних. Як бачимо, прогноз виглядає як пряма лінія та по суті намагається описати загальний тренд, однак похибка відносно загального тренд складає близько 20% (зазначена інформація не наведена на графіку).

На рисунку нижче (див. рис. 5.2), зелена лінія, яка теж описує прогнозоване значення показує більш точний прогноз

Рисунок 5.2 – Динаміка результатів прогнозування для алгоритму з передобробкою та врахування екзогенних змінних

Загалом можна зазначити, що модифікована імплементація алгоритму векторної авторегресії з інтегрованим ковзаючим середнім вказує на те, що дані у звичному алгоритмі не враховують коливання в цілому, натомість покращений алгоритм цю проблему пом’якшує.

5.2 Аналіз результатів експерименту

Тепер можемо перейти до отриманих результатів.

Почнемо показника часу прогнозування цільових змінних, інформація щодо яких наведена у таблиці нижче (див. табл. 5.1).

Врахуємо наступні позначення введені для кращого сприйняття:

– R: класична авторегресія,

– RS: сезонна авторегресія,

– RL: авторегресія розподіленого лагу,

– RMA: авторегресія ковзаючого середнього,

– RIMA: авторегресія інтегрованого ковзаючого середнього.

Таблиця 5.1 – Час прогнозування (у мілісекундах)

Послідовна версія					Версія MapReduce				
R	RS	RL	RMA	RIMA	R	RS	RL	RMA	RIMA
89	94	103	127	139	31	33	36	45	49
143	169	197	214	227	50	59	69	75	80
109	129	143	159	179	38	45	50	56	63
167	253	273	299	321	59	89	96	105	113
183	229	245	276	294	64	80	86	97	103

Знайдемо середнє значення для кожного випадку при послідовній версій:

– EC-VAR: 0.138 с;

– EC-VARS: 0.175 с;

– EC-VARL: 0.192 с;

– EC-VARMA: 0.215 с;

– EC-VARIMA: 0.232 с.

Як бачимо, в середньому алгоритми рухомого середнього та інтегрованого рухомого середнього є значно повільнішими. Це пояснюється тим, що вони враховують екзогенні змінні у повному обсязі і при цьому здійснюють врегулювання шумів.

Якщо ж порівнювати послідовне виконання з MapReduce версіями, то виграш у швидкості становить ~ 2.85 для кожної з моделі. При цьому, якщо покращити конфігурацію для MapReduce збільшивши кількість вузлів до 4, то виграш складе ~ 3.74. Якщо ж змінити кількість оперативної пам’яті на трьох вузлах, прискорення теж збільшиться, хоча і не настільки суттєво до ~ 2.98

Перейдемо до результатів замірів часу підготовки.

Варто зауважити, що час підготовки для паралелізованої версії та звичайної однаковий, оскільки цей процес було вирішено не розподіляти між вузлами (див. табл. 5.2).

Таблиця 5.2 – Час підготовки (у секундах)

R	RS	RL	RMA	RIMA
7.6	9.3	9.2	18.7	18.9
6.9	8.7	8.8	17.6	17.7
7.3	9.2	9.1	17.7	17.5
7.9	9.0	8.9	18.3	18.1
6.6	8.5	8.5	18.0	18.4

Маємо наступні середні значення показників:

– EC-VAR: 7.3 с;

– EC-VARS: 8.9 с;

– EC-VARL: 8.9 с;

– EC-VARMA: 18.1 с;

– EC-VARIMA: 18.1 с.

Різні значення часу підготовки пояснюються можливістю врахування екзогенних змінних.

Не всі вказані раніше часові ряди необхідно подавати як базове значення для створення екземпляру класу-алгоритму, потрібні тільки ті, що є детермінованими на певному часовому проміжку, інші зовнішні змінні подаються під час тренування моделі.

Як бачимо алгоритми, що враховують зовнішній вплив у повному обсязі, загалом працюють довше, оскільки для них необхідно виконати більшу кількість обчислень.

Перейдемо до результатів точності прогнозування.

Задля спрощення викладок у таблиці нижче (див. табл. 5.3) наведені результати для кожної з вибірок.

