Міністерство освіти і науки України
Сумський державний університет
Навчально-науковий інститут бізнесу, економіки та менеджменту
Кафедра економічної кібернетики

КВАЛІФІКАЦІЙНА МАГІСТЕРСЬКА РОБОТА

на тему «Моделювання та прогнозування трендів кібератак»

Виконала студентка 2 курсу, групи ЕК.м-11
Спеціальності 051 «Економіка»
(«Економічна кібернетика»)
Кобзенко Вікторія Вікторівна
Керівник доктор економічних наук, доцент
Яровенко Ганна Миколаївна

Суми – 2022 рік

РЕФЕРАТ

кваліфікаційної магістерської роботи на тему

«МОДЕЛЮВАННЯ СКОРИНГОВОЇ ОЦІНКИ КРЕДИТОСПРОМОЖНОСТІ ПОЗИЧАЛЬНИКІВ БАНКУ»

студентки Кобзенко Вікторії Вікторівни
(прізвище, ім'я, по батькові)

Актуальність теми, обраної для дослідження, визначається тим, що в реаліях сьогодення суспільство стикається кожен день з кібератаками, які можуть згубно подіяти на важливі сфери діяльності. Тому саме моделювання та прогнозування кібератак зможуть запобігти та захистити дані.

Мета кваліфікаційної магістерської роботи полягає у розробці математичних моделей та прогнозуванні трендів кібератак на основі панельних даних джерел, які звітують про кількість кібератак.

Об'єктом дослідження є кількість кібератак, а саме потоку даних зі шкідливими програмами у поштових додатках, підозрілий поштовий трафік та потік даних з виявлених мережевих атак.

Предметом дослідження є методи і моделі дослідження та прогнозування панельних даних.

Задачами дослідження є:

1) розкриття сутності об'єкту дослідження - кібератак та проаналізувати сучасні тренди кібербезпекової політики;

2) розробка концептуальної моделі моделювання та прогнозування трендів кібератак;

3) описання вхідних даних та проведення статистичного аналізу та візуалізацію даних;

4) побудова математичних моделей та інтерпретація отриманих результатів;

5) перевірка моделей на адекватність;

6) побудова прогнозних моделей та вибір релевантної.

Для досягнення поставленої мети та задач дослідження були використані такі методи дослідження: аналіз і синтез, дедукція, абстрагування, конкретизація, аргументація, порівняння, класифікація та метод узагальнення, за допомогою якого було зроблено загальні висновки, статистичні методи для здійснення розрахунків.

Інформаційною базою кваліфікаційної магістерської роботи є дані зібрані на сайті https://cybermap.kaspersky.com/.

Основний науковий результат кваліфікаційної магістерської роботи полягає у такому: були розроблені і перевірені на адеквантість моделі прогнозування трендів кібератак, обрано релевантну, що дозволяє отримати інформацію про тренд кібератаки на майбутній період.

Одержані результати можуть бути використані державними чи комерційним органами, чия інформація знаходиться під загрозою знищення та викрадення, для прийняття правильного рішення про захист даних та гарантування державної безпеки.

Ключові слова: кібератака, кібервійна, моделювання, прогнозування, фіксовані ефекти, випадкові ефекти, LSTM модель.

Зміст кваліфікаційної магістерської роботи викладено на 52 сторінках. Список використаних джерел із 40 найменувань, розміщений на 3 сторінках. Робота містить 4 таблиці, 75 рисунків, а також 1 додаткф, розміщених на 3 сторінках.

Рік виконання кваліфікаційної роботи – 2022 рік.

Рік захисту роботи – 2022 рік.

Міністерство освіти і науки України

Сумський державний університет

Навчально-науковий інститут бізнесу, економіки та менеджменту

Кафедра економічної кібернетики

ЗАТВЕРДЖУЮ

Завідувач кафедри

доцент

_____ В.В. Койбічук

“__” _____ 20__ р.

ЗАВДАННЯ НА КВАЛІФІКАЦІЙНУ МАГІСТЕРСЬКУ РОБОТУ

(спеціальність 051 Економіка «Економічна кібернетика»)

студенту 2 курсу, групи ЕК.м-11

Кобзенко Вікторії Вікторівні

(прізвище, ім’я, по батькові студента)

1. Тема роботи Моделювання та прогнозування трендів кібератак затверджена наказом по університету від «__» ____ 2022 року № _____

2. Термін подання студентом закінченої роботи «__» _____2022 року

3. Мета кваліфікаційної роботи - аналіз сучасної ситуації, пов’язаної з кібератаками, вивчення тенденцій для підтримки кібербезпеки та розробка математичних моделей, моделювання та прогнозування трендів кібератак на основі панельних даних джерел, які звітують про кількість кібератак, а саме потоку даних зі шкідливими програмами у поштових додатках, підозрілий поштовий трафік та потік даних з виявлених мережевих атак.

4. Об’єкт дослідження – кількість кібератак.

5. Предмет дослідження - методи і моделі дослідження та прогнозування панельних даних.

6. Кваліфікаційна робота виконується на матеріалах зібраних на сайті https://cybermap.kaspersky.com/.

7. Орієнтовний план кваліфікаційної роботи, терміни подання розділів керівникові та зміст завдань для виконання поставленої мети

Розділ 1 Теоретико-методологічні аспекти моделювання та прогнозування трендів кібератак

У розділі 1 необхідно розглянути поняття, зміст та види кібертак та сучасному етапі, проаналізувати статистичні дані кібератак та сучасні тренди кібербезпекової політики та побудувати концептуальну модель моделювання та прогнозування трендів кібератак.

Розділ 2 Розробка моделей прогнозування трендів кібератак

У розділі 2 необхідно провести опис вхідних даних, їх статистичний аналіз та візуалізацію та розглянути теоретичні засади моделювання на основі панельних даних

Розділ 3 МОДЕЛЮВАННЯ ТА ПРОГНОЗУВАННЯ ТРЕНДІВ КІБЕРАТАК

У розділі 3 необхідно побудувати регресійні моделі для всіх незалежних змінних, оцінити отримані результати, обрати релевантну модель для подальшого прогнозування та побудувати прогнозні моделі

8. Консультації з роботи:

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
1			
2			
3			

9. Дата видачі завдання «__» _____ 2022 року

Керівник кваліфікаційної роботи _____ _____
(підпис) (ініціали, прізвище)

Завдання до виконання одержав _____ _____
(підпис) (ініціали, прізвище)

ЗМІСТ

ВСТУП

Питання кібербезпекової політики стало нагальним з початком розвитку сучасних технологій, оскільки автоматизовані системи важливих державних та комерційних об’єктів знаходяться під загрозою, не тільки вірусів, а й кібератак. У таких умовах сучасності пошук нових можливостей гарантування державної безпеки набуває особливого значення.

Кібератака – це напад, який здійснюється кіберзлочинцями, з одного або декількох комп’ютерів проти одного чи декількох комп’ютерів чи мереж. Кібератака призводить до виведення з ладу пристроїв, використання зламаних даних, та точки запуску нових атак на інші пристрої.

Проникнення до інформаційного простору будь-якої компанії чи державного об’єкту вважається кримінальним злочином та називається кіберзлочинністю чи кібертероризмом.

Зважаючи на те, що проти України вже 8 рік ведеться гібридна війна з боку російської федерації, а з початком широкомаштабного вторгнення на територію України кібератаки нарощують свої обороти.

Основне завдання моделювання та прогнозування трендів кібератак є попередження наслідків та створення захисту, особливо для критичної інфраструктури країни.

Актуальність теми полягає в тому, що в реаліях сьогодення суспільство стикається кожен день з кібератаками, які можуть згубно подіяти на важливі сфери діяльності. Тому саме моделювання та прогнозування кібератак зможуть запобігти та захистити дані.

Метою даного дослідження є аналіз сучасної ситуації, пов’язаної з кібератаками, вивчення тенденцій для підтримки кібербезпеки та розробка математичних моделей, моделювання та прогнозування трендів кібератак на основі панельних даних джерел, які звітують про кількість кібератак, а саме

потоку даних зі шкідливими програмами у поштових додатках, підозрілий поштовий трафік та потік даних з виявлених мережевих атак.

Предметом дослідження виступають методи і моделі дослідження та прогнозування панельних даних.

Об’єктом дослідження є кількість кібератак.

Для досягнення поставленої мети необхідно реалізувати наступні задачі:

Для реалізації поставленої мети необхідно реалізувати наступні задачі:

1) розкрити сутність об’єкту дослідження - кібератак та проаналізувати сучасні тренди кібербезпекової політики;

2) розробити концептуальну модель моделювання та прогнозування трендів кібератак;

3) описати вхідні дані для моделювання та провести статистичний аналіз та візуалізацію даних;

4) побудувати математичні моделі та інтерпретувати отримані результати;

5) перевірити моделі на адекватність;

6) побудувати прогнозні моделі та обрати релевантну.

При дослідженні теми в роботі було використано такі загальнонаукові методи: аналіз і синтез, дедукція, абстрагування, конкретизація, аргументація, порівняння, класифікація та метод узагальнення, за допомогою якого було зроблено загальні висновки, статистичні методи для здійснення розрахунків.

Інформаційно-фактологічну базу склали: набір емпіричних статистичних даних, який допоміг зробити спостереження для моделювання та прогнозування трендів кібератак; документація по мові програмування Python, за допомогою якої здійснювалися розрахунки.

Результатом роботи повинна стати прогнозна модель кібератак, яка зможе попередити про збільшення атак вчасно та допомогти розробити відповідні заходи щодо запобігання чи зменшення наслідків атак.

РОЗДІЛ 1 ТЕОРЕТИКО-МЕТОДОЛОГІЧНІ АСПЕКТИ МОДЕЛЮВАННЯ ТА ПРОГНОЗУВАННЯ ТРЕНДІВ КІБЕРАТАК

1.1 Поняття, зміст та види кібератак на сучасному етапі.