Таблиця 5.3 – Точність прогнозування

Набір	R	RS	RL	RMA	RIMA
Walmart	0.85	0.90	0.87	0.93	0.94
Amazon	0.83	0.89	0.91	0.97	0.96
Big Mart	0.84	0.91	0.90	0.96	0.96
Alibaba	0.84	0.90	0.88	0.94	0.94

Варто зауважити, що вказані значення вже є нормованими за правилом наведеним раніше, а також необхідно зазначити, що описане вище проблема з некоректним визначенням еталонного значення не була виявлена на обраних наборах даних.

Як видно з таблиці алгоритми, що враховують усі можливі зовнішні змінні працюють точніше для кожної з вибірок. При цьому напевно сказати яка саме з моделей EC-VARMA чи EC-VARIMA точніше, з огляду на результати, не можна.

Тепер можемо навести значення критеріїв для кожної альтернативи (див. табл. 5.4). Усі параметри часу наведені в секундах.

Таблиця 5.4 – Значення критеріїв для кожної альтернативи

Модель	Час прогнозування	Точність	Зовнішні змінні	Час підготовки	Шуми
R	0.138	0.84	2	7.3	0
RS	0.175	0.90	3	8.9	0
RL	0.192	0.89	3	8.9	0
RMA	0.215	0.95	5	18.1	1
RIMA	0.232	0.95	5	18.1	1

Перед тим як перейти до застосування вагових коефіцієнтів приведемо отримані дані до одного принципу оптимізації (див. табл. 5.5). Для цього необхідно замінити показники часу на показники економії часу.

Таблиця 5.5 – Значення критеріїв для кожної альтернативи приведені до одного принципу оптимізації

Модель	Економія часу прогнозування	Точність	Зовнішні змінні	Економія часу підготовки	Шуми
R	0.094	0.84	2	10.8	0
RS	0.057	0.90	3	9.2	0
RL	0.040	0.89	3	9.2	0
RMA	0.017	0.95	5	0.0	1
RIMA	0.000	0.95	5	0.0	1

Виходячи з отриманих результатів, визначимо множину Парето (див. табл. 5.6). Для цього достатньо викреслити альтернативи, які за всіма критеріями «гірші» за будь-яку іншу з альтернатив. У цьому випадку подібними альтернативами є авторегресія розподіленого лагу та авторегресія інтегрованого ковзаючого середнього.

Таблиця 5.6 – Множина Парето

Модель	Економія часу прогнозування	Точність	Зовнішні змінні	Економія часу підготовки	Шуми
R	0.094	0.84	2	10.8	0
RS	0.057	0.90	3	9.2	0
RMA	0.017	0.95	5	0.0	1

Перейдемо до нормування критеріїв без еталонів.

З огляду на дані, найкращим значенням економії часу прогнозування є 0.094 с, найгіршим – 0.017 с. Таким чином отримуємо:

– EC-VAR: 1;

– EC-VARS: 0.52.

– EC-VARS: 0.

Найкращим значенням економії часу підготовки є 10.8 с, найгіршим – 0 с. Таким чином отримуємо:

– EC-VAR: 1;

– EC-VARS: 0.85.

– EC-VARS: 0.

Для показників точності та можливості врахування екзогенних змінних є еталони 100 і 5 відповідно.

Показник врахування шумів вже є нормованим, оскільки приймає значення або 0 або 1. Таким чином маємо наступні результати (див. табл. 5.7).

Таблиця 5.7 – Множина Парето з нормованими значеннями

Модель	Економія часу прогнозування	Точність	Зовнішні змінні	Економія часу підготовки	Шуми
R	1.00	0.84	0.40	1.00	0
RS	0.52	0.90	0.60	0.85	0
RMA	0.00	0.95	1.00	0.00	1

Нормувавши всі значення для кожної альтернативи та взявши до уваги вагові коефіцієнти можемо знайти значення згортки, враховуючи лише оптимальні за Парето альтернативи:

– для EC-VAR отримуємо: 0.82;

– для EC-VARS отримуємо: 0.82;

– для EC-VARMA отримуємо: 0.84.

Виходячи з результатів, при прогнозуванні показників ринкової активності для компаній на ринку економічної комерції, необхідно обрати модель EC-VARMA, що є логічним, оскільки точність моделі суттєво вища за інші, при цьому програш у економії часу не є впливовим.

Окрім цього варто зауважити, що вплив часу можна зменшити за допомогою MapReduce у 2.85 рази при 3 вузлах і 3.74 при 4 вузлах. Таким чином можемо констатувати високу ефективність використання в рамках галузі

дослідження модифікованої паралелізованої авторегресійної моделі рухомого середнього з екзогенними регресорами.