Кібератака – це спроба реалізації кіберзагрози, тобто, отримати несанкціонований доступ до обчислювальної техніки чи комп'ютерної мережі з метою заволодіння даними або з наміром завдати шкоди. Кібер загрози мають широкий спектр атак, починаючи від крадіжки персональних даних, маніпулювання ними, запуску вірусів до мережі, порушення цифрового добробуту та стабільності господарчих підприємств та підприємств критичної інфраструктури держав [1].

Фактично кібератаки впливають на кожен аспект нашого життя. Вони можуть призвести до відключення електроенергії, виведення з ладу військового обладнання та викрадення конфіденційних даних, які порушують національну безпеку країни.

Людей, які здійснюють такі атака, називають хакерами, кіберзлочинцями. Злочинці можуть діяти як поодиноко, так і групами. Також, вони можуть належати до злочинних угрупувань, працюючі з іншими загрозливими суб'єктами, щоб більш детально знаходити слабкі місця в комп'ютерних мережах.

ІТ підрозділи, які фінансуються державами, також можуть здійснювати кібератаки. Тоді їх ідентифікують як злочинців, які загрожують національній безпеці держави.

Кібератаки призначені для завдання шкоди та переслідують різні цілі, а саме:

— фінансова - отримання фінансової вигоди. Здійснюються проти комерційних підприємств та фізичних осіб, спрямовані на викрадення

персональних даних та номерів кредитних карток, які потім використовується для доступу до грошей;

— помста - здіснюються для сіяння хаосу та плутанини, розчарування та недовіру. Прикладом таких угрупувань є Anonymous. Вони вважаються кіберактивістами, які боряться за справедливість та національні інтереси жертв;

— кібервійна - участь у таких атаках беруть всі країни світу. Кібервійни відбуваються у рамках економічних, політичних чи соціальних суперечок [2].

До найпоширеніших видів кібератак належать:

— Denial of service (DoS attack) - це мережева атака, під час якої здійснюється перенавантаження компонентів комп'ютерних систем та прагне зробити пристрій недоступним для користувачів.

— Phishing - це атака, яка використовує засоби соціальної інженерії для викрадення персональних через створення копій відомих сайтів.

— Malware - це поширена кібератака, через запуск усередину комп'ютера шкідливого програмного забезпечення (віруси, трояни), яке виконує неавторизовані дії в системі жертви.

— Ransomware - це атака, яка здійснює запуск всередину комп'ютера шкідливого програмного забезпечення, що блокує та шифрує дані жертви, або робить їх копію задля подальшого шантажу.

— Man-in-the-Middle - це мережева атака, в ході якої злочинець отримує дані, які передаються між двома користувачами та замінює їх фіктивними.

— Zero-day exploit - це атака, яка здійснюється на вразливі місця ліцензованого програмного забезпечення, які ще невідомі розробнику.

— Cross-site scripting (XSS) - це атака, яка здійснюється шляхом додавання до сайту небезпечного коду. Під час користування таким сайтом дані користувача можуть передаватися хакерам.

— Logic bombs – це атака, яка має набір інструкцій у програмі, що несе зловмисне навантаження та може атакувати операційну систему [3].

1.2 Статистика кібератак та сучасні тренди кібербезпекової політики.

Становлення сучасного інформаційного суспільства дає змогу не лише будувати більш ефективний соціум, а й надає нових викликів традиційним національним загрозам безпеки держави та створює нові складнощі для системи національної безпеки.

Більшість держав світу працює над модернізацією та удосконаленням власних секторів безпеки, щоб бути готовими до нових викликів сучасності. Цей процес відбувається завдяки активним реформуванням системи управління безпекою держави, впорядкування нормативних норм, щодо понять кібербезпеки, активною роз'яснювальною роботою серед населення країни щодо кіберзагроз, збільшенням чисельності спеціальних підрозділів, структур у системі кіберзахисту та посиленням контролю за національною інформаційною безпекою держави [4].

Нинішні реалії та тренди, а саме: віддалена робота, спричинена вірусом COVID-19, спричиняє поширення хмарних операцій; розширення мереж 5G - підключає пристрої на вищих швидкостях і більшій пропускній здатності; криптовалюти вибухнули в популярності, і тепер вони купуються, продаються та торгуються людьми у більших масштабах, ніж будь-коли раніше. Ці всі сьогоденні реалії спричинили більш прискіпливу увагу до політики безпеки компанії, держави та взагалі суспільства.

Тому, очікуються на закріпляються наступні тенденції кіберзагроз 2022 року у світі:

— щорічне зростання користувачів Інтернету призводить до підвищення кіберзлочинності та її жертв;

— криптовалюти підлягатимуть жорсткішому регулюванню, оскільки їх впровадження зростатиме;

— організації соціальних мереж працюватимуть над суворішим наглядом за обміном інформацією;

— віддалені працівники залишаться мішенню для кіберзлочинців;

— через віддалену робочу силу збільшуватимуться випадки проникнення в хмару;

— навички кібербезпеки являються дефіцитною проблемою, оскільки все більше вакансій залишатимуться незаповненими.

— пристрої стануть більш вразливими до кібератак, оскільки 5G збільшує пропускну здатність підключених пристроїв [5].

У 2020 році кібератаки посіли п'яте місце серед найбільших ризиків (економічних, ризики навколишнього середовища, геополітичних, соціальних) [6].

У 2022 році ця тенденція зберігається та продовжує розвиватися.

Через пандемію випадків кібератак збільшилось на 600% [7].

Cybersecurity Ventures (міжнародний експерт з кібербезпеки) повідомляє, що до 2025 року кіберзлочинність коштуватиме компаніям у всьому світі приблизно 10,5 трлн доларів США щорічно, порівняно з 3 трлн доларів у 2015 році [8].

Корпорація Microsoft опублікувала у 2021 році Digital Defence Report – звіт, у якому було висвітлено дослідження та надано висновки про державні загрози та найбільші цільові сектори. Виявилося, що майже 80% кібератак здійснюється на національну державу та спрямовані проти державних установ, аналітичних центрів і неурядових організацій (див. рис. 1.1):

Рисунок 1.1- Сектори, на які націлені кібеатаки (липень 2020 року – червень 2021 року)

Сполучені Штати Америки є найбільш цільовою країною для кібератак останні роки. Російська компанія NOBELIUM здійснює кібератаки на держави, а також сильно націлилася на Україну, особливо зосередившись на державних інтересах, пов'язанихіз російсько-українською війною (див. рис. 1.2) [9].

Рисунок 1.2 – Країни, на які найбільше націлено кібератак (липень 2020 року – червень 2021 року)

З початком повномасштабного вторження росії на територію України було зафіксовано понад 3,5 тис. кібератак (жовтень 2022 року). Кіберфахівці нейтралізували ці атаки, які були націлені на електронні системи центральних органів влади та об'єкти критичної інфраструктури України. З 3,5 тис. атак 1650 було виявлено у режимі «реального часу» за допомогою системи

управління кіберзагрозами, що створена на базі служби безпеки України (СБУ). Було встановлено, що переважна більшість атак була спрямована на повне знищення цифрових сервісів, або дестабілізувати стратегічно важливі підприємства енергетичної галузі (див. рис. 1.3).

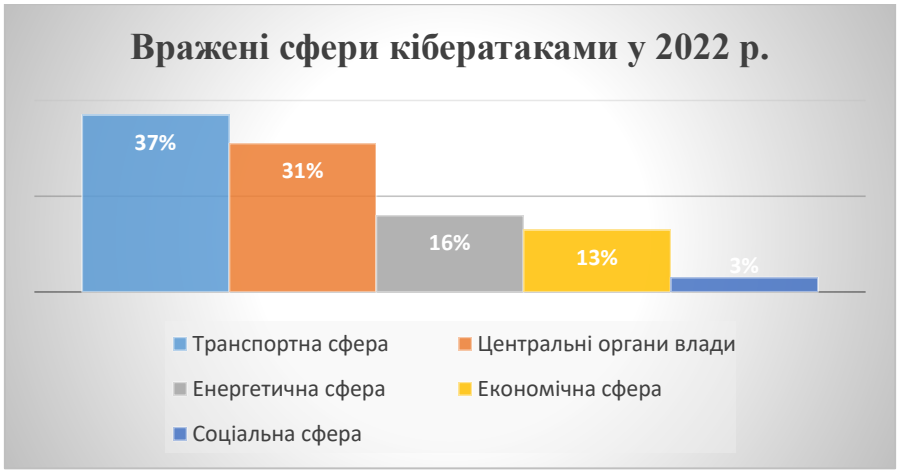


Рисунок 1.3 – Сфери, які були вражені росією у 2022 р. в Україні

До організації та проведення кібератак причетні виключно російські спецслужби та підконтрольні їм хакерські групи [10].

Останнім часом спостерігаються п’ять основних тенденцій, які стають все більш вирішальними для підтримки кібербезпеки комерційних та державних установ. Ці кроки є дієвими, які компанії можуть запроваджувати, щоб підвищити поточний рівень кібербезпеки та забезпечити надійний захист вразливих систем.

До п’яти основних тенденцій належать:

— безпека блокчейну.

Блокчейн – це нова технологія обробки та зберігання інформації, вибудований за певними правилами ланцюгів, розташованих навколо мережі. Блокчейн є безпечним методом захисту даних, оскільки він вимагає одночасної зміни принаймні половини вузлів у ланцюжку, щоб повністю змінити запис [11];

— покращена безпека кінцевої точки та пристрою.

Безпека кінцевої точки застосовується до будь-якого пристрою, який знаходиться в хмарній мережі. Кожен пристрій, який має доступ до системи, слід вважати кінцевою точкою. Однак використання антивірусної безпеки для захисту цих пристроїв не є ефективною політикою кібербезпеки, оскільки загрози для кінцевих точок розвиваються щодня. Сучасні методи захисту кінцевих точок і пристроїв включають машинне навчання та штучний інтелект (AI), щоб допомогти розпізнавати зловмисні підписи та ізолювати їх, доки пристрій не отримає перевірку від сервера;

— підвищений акцент на обізнаності користувачів.

Одним із найбільш типових методів отримання несанкціонованого доступу до даних є соціальна інженерія. Основна передумова цього методу полягає в тому, що люди схильні бути довірливими. Завоювавши довіру людини, хакер потенційно може отримати пароль цієї особи. Тому необхідно проводити навчання як серед співробітників компаній, так і серед звичайного населення, щоб підвищити обізнаність суспільства;

— більш часті й ефективні патчі програмного забезпечення.

Під словом патч розуміється значення «виправити». Зазвичай це невеликий програмний код, призначений для виправлення проблеми (помилки), в межах операційної системи чи програми [12];

— розширене розкриття інформації про порушення з боку компаній.

Визнавати помилки – властива не тільки людям, а й компаніям. Проблема з’являється, коли компанія має право на витік даних, але при цьому не інформує про це своїх клієнтів - постраждалих осіб. Несвоєчасне повідомлення про витік даних може мати жахливі наслідки для осіб, чиї дані було втрачено.

Забезпечення інформацією користувачів, про те що хакери мають їхні дані, допомагає їм краще впоратися з ситуацією. Користувачі можуть вчасно змінити пароль, або зв’язатися з банком, щоб змінити дані. Однак, якщо користувачі не знають, що їхня інформація знаходиться в руках зловмисника, вони не можуть відреагувати належним чином [13].

1.3 Побудова концептуальної моделі моделювання трендів кібератак.

Головна мета побудови концептуальної моделі - це моделювання трендів кібератак у світі та прогнозування їх.

Побудова концептуальної моделі являє собою масштабний процес, який включає низку етапів, починаючи з виявлення та аналізу реальних проблем, до побудови моделі для моделювання та прогнозування трендів кібератак.

Даний процес може бути реалізований за допомогою наступних кроків:

– постановка задачі;

– підготовка даних;

– статистичний аналіз та візуалізація;

– побудова трьох моделей, де залежними змінними являються кожен із джерел виявлення кібератаки;

– прогнозування за допомогою моделей (див. рис. 1.4).

Побудова даної моделі передбачає використання статистичних даних про кібератаки 40 країн світу. Обрано по 10 країн з Європи, Азії, Африки та по 5 країн з Північної та Південної Америки.

Обрано три джерела, які збирають інформацію про кількість кібератак, а саме:

- MAV (Mail Anti Virus) – поштовий антивірус показує потік даних за шкідливими програмами, виявленими серед нових об'єктів у поштових додатках. Поштовий антивірус перевіряє вхідні повідомлення та запускає автоматичну перевірку при збереженні вкладених файлів на диск.
- KAS (Kaspersky Anti-Spam) - Касперський Анти-Спам показує підозрілий та небажаний поштовий трафік, виявлений за допомогою технологій репутаційної фільтрації «Лабораторії Касперського».
- IDS (Intrusion Detection Scan) - система виявлення вторгнень показує потік даних з виявлених мережевих атак [14].

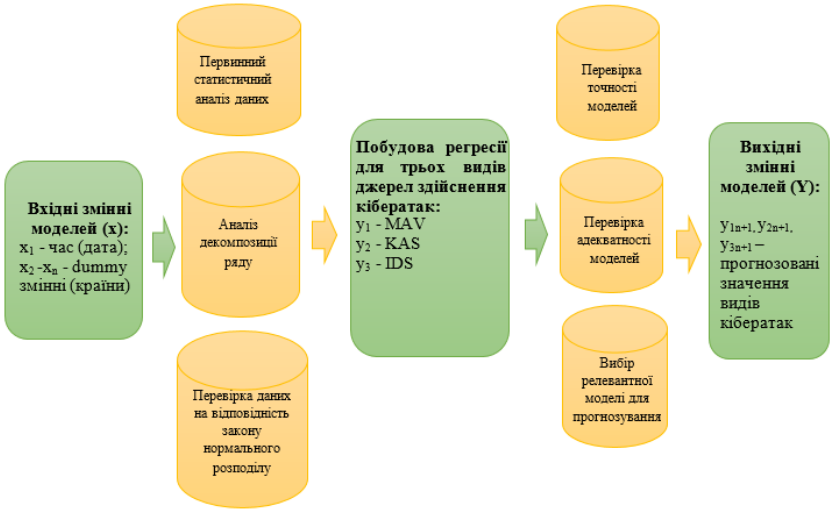


Рисунок 1.4 - Концептуальна модель моделювання та прогнозування трендів кібератак

Розглянемо алгоритм моделювання та прогнозування трендів кібератак:

– визначення вхідних даних для побудови регресії панельних даних для визначення залежності кібератак від часу та країни (незалежних даних);

– дослідження вхідних даних, аналіз та візуалізація, виявлення тенденцій та взаємозв’язків між змінними, проведення коригування даних за необхідності;

– побудова математичних моделей у вигляді представлення математичних залежностей, які описуватимуть вхідні дані;

– моделювання результуючої величини та отримання даних, які ймовірно впливатимуть на прогнозування трендів;

– інтерпретація отриманих результатів моделювання;

– формування висновків - аналіз результатів та узагальнення тенденції.

Таким чином, використання розробленої моделі дасть змогу змоделювати, прогнозувати та попередити типові кіберзагрози з метою попередження користувачів.

РОЗДІЛ 2 РОЗРОБКА МОДЕЛЕЙ ПРОГНОЗУВАННЯ ТРЕНДІВ КІБЕРАТАК

2.1 Опис вхідних даних, їх статистичний аналіз та візуалізація.

Вхідними даними для побудови моделей та прогнозування трендів кібератак є статистичні дані 40 країн світу, обраних рандомним методом. База даних містить інформацію про назву країни («Country», дату («Date») – щоденна статистика про кількість кібератак з 14 серпня 2022 року до 13 вересня 2022 року, та дані по трьом видам джерел фіксування кібератак, а саме (табл. 2.1):

- MAV (Mail Anti Virus) – поштовий антивірус показує потік даних за шкідливими програмами, виявленими серед нових об'єктів у поштових додатках. Поштовий антивірус перевіряє вхідні повідомлення та запускає автоматичну перевірку при збереженні вкладених файлів на диск.
- KAS (Kaspersky Anti-Spam) - Касперський Анти-Спам показує підозрілий та небажаний поштовий трафік, виявлений за допомогою технологій репутаційної фільтрації «Лабораторії Касперського».
- IDS (Intrusion Detection Scan) - система виявлення вторгнень показує потік даних з виявлених мережевих атак [14].

Таблиця 2.1 – Опис вхідних даних

Назва змінної	Опис	Тип даних
Country	Країна дослідження	категоріальні
Date	Дата спостереження	дата
MAV	Mail Anti Virus	число
KAS	Kaspersky Anti-Spam	число
IDS	Intrusion Detection Scan	число
Date_num	Період	число

Статистичний аналіз даних та візуалізація буде виконуватися за допомогою мови програмування Python.

Для початку роботи з базою даних необхідно імпортувати необхідні бібліотеки:

```
import pandas as pd
import numpy as np
```

Рисунок 2.1 – Імпортування бібліотек

Бібліотека Pandas – це потужний інструмент, який працює з даними. Numpy – це бібліотека, яка полегшує ефективні числові операції з великими обсягами даних, виконання інших операцій з рядками та стовпчиками таблиць [15].

Всі необхідні бібліотеки завантажені, тепер потрібно імпортувати базу даних для її візуалізації [16]:

```
df = pd.read_excel('diplom_work.xlsx')
df
```

	Date	Country	Date_num	Country_id	MAV	KAS	IDS
0	2022-08-14	Ukraine	1	1	121	1177500	20174
1	2022-08-15	Ukraine	2	1	217	1658000	20195
2	2022-08-16	Ukraine	3	1	243	1624000	21013
3	2022-08-17	Ukraine	4	1	570	1794000	21536
4	2022-08-18	Ukraine	5	1	436	1646000	20052
...	...	...	...	...	...	...	...
1235	2022-09-09	Brazil	27	40	35520	28600000	318400
1236	2022-09-10	Brazil	28	40	5682	22336000	290281
1237	2022-09-11	Brazil	29	40	4052	18769000	298359
1238	2022-09-12	Brazil	30	40	28461	22110000	336290
1239	2022-09-13	Brazil	31	40	24481	17919500	358940

1240 rows × 7 columns

Рисунок 2.2 – Імпортування бази даних

Після імпорту бази даних, треба зробити низку операцій, щодо перевірки інформації про наявні рядки та стовпчики, відсутність даних.

Перевіряєм базу даних на відсутність даних:

```
df.isna().sum()
```

```
Date          0
Country       0
Date_num      0
Country_id    0
MAV           0
KAS           0
IDS           0
dtype: int64
```

Рисунок 2.3 – Перевірка на відсутність даних

На рис. 2.3 зображено вихідна інформація, про те, що база даних не має відсутніх даних.