5.3 Подальше дослідження

Для подальшого дослідження обраної тематики роботи можна сформувати наступний набір ідей:

- розширення спектру використаних сімейств алгоритмів, зокрема розгляд ефективності використання нейронних мереж: згорткових, рекурентних з довгостроковою пам’яттю, їх поєднання тощо;
- збільшення кількості метрик для оцінки ефективності моделі, шляхом проведення більш масштабного експертного оцінювання;
- визначення способів отримання та формування екзогенних змінних для майбутнього та проведення корекції результатів точності, врахувавши достовірність цих значень;
- розширення набору цільових факторів для прогнозування та вихід за межі ринку електронної комерції, зокрема на ті сфери, що є відносно стабільними, навіть під час соціальних катастроф та ті, що навпаки є вразливими для будь-яких ринкових коливань;
- дослідження можливості побудови активної системи підтримки прийняття рішень, наприклад, за допомогою поєднання найбільш ефективного (виходячи із наведених раніше метрик) прогностичного алгоритму зі штучним інтелектом, який зможе на основі отриманого прогнозу зробити висновки, корисні для користувача такої системи.

ВИСНОВКИ

У ході виконання чинної роботи було визначено проблему у важкості прогнозування економічних показників під час надзвичайних ситуацій, що викликають соціальні катастрофи.

Після проведення аналізу комерційних та наукових рішень було з’ясована, що загалом прийнято виділяти три способи прогнозування:

– за допомогою авторергресії;

– за допомогою нейронних мереж;

– за допомогою байєсового підходу.

Визначено, що оскільки найважливішим є точність та швидкість прогнозу, авторегреісійні моделі є найбільш ефективними з-поміж названих. Окрім цього встановлено, що серед них найчастіше вживаними є:

– модель класичної векторної авторегресії;

– модель векторної сезонної авторегресії;

– модель векторної авторегресії розподіленого лагу;

– модель векторної авторегресії рухомого середнього;

– модель векторної авторегресії інтегрованого рухомого середнього.

При цьому кожна модель розглядається у модифікації з корегування помилок, яка передбачає використання екзогенних регресорів.

Задля кращого розуміння сутності моделей було проведено теоретичне ознайомлення з ними та визначено, що соціальну катастрофу можна розглядати як зовнішню змінну.

Отриманий висновок дозволив сформувати послідовність кроків перетворення соціальної катастрофи на кількісний показник за допомогою використання принципів контент-аналізу. При цьому визначення характеристик явища, які необхідно врахувати, було покладено на результати експертної оцінки соціологів, ризик-менеджерів та спеціалістів з ринку електронної комерції декількох міст України.

Було встановлено необхідність використання не лише соціальної катастрофи як такої, а і профілю цільової аудиторії і мікроекономічного профілю ринкової кон’юнктури, задля врахування впливу надзвичайної ситуації на всіх суб’єктів ринку: домашні господарства, фірми, держава.

Як результат зазначених дій у якості зовнішніх змінних для сімейства модифікованих моделей векторної авторегресії було вирішено розглянути наступні ряди:

- характеристика цільової аудиторії;
- характеристика діяльності компанії на ринку;
- стан світової економіки виражений трьома факторами;
- стан галузі виражений портфелем акцій 5 найбільших компаній на ринку електронної комерції.

Окрім зазначеного, було окреслену проблеми, які можуть виникнути при застосуванні обраних моделей на великих обсягах даних, зокрема це значний час виконання. Задля пом’якшення цієї проблеми запропоновано використовувати технологію MapReduce на основі Hadoop. Аби покращити розуміння цієї технології було наведено алгоритми для двох ключових функцій – мапінгу та редукції.

У ході проведення експерименту виявлено, що алгоритм рухомого середнього (як класичного, так і інтегрованого) дає найточніший результат, однак має найбільші вимоги за часом виконання та обробки. Однак застосування вагових коефіцієнтів показало, що загальна ефективність цих моделей вища за інших. Окрім цього з’ясовано, що виграш у економії часу прогнозування при застосуванні технології MapReduce може досягати 3.74.

Виходячи з вищезазначеного, можемо стверджувати виконання поставленого перед чинною кваліфікаційною роботою завдання –визначення ефективності використання модифікованих авторегресійних моделей для прогнозування даних ринку електронної комерції під час соціальних катастроф.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Impact of the COVID-19 pandemic on trade and development. New York : United Nations Publications, 2020. 113 p.