Після цього ми використаємо функцію .describe(), яка може надати більшість статистичних елементів:

```
df.describe() #statistics
```

	MAV	KAS	IDS	Weekday
count	1240.000000	1.240000e+03	1.240000e+03	1240.000000
mean	4647.270968	7.617245e+06	1.521110e+05	3.935484
std	7700.384493	2.009227e+07	2.515326e+05	2.078919
min	1.000000	3.500000e+03	2.000000e+00	1.000000
25%	286.000000	1.403750e+05	8.927000e+03	2.000000
50%	1630.500000	7.640000e+05	4.623700e+04	4.000000
75%	5174.000000	4.785625e+06	2.133608e+05	6.000000
max	77612.000000	1.810050e+08	2.643943e+06	7.000000

Рисунок 2.4 – Статистичний аналіз даних

Так, як панельні дані являються і часовим рядом, необхідно зробити декомпозицію часового ряду та дослідити дані на нормальний розподіл даних.

```
from statsmodels.tsa.seasonal import seasonal_decompose
decompose = seasonal_decompose(df['MAV'],model='additive', period=30)
decompose.plot()
plt.show()
```

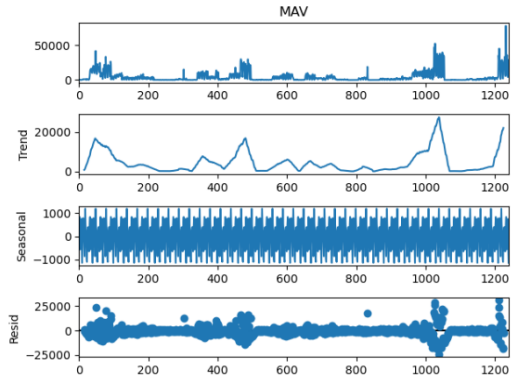


Рисунок 2.5 – Декомпозиція трендів для змінної «MAV»

```
from statsmodels.tsa.seasonal import seasonal_decompose
decompose = seasonal_decompose(df['KAS'],model='additive', period=30)
decompose.plot()
plt.show()
```

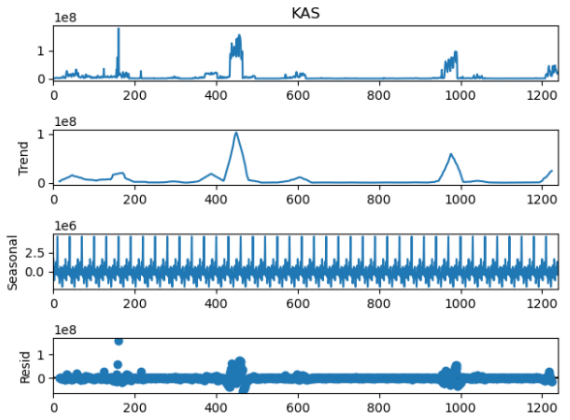


Рисунок 2.6 – Декомпозиція трендів для змінної «KAS»

```
from statsmodels.tsa.seasonal import seasonal_decompose
decompose = seasonal_decompose(df['IDS'],model='additive', period=30)
decompose.plot()
plt.show()
```

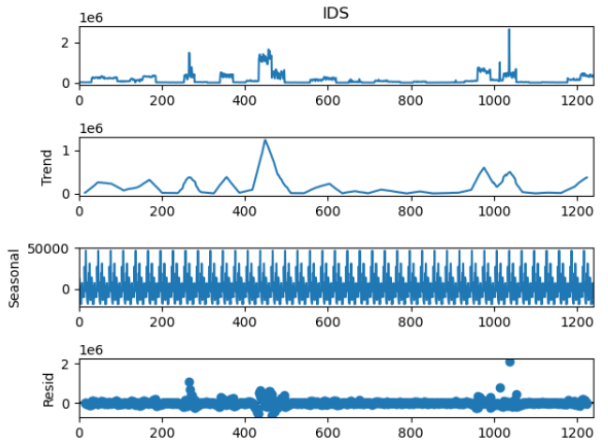


Рисунок 2.7 – Декомпозиція трендів для змінної «IDS»

Декомпозиція трендів є одним корисним способом візуалізації тенденцій у даних часових рядів. Щоб продовжити, давайте імпортуємо seasonal_decompose з пакету statsmodels (див. рис. 2.5-2.7) [17].

Перевірку на нормальність - відповідність даних нормальному розподілу виконаємо двома методами: методом побудови гістограм та обчислення тесту Харке-Бера:

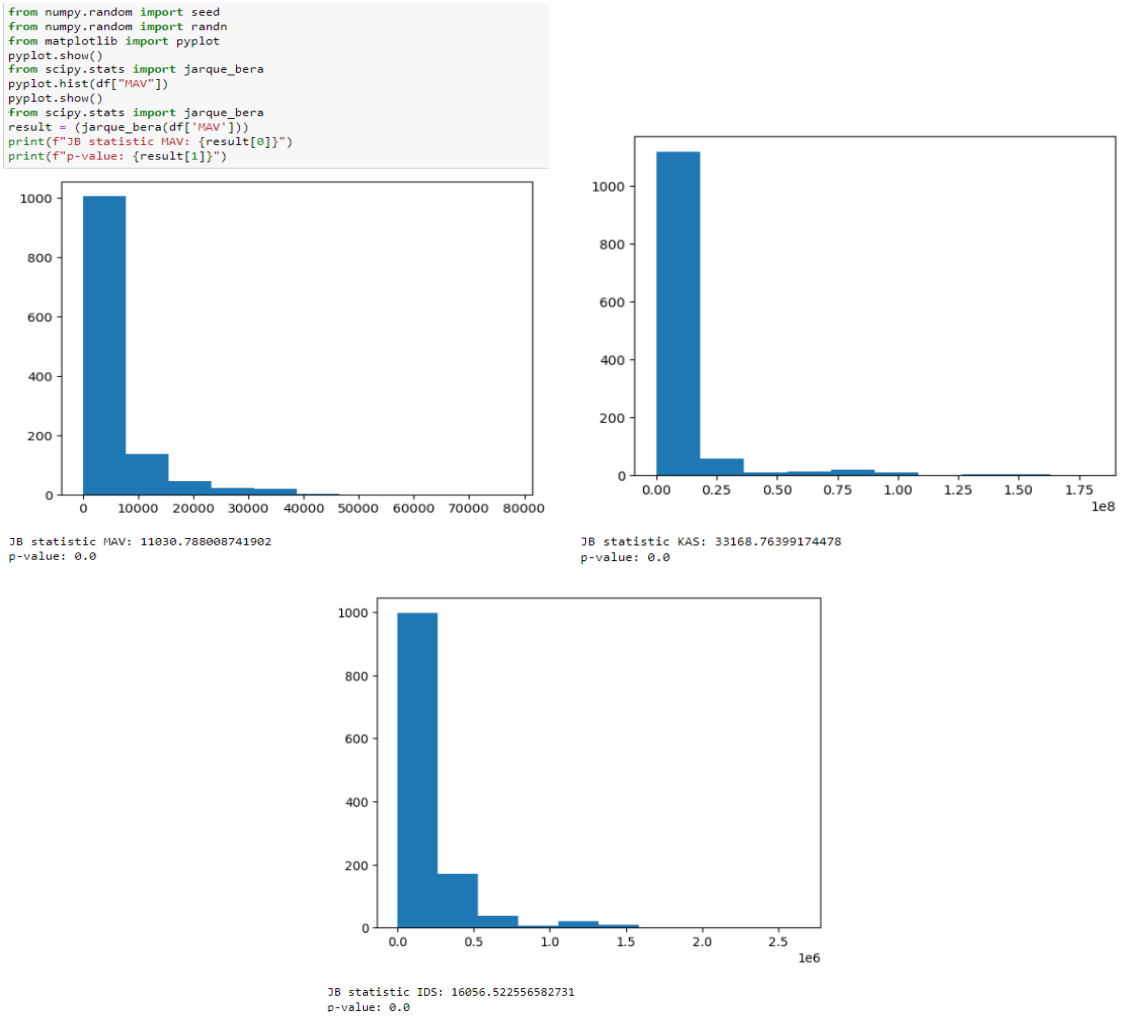


Рисунок 2.8 – Перевірка на нормальний розподіл змінних

Якщо гістограма має приблизно «дзвіноподібну форму», то дані вважаються нормально розподіленими. В цьому випадку змінні не мають такої форми [18].

Статистика тесту Харка-Бера завжди є позитивним числом, і якщо вона далека від нуля, це вказує на те, що вибіркові дані не мають нормального розподілу.

Розрахунок тесту Харка-Бера показів наступні результати: для змінної MAV – 144.3 при p-value 0.0, для змінної KAS – 23.94 при p-value 0.0 та для IDS – 123.23 при p-value 0.0.

Побудувавши гістограми та розрахувавши тест Харка-Бера можна зробити висновок, що не одна змінна не відповідає нормальному розподілу даних (див. рис. 2.8) [19].

Для наближення даних до нормального розподілу можна виконати логарифмування. Перетворення даних з x на log(x) [20].

```
import matplotlib.pyplot as plt

#create log-transformed data
logMAV = np.log(df["MAV"])
#define grid of plots
fig, axs = plt.subplots(nrows=1, ncols=2)
#create histograms
axs[0].hist(df["MAV"], edgecolor='black')
axs[1].hist(logMAV, edgecolor='black')

#add title to each histogram
axs[0].set_title('Original Data')
axs[1].set_title('Log-Transformed Data')
```

```
Text(0.5, 1.0, 'Log-Transformed Data')
```

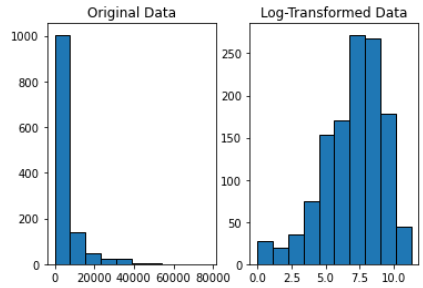


Рисунок 2.9 – Трансформування даних з x на log(x)

Такі перетворення здійснюємо для інших змінних, які не відповідали нормальному розподілу (Додаток А).

Зобразимо розподіл кожного з джерел виявлення кібератаки за країною. Для прикладу обираємо країну - Україну та будуємо графік (рис. 2.10):

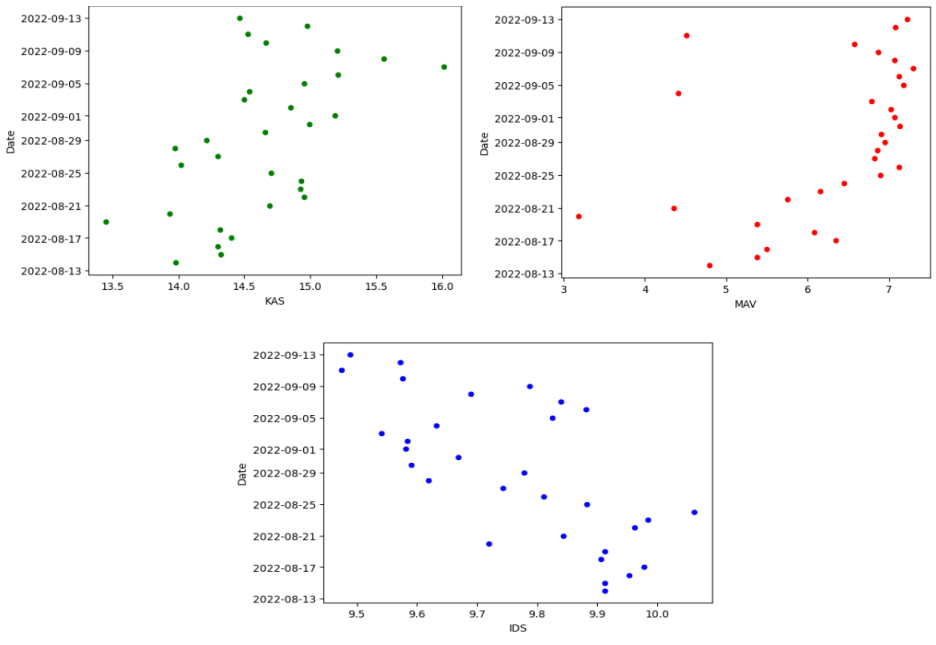


Рисунок 2.10 – Графік здійснення кібератак в Україні

Статистичний аналіз виконано, в результаті якого, незалежні змінні, які будуть виступати в ролі y1, y2 та y3 – прологарифмовані, для відповідності закону нормального розподілу даних.

2.2 Теоретичні засади моделювання на основі панельних даних.

Для врахування всіх особливостей розвитку соціально-економічних процесів, практикують об'єднання динамічних та просторових рядів. Найчастіше застосовують моделі на основі панельних даних.

Панельні дані – це дані, які містять статистичну інформацію про одну і ту ж множину об'єктів за ряд послідовних періодів часу.

Панельні дані часто передбачають одночасне спостереження за багатьма змінними, щоб максимізувати обсяг аналізу та встановити тенденції між змінними. Країни, окремі регіони, демографічні групи, економічні показники, організації та окремі особи є типовими прикладами суб'єктів панельних даних. Зазвичай панельні дані використовуються для статистичних, економічних або фінансових дослідженнях.

Як правило, для кожної одиниці вимірюється один або декілька параметрів (також регресійні змінні або ефекти) за кожен період часу. Набір точок даних, що відносяться до однієї одиниці (однієї країни), називається групою.

Якщо всі одиниці відстежуються протягом однакової кількості періодів часу, панель даних називається збалансованою панеллю. В іншому випадку це називається незбалансованою панеллю. Якщо один і той самий набір одиниць відстежується протягом дослідження, це називається фіксованою панеллю, але якщо одиниці змінюються під час дослідження, це називається обертовою панеллю [21].

Масиви панельних даних поєднують у собі дані часових рядів, так і просторових. Тобто для кожного об'єкта з вибірки ми маємо сукупність

часових показників (часового ряду) та для кожного моменту часу – просторову вибірку (перехресні дані) [22].

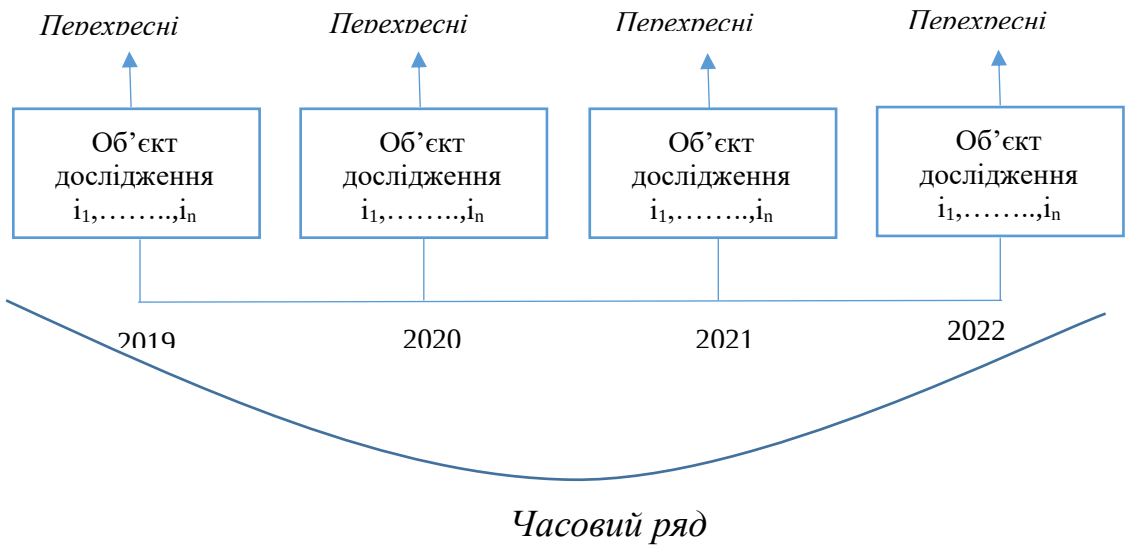


Рисунок 2.11 - Ілюстрація дизайну панельних даних

Динамічні та просторові змінні об'єднуються в один статистичний масив, одиницею якого є об'єкто-період. Припускається, що всі елементи мають однакову поведінку, тому панельні дані не мають лагових змінних [16].

Головною особливістю статично-динамічної інформації – це залежність між спостереженнями. Залежними виявляються не лише динамічні рядди, але й ряди в цілому, як просторові, так і часові.

Так, залежність між рядами динаміки - це результат просторової варіації, яка через інерційність процесів зберігається певний час. Залежність просторових рядів відображає синхронність динаміки показників по окремих об'єктах, зумовлену спільними умовами розвитку. Ігнорування цих особливостей інформаційної бази моделювання призводить до помилкових висновків.

Для аналізу панельних даних використовують три типи моделей регресії:

— Об'єднана регресійна модель OLS (Pooled OLS Regression Model);

— Модель регресії фіксованих ефектів (The Fixed Effects Regression Model);

— Модель регресії випадкових ефектів (The Random Effects Regression Model).

Об’єднана регресійна модель OLS (надалі Pooled OLS) є загальноприйнятою для панельних даних. Ця модель використовується як базова (відправна точка) або еталонна модель для порівняння з результатами інших моделей. Pooled OLS можна описати як просту лінійну регресійну модель OLS (метод найменших квадратів). Модель ігнорує час на індивідуальні характеристики, якщо їх не включити в саме рівняння регресії, та зосереджується лише на залежності між заданими змінними.

Рівняння регресії представлене у наступному математичному вигляді (2.1):

$$y_i = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n + \epsilon_{it} \,, \qquad (2.1)$$

де y_i – це залежна змінна;

x_1, x_n – незалежні змінні;

β_0 – індивідуальний ефект;

β_1, β_n – це параметри рівняння регресії;

ϵ_{it} – це похибка.

Проте таке регресійна модель вимагає відсутності кореляції між незалежними змінними. Проблема з Pooled OLS полягає в тому, що індивідуальний ефект β_0 може мати послідовну кореляцію з часом. Pooled OLS здебільшого не підходить для даних панелей [23].

Модель фіксованих ефектів (надалі FE-модель) – це модель , яка визначає окремі ефекти неспостережуваних незалежних змінних як постійні – «фіксовані» протягом усього часу. У FE-моделях зв’язок між незалежними і неспостережуваними змінними може існувати. Ця модель допускає

неоднорідність всередині моделі. Рівняння FE-моделі представлене у наступному математичному вигляді [24] (2.2):

$$y_{it} = x_{it}\beta_i + d_{it}c_i + \epsilon_{it} \quad (2.2)$$

де c_i – це значення фіксованих ефектів;

d_{it} – це вектор фіктивних змінних.

Модель випадкових ефектів (надалі RE-модель) – це модель, яка визначає окремі ефекти неспостережуваних незалежних змінних як випадкових змінних з часом. Модель здатна «перекликатися між OLS-моделлю та FE-моделлю, а отже, може зосередитися на обох [25].

У моделі випадкових ефектів ми припускаємо, що специфічні для одиниць ефекти для всіх одиниць розподіляються навколо загального середнього значення відповідно до деякого невідомого розподілу ймовірностей. Крім того, це загальне середнє є постійним протягом усіх періодів часу на панелі даних.

З точки зору нотації, ці припущення породжують наступне умовне очікування для члена *Ziγi*, який призначений для охоплення всіх неспостережуваних специфічних ефектів одиниці (2.3):

$$E(Z_i y_i | x_i) = \alpha \quad (2.3)$$

Тепер додаємо та віднімаємо очікуване значення рівняння регресії моделі (2.4):

$$y_i = x_i\beta_i + Z_i y_i - E(Z_i y_i | x_i) + E(Z_i y_i | x_i) + \epsilon_i \quad (2.4)$$

У наведеному вище рівнянні член $Z_i y_i - E(Z_i y_i | x_i) + E(Z_i y_i | x_i)$ є зміною специфічного ефекту одиниці навколо його середнього. Будемо позначати цю варіацію терміном μ_i

Рівняння RE-моделі представлене у наступному математичному вигляді (2.5):

$$y_{it} = x_{it}\beta_i + a + (\mu_i + \epsilon_{it}) \quad (2.5)$$

Для побудови RE-моделі необхідно виконати наступні кроки:

1) Оцінити компоненти дисперсії $\sigma^2\epsilon$ і σ^2u. $\sigma^2\epsilon$ і $\sigma^2\mu$ - дисперсії компонентів похибки μ і ϵ.