2. Tollefson J. What the war in Ukraine means. Nature. URL: https://www.nature.com/articles/d41586-022-00969-9 (дата звернення: 30.04.2023).

3. Schröter R., Jovanovic A., Renn O. Social Unrest: A Systemic Risk Perspective. GRF Davos Planet@Risk. 2014. Т. 4, № 2. P. 125–134.

4. Cambridge Centre for Risk Studies. Social Unrest: Stress Test Scenario. Cambridge : Cambridge Press, 2014. 53 p.

5. Burton C. G., Rufat S., Tate E. Social vulnerability: conceptual foundations and geospatial modeling. Cambridge University Press. 2018. P. 53–81.

6. Osuszek L., Ledzianowski J. Decision support and risk management in business context. Journal of Decision Systems. 2020. P. 1–12.

7. Building of Regression Models for Cryptocurrency Price Prediction / K. Smelyakov et al. Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022). Volume I: Main Conference, Gliwice, 12–13 May 2022. 2022. P. 1216–1232.

8. Autoregressive Integrated Moving Average (ARIMA) Model of Forecast Demand in Distribution Centre / I. Rizkya et al. IOP Conference Series: Materials Science and Engineering. 2019. Vol. 598. P. 012071.

9. E-commerce Sales Forecast Based on Ensemble Learning / J. Li et al. IEEE International Symposium on Product Compliance Engineering-Asia. 2020.

10. Chang W.-L., Yuan S.-T. A synthesized model of markov chain and erg theory for behavior forecast in collaborative prototyping. Journal of Information Technology Theory and Application. 2008. Vol. 9, no. 2. P. 45–63.

11. Shim K. MapReduce algorithms for big data analysis. Proceedings of the VLDB Endowment. 2012. Т. 5, № 12. С. 2016–2017.

12. Automated risk management system / G. Henderson et al Ottawa : DRDC Ottawa TR, 2012. 142 p.

13. Arcaya M., Raker E. J., Waters M. C. The Social Consequences of Disasters. Annual Review of Sociology. 2020. Vol. 46, no. 1. P. 671–691.

14. Srinivas K. Process of Risk Management. IntechOpen, 2019. 210 c. URL: https://doi.org/10.5772/intechopen.80804 (дата звернення: 29.04.2023).

15. Tereshchenko G., Gruzdo I. Overview and Analysis of Existing Decisions of Determining the Meaning of Text Documents. 2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T), Kharkiv, 9–12 October 2018. 2018. P. 645–653.

16. Best IT risk management software. G2. URL: https://www.g2.com/categories/it-risk-management (дата звернення: 29.04.2023).

17. Engström F., Rojas D. N. Prediction of the future trend of e-commerce. Stockholm : KTH, 2021. 17 p.

18. Faehnle A., Guidolin M. Dynamic Pricing Recognition on E-Commerce Platforms with VAR Processes. Forecasting. 2021. T. 3. P. 166–180.

19. Home Page. Hyperproof. Automate Compliance & Risk Workflows. URL: https://hyperproof.io/ (дата звернення: 30.04.2023).

20. Home Page. Soterion. GRC Solutions for SAP. URL: https://soterion.com/ (дата звернення: 30.04.2023).

21. Home Page. Whistic.The best way to assess, publish, and share vendor security information. URL: https://www.whistic.com/ (дата звернення: 30.04.2023).

22. Akkaya M. Vector Autoregressive Model and Analysis. Handbook of Reseach on Emerging Theories, MAFE. 2021. P. 197–214.

23. Ding J., Tarokh V., Yang Y. Bridging AIC and BIC: a new criterion for autoregression. Transactions on Information Theory. 2018. Vol. 64, № 6. P. 4024–4043.

24. Jungwirth D., Weninger C. A., Haluza D. Fitness and the Crisis: Impacts of COVID-19 on Active Living and Life Satisfaction in Austria. International Journal of Environmental Research and Public Health. 2021. Vol. 18.

25. Generalized Semantic Analysis Algorithm of Natural Language Texts for Various Functional Style Types / N. Sharonova et al. Proceedings of the 6th

International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022), Gliwice, 12–13 May 2022. 2022. P. 16–26.