2) Оцінити групові середні

3) Розрахунок центрованої панелі даних.

Після оцінки результатів побудованих моделей, необхідно обрати найкращу та побудувати на основі обраної моделі прогнозну. Для порівняння результатів також побудуємо LSTM модель.

LSTM модель (Long short-term memory) – це особливий вид рекурентних нейронних мереж (нейронні мережі, розроблені для роботи з тимчасовими даними), яка здатна вивчати довгострокові залежності в даних. Це досягається тим, що повторюваний модуль моделі має комбінацію з чотирьох шарів, що взаємодіють між собою. LSTM модель – це одна з найкращих видів моделей для прогнозування часових рядів [26].

Стан клітини в LSTM допомагає інформації протікати через одиниці, не змінюючись, дозволяючи лише кілька лінійних взаємодій. Кожен блок має вхід, вихід і забуті ворота, які можуть додавати або видаляти інформацію до стану комірки. Забуті ворота вирішують, про яку інформацію з попереднього стану клітини слід забути, для чого вона використовує сигмовидну функцію. Вхідний затвор контролює потік інформації до поточного стану комірки за допомогою точкової операції множення 'sigmoid' і 'tanh' відповідно. Нарешті, вихідні ворота вирішують, яку інформацію слід передати в наступний прихований стан [27].

Існує два типи нормалізуючих рівнянь, які використовуються в LSTM. Перша -це сигмовидна функція (представлена сигмою нижнього регістру), а друга -функція танга [28].

РОЗДІЛ 3 МОДЕЛЮВАННЯ ТА ПРОГНОЗУВАННЯ ТРЕНДІВ КІБЕРАТАК

3.1 Побудова регресійних моделей для змінної «MAV».

3.1.1 Побудова об’єднаної регресійної моделі (Pooled OLS).

Як згадувалося раніше, регресійна модель Pooled OLS часто є хорошою відправною точкою та еталонною моделлю для кількох наборів панельних даних.

Для побудови використовуємо OLS клас statsmodels для побудови та адаптації регресійної моделі OLS [29].

Для початку імпортуємо необхідні бібліотеки [21]:

```
from linearmodels import PooledOLS
import statsmodels.api as sm
```

Рисунок 3.1 – Імпорт бібліотек

Необхідно визначити залежну та незалежні змінні. Залежна змінна – це MAV, яка показує потік даних за шкідливими програмами, виявленими серед нових об'єктів у поштових додатках, незалежними змінними Data_num – показує порядок днів у місяці, так як Python не розуміє типу даних Дата, та перетворює їх в числа з 0 до кінцевого значення по порядку. Дані змінні представлені в таблиці 2.1.

Також необхідно створити dummy змінні, які будуть виступати незалежними змінними, тобто кожна країна – це окрема булева змінна та приєднуємо їх до основної бази даних:

```
df_dummies = pd.get_dummies(df[unit_col_name])
df_panel_with_dummies = df.join(df_dummies)
df_panel_with_dummies
```

	Date	Country	Date_num	Country_id	MAV	KA$	ID$	Afghanistan	Armenia	Azerbaijan	...	Togo	Tynisia	UK	Uganda	Ukraine	Unit sta
0	2022-08-14	Ukraine	1	1	4.795791	13.978904	9.912150	0	0	0	...	0	0	0	0	1	
1	2022-08-15	Ukraine	2	1	5.379897	14.321123	9.913190	0	0	0	...	0	0	0	0	1	
2	2022-08-16	Ukraine	3	1	5.493061	14.300403	9.952897	0	0	0	...	0	0	0	0	1	
3	2022-08-17	Ukraine	4	1	6.345636	14.399956	9.977481	0	0	0	...	0	0	0	0	1	
4	2022-08-18	Ukraine	5	1	6.077642	14.313859	9.908084	0	0	0	...	0	0	0	0	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1235	2022-09-09	Brazil	27	40	10.477851	17.168917	12.671064	0	0	0	...	0	0	0	0	0	
1236	2022-09-10	Brazil	28	40	8.645059	16.921710	12.578605	0	0	0	...	0	0	0	0	0	
1237	2022-09-11	Brazil	29	40	8.308966	16.747717	12.608053	0	0	0	...	0	0	0	0	0	
1238	2022-09-12	Brazil	30	40	10.256290	16.911541	12.725729	0	0	0	...	0	0	0	0	0	
1239	2022-09-13	Brazil	31	40	10.105653	16.701400	12.790911	0	0	0	...	0	0	0	0	0	

1240 rows × 47 columns

Рисунок 3.2 – Створення dummy змінних

Визначаємо залежну та незалежні змінні:

```
y = 'MAV'
x = df_panel_with_dummies.drop(['MAV', 'KAS', 'IDS', 'Date', 'Country'], axis=1)
```

Рисунок 3.3 – Вхідні змінні моделі

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    MAV   R-squared:                       0.776
Model:                            OLS   Adj. R-squared:                  0.769
Method:                 Least Squares   F-statistic:                     104.0
Date:                Sun, 27 Nov 2022   Prob (F-statistic):               0.00
Time:                        12:52:01   Log-Likelihood:                -1785.7
No. Observations:                1240   AIC:                             3653.
Df Residuals:                    1199   BIC:                             3863.
Df Model:                          40
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          6.7728      0.059    114.842      0.000       6.657       6.888
Date_num       0.0038      0.003      1.147      0.251      -0.003       0.010
Afghanistan   -1.9391      0.184    -10.527      0.000      -2.301      -1.578
Armenia       -1.6596      0.184     -9.010      0.000      -2.021      -1.298
Azerbaijan    -1.5556      0.184     -8.445      0.000      -1.917      -1.194
Brazil         2.7729      0.184     15.053      0.000       2.412       3.134
Canada         0.7021      0.184      3.812      0.000       0.341       1.064
Chile          0.8461      0.184      4.593      0.000       0.485       1.207
China          1.8793      0.184     10.202      0.000       1.518       2.241
Colombia       2.1974      0.184     11.929      0.000       1.836       2.559
Costa Rica    -0.7608      0.184     -4.130      0.000      -1.122      -0.399
Cuba          -2.4545      0.184    -13.325      0.000      -2.816      -2.093
Czech Republic -0.2331     0.184     -1.266      0.206      -0.595       0.128
Egypt          1.2235      0.184      6.642      0.000       0.862       1.585
France         0.9883      0.184      5.365      0.000       0.627       1.350
Germany        2.8093      0.184     15.251      0.000       2.448       3.171
Hungary        0.7274      0.184      3.949      0.000       0.366       1.089
India          1.5464      0.184      8.395      0.000       1.185       1.908
Indonezia      1.8612      0.184     10.104      0.000       1.500       2.223
Iran           1.2874      0.184      6.989      0.000       0.926       1.649
Italy          2.3567      0.184     12.794      0.000       1.995       2.718
Japan          1.4038      0.184      7.620      0.000       1.042       1.765
Kenya          1.2794      0.184      6.945      0.000       0.918       1.641
Mexico         3.1120      0.184     16.894      0.000       2.751       3.473
Moldova       -2.0779      0.184    -11.280      0.000      -2.439      -1.716
Paraguay      -0.1887      0.184     -1.025      0.306      -0.550       0.173
Poland         0.6754      0.184      3.667      0.000       0.314       1.037
Slovakia      -2.0631      0.184    -11.200      0.000      -2.424      -1.702
Somalia       -5.6288      0.184    -30.557      0.000      -5.990      -5.267
South Korea    0.0173      0.184      0.094      0.925      -0.344       0.379
Sudan         -0.9073      0.184     -4.925      0.000      -1.269      -0.546
Tanzania      -0.1039      0.184     -0.564      0.573      -0.465       0.257
Togo          -3.0726      0.184    -16.680      0.000      -3.434      -2.711
Tynisia        0.2607      0.184      1.415      0.157      -0.101       0.622
UK             1.6226      0.184      8.808      0.000       1.261       1.984
Uganda        -0.3444      0.184     -1.870      0.062      -0.706       0.017
Ukraine       -0.5675      0.184     -3.081      0.002      -0.929      -0.206
United States  2.2485      0.184     12.206      0.000       1.887       2.610
Venezuala     -2.6326      0.184    -14.291      0.000      -2.994      -2.271
Vietnam        2.5526      0.184     13.857      0.000       2.191       2.914
Zamdia        -1.7266      0.184     -9.373      0.000      -2.088      -1.365
Zimbabwe       0.3189      0.184      1.731      0.084      -0.042       0.680
==============================================================================
Omnibus:                      245.994   Durbin-Watson:                   1.499
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              441.868
Skew:                          -1.216   Prob(JB):                     1.12e-96
Kurtosis:                       4.623   Cond. No.                     1.72e+16
==============================================================================
```

Рисунок 3.4 – Результати побудови об’єднаної моделі для змінної MAV

Після побудови об’єднаної регресійної моделі, отримуємо наступне рівняння регресії (3.1):

$$MAV = 6.77 + 0.0038 \cdot Data_{num} - 4.83 \cdot Afganistan + 5.11 \cdot Armenia + 5.22 \cdot Azerbaijan \ldots + Zandia \cdot 5.05 + 7.09 \cdot Zimbabwe + \epsilon \quad (3.1)$$

Щоб проаналізувати, чи є об'єднана модель OLS адекватною моделлю для нашої проблеми регресії, необхідно провести аналіз таких показників R-квадрат і F-тест, логарифм правдоподібності та балів AIC, а також опосередковано через аналіз залишків.

Скоригований R-квадрат, який вимірює частку загальної дисперсії в y, яка пояснюється X після врахування ступенів свободи, втрачених через включення змінних регресії, становить 0.769 або близько 76.9 %. Це, безумовно гарний результат.

F - тест для регресії, який вимірює спільну значущість параметрів моделі, дав тестову статистику 104.0 із значенням p = 0.00, що дозволяє зробити висновок, що оцінки коефіцієнтів моделі є спільно значущими при p < 0.001 [30].

Log-правдоподібність регресійної моделі - це спосіб вимірювання користі придатності для моделі. Чим вище значення лог-ймовірності, тим краще модель підходить для набору даних [31].

Log-правдоподібність моделі становить1785.7, а показник AIC 3653. Ці значення придатності самі по собі не мають сенсу, якщо ми не порівняємо їх із показниками конкуруючої моделі [32].

Проаналізуємо залишкові похибки моделі для нормальності**,** гетероскедастичності та кореляції - трьох властивостей, які впливають на відповідність лінійної моделі.

Залишкова стандартна похибка - це міра, яка використовується для оцінки того, наскільки добре модель лінійної регресії відповідає даним [33].

```
print(pooled_olsr_model_results.resid)
```

```
0       -1.413282
1       -0.832959
2       -0.723579
3        0.125212
4       -0.146566
           ...
1235     0.829994
1236    -1.006582
1237    -1.348459
1238     0.597081
1239     0.442660
Length: 1240, dtype: float64
```

Рисунок 3.5 – Залишкові похибки моделі

```
print('Mean value of residual errors='+str(pooled_olsr_model_results.resid.mean()))
```

```
Mean value of residual errors=5.80431760003223e-15
```

Рисунок 3.6 – Середні значення похибок моделі

Це говорить нам про те, що регресійна модель прогнозує MAV із середньою похибкою близько 5.80 e-15.

3.1.2 Побудова моделі регресії фіксованих ефектів.

Для побудови моделі регресії фіксованих ефектів, необхідно створити фіктивні змінні:

```
unit_col_name= 'Country'
time_period_col_name='Date'
```

Рисунок 3.7 – Створення фіктивних змінних

Визначаємо залежну та незалежні змінні (див. рис. 3.2).

Визначаємо всі dummy змінні – країни:

```
unit_names = ['Germany','Italy','UK' ,'Poland','France' ,'Hungary' ,'Moldova', 'Slovakia', 'Afghanistan','Indonezia','Japan','Chi
```

```
lsdv_expr = y_var_name + ' ~ '
i = 0
for X_var_name in X_var_names:
    if i > 0:
        lsdv_expr = lsdv_expr + ' + ' + X_var_name
    else:
        lsdv_expr = lsdv_expr + X_var_name
    i = i + 1
for dummy_name in unit_names[:-1]:
    lsdv_expr = lsdv_expr + ' + ' + dummy_name

print('Regression expression for OLS with dummies=' + lsdv_expr)
```

```
Regression expression for OLS with dummies=MAV ~ Date_num + Germany + Italy + UK + Poland + France + Hungary + Moldova + Slovak
ia + Afghanistan + Indonezia + Japan + China + Vietnam + Armenia + Azerbaijan + Iran + India + Zamdia + Kenya + Zimbabwe + Egyp
t + Sudan + Somalia + Tynisia + Togo + Uganda + Tanzania + Canada + Colombia + Mexico + Cuba + Paraguay + Chile + Brazil
```

Рисунок 3.8 – Побудова рівняння регресії

OLS Regression Results

Dep. Variable:	MAV	R-squared:	0.710
Model:	OLS	Adj. R-squared:	0.702
Method:	Least Squares	F-statistic:	84.29
Date:	Fri, 25 Nov 2022	Prob (F-statistic):	6.31e-295
Time:	13:51:22	Log-Likelihood:	-1946.1
No. Observations:	1240	AIC:	3964.
Df Residuals:	1204	BIC:	4149.
Df Model:	35		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.4514	0.105	61.309	0.000	6.245	6.658
Date_num	0.0038	0.004	1.010	0.313	-0.004	0.011
Germany	3.1307	0.229	13.680	0.000	2.682	3.580
Italy	2.6781	0.229	11.702	0.000	2.229	3.127
UK	1.9439	0.229	8.494	0.000	1.495	2.393
Poland	0.9968	0.229	4.356	0.000	0.548	1.446
France	1.3097	0.229	5.723	0.000	0.861	1.759
Hungary	1.0487	0.229	4.583	0.000	0.600	1.498
Moldova	-1.7565	0.229	-7.675	0.000	-2.206	-1.308
Slovakia	-1.7417	0.229	-7.611	0.000	-2.191	-1.293
Afghanistan	-1.6177	0.229	-7.069	0.000	-2.067	-1.169
Indonezia	2.1826	0.229	9.537	0.000	1.734	2.632
Japan	1.7251	0.229	7.538	0.000	1.276	2.174
China	2.2007	0.229	9.616	0.000	1.752	2.650
Vietnam	2.8740	0.229	12.558	0.000	2.425	3.323
Armenia	-1.3383	0.229	-5.848	0.000	-1.787	-0.889
Azerbaijan	-1.2342	0.229	-5.393	0.000	-1.683	-0.785
Iran	1.6088	0.229	7.030	0.000	1.160	2.058
India	1.8678	0.229	8.162	0.000	1.419	2.317
Zamdia	-1.4052	0.229	-6.140	0.000	-1.854	-0.956
Kenya	1.6007	0.229	6.995	0.000	1.152	2.050
Zimbabwe	0.6403	0.229	2.798	0.005	0.191	1.089
Egypt	1.5449	0.229	6.751	0.000	1.096	1.994
Sudan	-0.5859	0.229	-2.560	0.011	-1.035	-0.137
Somalia	-5.3074	0.229	-23.192	0.000	-5.756	-4.858
Tynisia	0.5821	0.229	2.544	0.011	0.133	1.031
Togo	-2.7512	0.229	-12.022	0.000	-3.200	-2.302
Uganda	-0.0231	0.229	-0.101	0.920	-0.472	0.426
Tanzania	0.2175	0.229	0.950	0.342	-0.232	0.666
Canada	1.0235	0.229	4.472	0.000	0.575	1.472
Colombia	2.5187	0.229	11.006	0.000	2.070	2.968
Mexico	3.4334	0.229	15.003	0.000	2.984	3.882
Cuba	-2.1331	0.229	-9.321	0.000	-2.582	-1.684
Paraguay	0.1327	0.229	0.580	0.562	-0.316	0.582
Chile	1.1674	0.229	5.101	0.000	0.718	1.616
Brazil	3.0943	0.229	13.521	0.000	2.645	3.543

Omnibus:	246.665	Durbin-Watson:	1.173
Prob(Omnibus):	0.000	Jarque-Bera (JB):	603.618
Skew:	-1.067	Prob(JB):	8.43e-132
Kurtosis:	5.670	Cond. No.	303.

Рисунок 3.9 – Регресія фіксованих змінних для змінної MAV

Рівняння регресії з фіксованими змінними виглядає наступним чином (3.2):

$$MAV = 6.45 + 0.0038 \cdot Data_{num} - 3.13 \cdot Germany + 2.68 \cdot Italy + \cdots + 2.17 \cdot Cuba + 1.18 \cdot Chile + 3.09 \cdot Brazil + \epsilon \quad (3.2)$$

Скоригований R-квадрат, який вимірює частку загальної дисперсії в y, яка пояснюється X після врахування ступенів свободи, втрачених через включення змінних регресії, становить 0.702 або близько 70.2 %. Це, безумовно гарний результат.

F - тест для регресії, який вимірює спільну значущість параметрів моделі, дав тестову статистику 84.29 із значенням p = 0.00, що дозволяє зробити висновок, що оцінки коефіцієнтів моделі є спільно значущими при p < 0.001.

Log-правдоподібність моделі становить 1946.1, а показник AIC 3964. Ці значення придатності самі по собі не мають сенсу, якщо ми не порівняємо їх із показниками конкуруючої моделі.

Проаналізуємо залишкові похибки підігнаної моделі для нормальності, гетероскедастичності та кореляції - трьох властивостей, які впливають на відповідність лінійної моделі [34].

```
print(lsdv_model_results.resid)
print('Mean value of residual errors='+str(lsdv_model_results.resid.mean()))
```

```
0      -1.659387
1      -1.079064
2      -0.969684
3      -0.120892
4      -0.392670
          ...
1235    0.829994
1236   -1.006582
1237   -1.348459
1238    0.597081
1239    0.442660
Length: 1240, dtype: float64
Mean value of residual errors=4.410210548010421e-13
```

Рисунок 3.10 - Залишкові похибки моделі

3.1.3 Побудова моделі регресії випадкових ефектів.

Першим кроком для побудови моделі випадкових ефектів, необхідно розрахувати $\sigma^2\_\ \epsilon$ і $\sigma^2\_\ \mu$ - дисперсії компонентів похибки μ і ϵмоделі фіксованих ефектів та об'єднаної моделі та знайти різницю між ними.