26. Effectiveness of Preprocessing Algorithms for Natural Language Processing Applications / K. Smelyakov et al. 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T), Kharkiv, 6–9 October 2020. 2020. P. 187–191.

27. Divya K., Siddhartha B. S. An Interpretation of Lemmatization and Stemming in Natural Language Processing. Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology. 2021. Vol. 22, № 10. P. 350–357.

28. Qaiser S., Ali R. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. IJCA. 2018. Vol. 181, № 1. P. 25–29.

29. Zollman K. J. S. Social structure and the effects of conformity. Synthese. 2010. Vol. 172, № 3. P. 317–340.

30. Investigation of the deep learning approaches to classify emotions in texts / D. Nazarenko et al. Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021). Volume I: Main Conference, Lviv, 22–23 April 2021. 2021. P. 206–224.

31. Karimova H. The Emotion Wheel: How to Use It. PositivePsychology. URL: https://positivepsychology.com/emotion-wheel/ (дата звернення: 30.04.2023).

32. Matsumoto A., Szidarovszky F., Merlone U. Some notes on applying the Herfindahl-Hirschman Index. Applied Economics Letters. 2011. P. 181–184.

33. Wang R., Tan J., Yao S. Are natural resources a blessing or a curse for economic development? The importance of energy innovations. Resources Policy. 2021. Vol. 72. P. 102042.

34. Nelson W. R., Perli R. Selected Indicators of Financial Stability. Division of Monetary Affairs, 2020. 30 p.

35. Largest e-commerce companies by market cap. Global ranking. URL: https://companiesmarketcap.com/e-commerce/largest-e-commerce-companies-by-market-cap/ (дата звернення: 30.04.2023).

36. Sinha A., Jana P. K. MRF: MapReduce based Forecasting Algorithm for Time Series Data. Procedia Computer Science. 2018. Vol. 132. P. 92–102.

37. Lin J., Dyer C. Data-Intensive Text Processing with MapReduce. College Park : University of Maryland, 2010. 175 p.

38. MapReduce Tutorial: Apache Hadoop. The Apache Software Foundation. 2008. 42 p. URL: https://hadoop.apache.org (дата звернення: 28.04.2023).

39. Brownlee J. Regression Metrics for Machine Learning. Machine Larning Mastery. URL: https://machinelearningmastery.com/regression-metrics-for-machine-learning (дата звернення: 30.04.2023).

40. Roveda J. H. Historical Amazon stock prices. . Kaggle: Your Data Science Community. URL: https://www.kaggle.com/datasets/josehenriqueroveda/historical-amazon-stock-prices (дата звернення: 30.04.2023).

41. Verma A. Alibaba Stock Data. Kaggle: Your Machine Learning and Data Science Community. URL: https://www.kaggle.com/datasets/varpit94/alibaba-stock-data (дата звернення: 30.04.2023).

42. Ahmedov A. Walmart Sales Forecast. Your Machine Learning and Data Science Community. URL: https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast (дата звернення: 30.04.2023).

43. Kuila A. Big Mart Sales. Kaggle: Your Machine Learning and Data Science Community. URL: https://www.kaggle.com/datasets/akashdeepkuila/big-mart-sales (дата звернення: 30.04.2023).

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ ЗА НАУКОВИМИ НАПРЯМАМИ КЕРІВНИКА ТА НАУКОВЦІВ КАФЕДРИ ПРОГРАМНОЇ ІНЖЕНЕРІЇ

7. Building of Regression Models for Cryptocurrency Price Prediction / K. Smelyakov et al. Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022). Volume I: Main Conference, Gliwice, 12–13 May 2022. 2022. P. 1216–1232.

15. Tereshchenko G., Gruzdo I. Overview and Analysis of Existing Decisions of Determining the Meaning of Text Documents. 2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T), Kharkiv, 9–12 October 2018. 2018. P. 645–653.

22. Generalized Semantic Analysis Algorithm of Natural Language Texts for Various Functional Style Types / N. Sharonova et al. Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022), Gliwice, 12–13 May 2022. 2022. P. 16–26.

23. Effectiveness of Preprocessing Algorithms for Natural Language Processing Applications / K. Smelyakov et al. 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T), Kharkiv, 6–9 October 2020. 2020. P. 187–191.

27. Investigation of the deep learning approaches to classify emotions in texts / D. Nazarenko et al. Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021). Volume I: Main Conference, Lviv, 22–23 April 2021. 2021. P. 206–224.