```
sigma2_epsilon = lsdv_model_results.ssr/(n*T-(n+k+1))
print('sigma2_epsilon = ' + str(sigma2_epsilon))
```

```
sigma2_epsilon = -45.283594237287836
```

```
sigma2_pooled = pooled_olsr_model_results.ssr/(n*T-(k+1))
print('sigma2_pooled = ' + str(sigma2_pooled))
```

```
sigma2_pooled = -646.7554791838006
```

```
sigma2_u = sigma2_pooled - sigma2_epsilon
print('sigma2_u = ' + str(sigma2_u))
```

```
sigma2_u = -601.4718849465128
```

Рисунок 3.11 – Розрахунок значень дисперсії компонентів похибки моделей

Обчислюємо середні значення y та X для кожної групи (тобто кожної одиниці *i*) на панелі даних таким чином:

```
df_group_means = df_panel_with_dummies.groupby(unit_col_name).mean()

print(df_group_means)
```

Country	Date_num	Country_id	MAV	KAS	IDS
Afghanistan	16.0	11.0	4.894210	9.706546	7.178635
Armenia	16.0	17.0	5.173679	11.371538	8.767991
Azerbaijan	16.0	18.0	5.277701	11.792697	8.177036
Brazil	16.0	40.0	9.606234	16.871789	12.799997
Canada	16.0	31.0	7.535446	14.592936	11.297106
Chile	16.0	39.0	7.679380	13.193566	11.915792
China	16.0	15.0	8.712614	18.418690	14.017562
Colombia	16.0	33.0	9.030679	13.974321	11.785556
Costa Rica	16.0	37.0	6.072505	11.769029	9.744092
Cuba	16.0	36.0	4.378815	9.496146	6.163067
Czech Republic	16.0	10.0	6.600173	14.659354	10.278660
Egypt	16.0	24.0	8.056791	12.712312	11.350442
France	16.0	6.0	7.821634	16.351361	12.645950
Germany	16.0	2.0	9.642643	16.479932	12.442787
Hungary	16.0	7.0	7.560679	13.582649	9.446409
India	16.0	20.0	8.379729	16.105214	12.285693
Indonezia	16.0	12.0	8.694488	14.663417	12.816168

Рисунок 3.12 – Обчислення середніх значень

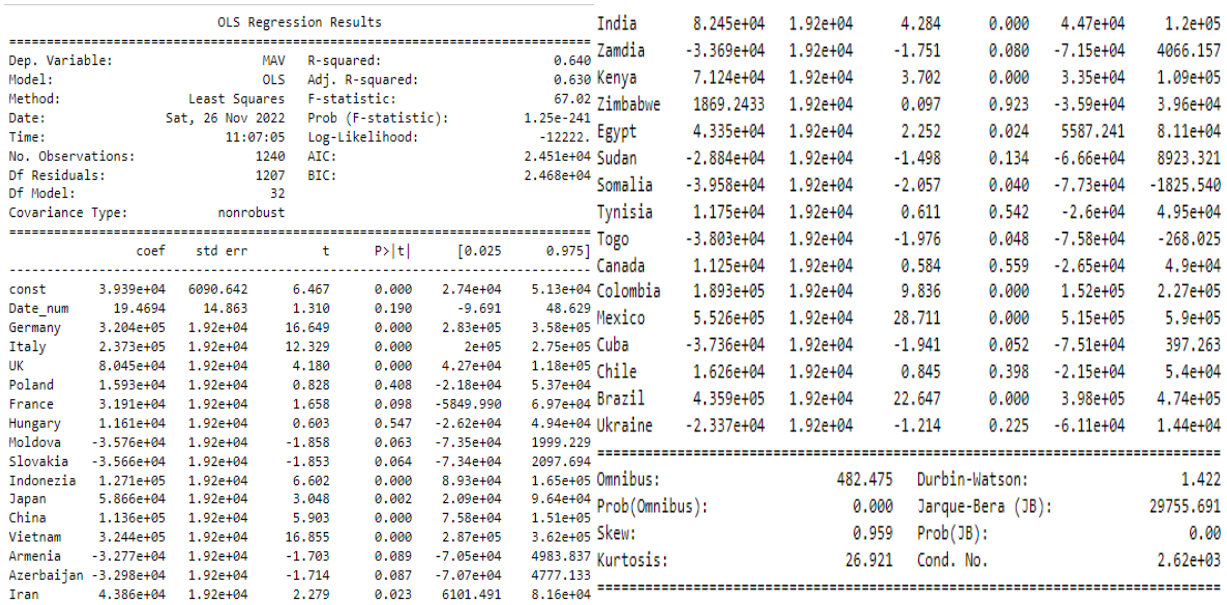
Наступним кроком є зменшення середніх значень всіх значень y та X для кожної одиниці, використовуючи масштабовану версію відповідного середнього значення для конкретної групи, обчислених на другому кроці.

```
theta = 1 - math.sqrt(c/(u))
print('theta = ' + str(theta))
```

```
theta = 0.9539504917492148
```

Рисунок. 3.13 – Обчислення показника Тета

Тепер будуємо модель випадкових ефектів (рис. 314):



OLS Regression Results

Dep. Variable:	MAV	R-squared:	0.640
Model:	OLS	Adj. R-squared:	0.630
Method:	Least Squares	F-statistic:	67.02
Date:	Sat, 26 Nov 2022	Prob (F-statistic):	1.25e-241
Time:	11:07:05	Log-Likelihood:	-12222.
No. Observations:	1240	AIC:	2.451e+04
Df Residuals:	1207	BIC:	2.468e+04
Df Model:	32		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.939e+04	6090.642	6.467	0.000	2.74e+04	5.13e+04
Date_num	19.4694	14.863	1.310	0.190	-9.691	48.629
Germany	3.204e+05	1.92e+04	16.649	0.000	2.83e+05	3.58e+05
Italy	2.373e+05	1.92e+04	12.329	0.000	2e+05	2.75e+05
UK	8.045e+04	1.92e+04	4.180	0.000	4.27e+04	1.18e+05
Poland	1.593e+04	1.92e+04	0.828	0.408	-2.18e+04	5.37e+04
France	3.191e+04	1.92e+04	1.658	0.098	-5849.990	6.97e+04
Hungary	1.161e+04	1.92e+04	0.603	0.547	-2.62e+04	4.94e+04
Moldova	-3.576e+04	1.92e+04	-1.858	0.063	-7.35e+04	1999.229
Slovakia	-3.566e+04	1.92e+04	-1.853	0.064	-7.34e+04	2097.694
Indonezia	1.271e+05	1.92e+04	6.602	0.000	8.93e+04	1.65e+05
Japan	5.866e+04	1.92e+04	3.048	0.002	2.09e+04	9.64e+04
China	1.136e+05	1.92e+04	5.903	0.000	7.58e+04	1.51e+05
Vietnam	3.244e+05	1.92e+04	16.855	0.000	2.87e+05	3.62e+05
Armenia	-3.277e+04	1.92e+04	-1.703	0.089	-7.05e+04	4983.837
Azerbaijan	-3.298e+04	1.92e+04	-1.714	0.087	-7.07e+04	4777.133
Iran	4.386e+04	1.92e+04	2.279	0.023	6101.491	8.16e+04
India	8.245e+04	1.92e+04	4.284	0.000	4.47e+04	1.2e+05
Zamdia	-3.369e+04	1.92e+04	-1.751	0.080	-7.15e+04	4066.157
Kenya	7.124e+04	1.92e+04	3.702	0.000	3.35e+04	1.09e+05
Zimbabwe	1869.2433	1.92e+04	0.097	0.923	-3.59e+04	3.96e+04
Egypt	4.335e+04	1.92e+04	2.252	0.024	5587.241	8.11e+04
Sudan	-2.884e+04	1.92e+04	-1.498	0.134	-6.66e+04	8923.321
Somalia	-3.958e+04	1.92e+04	-2.057	0.040	-7.73e+04	-1825.540
Tynisia	1.175e+04	1.92e+04	0.611	0.542	-2.6e+04	4.95e+04
Togo	-3.803e+04	1.92e+04	-1.976	0.048	-7.58e+04	-268.025
Canada	1.125e+04	1.92e+04	0.584	0.559	-2.65e+04	4.9e+04
Colombia	1.893e+05	1.92e+04	9.836	0.000	1.52e+05	2.27e+05
Mexico	5.526e+05	1.92e+04	28.711	0.000	5.15e+05	5.9e+05
Cuba	-3.736e+04	1.92e+04	-1.941	0.052	-7.51e+04	397.263
Chile	1.626e+04	1.92e+04	0.845	0.398	-2.15e+04	5.4e+04
Brazil	4.359e+05	1.92e+04	22.647	0.000	3.98e+05	4.74e+05
Ukraine	-2.337e+04	1.92e+04	-1.214	0.225	-6.11e+04	1.44e+04

Omnibus:	482.475	Durbin-Watson:	1.422
Prob(Omnibus):	0.000	Jarque-Bera (JB):	29755.691
Skew:	0.959	Prob(JB):	0.00
Kurtosis:	26.921	Cond. No.	2.62e+03

Рисунок. 3.14 – Побудова моделі регресії випадкових ефектів

$$MAV = 39390 + 19.47 \cdot Data_{num} - 3.204e + 04 \cdot Germany + 2.373t + 05 \cdot Italy + \cdots + 4.359e + 05 \cdot Brazil - 2.337\text{e} + 04 \cdot Ukraine + \epsilon \quad (3.3)$$

Обчислена дисперсія σ^2_u було оцінено як -601.47, а σ^2_ϵ було оцінено як -45.28. Таким чином, частка загальної дисперсії, яка може бути віднесена до випадкового ефекту окремої одиниці, дорівнює (3.4):

$$\frac{-601.47}{-601.47+(-45.28)} = 0{,}93 \quad (3.4)$$

Це означає, що на 93% присутній випадковий ефект у моделі, але дивлячись на показник скоригованого R, який дорівнює 0.63, можна зробити висновок, що ця модель не є кращою за модель фіксованого ефекту та об’єднану модель.

Ми можемо використовуватитест Breusch-Pagan LM для перевірки значущості випадкового ефекту [35].

Нульова гіпотеза тесту Бреуша-Пагана LM полягає в тому, що одинична дисперсія σ^2_u дорівнює нулю.

```
df_pooled_olsr_resid_with_unitnames = pd.concat([df_data[unit_col_name],pooled_olsr_model_results.resid], axis=1)
df_pooled_olsr_resid_group_means = df_pooled_olsr_resid_with_unitnames.groupby(unit_col_name).mean()
ssr_grouped_means=(df_pooled_olsr_resid_group_means[0]**2).sum()
ssr_pooled_olsr=pooled_olsr_model_results.ssr
LM_statistic = (n*T)/(2*(T-1))*math.pow(((T*T*ssr_grouped_means)/ssr_pooled_olsr - 1),2)
print('BP LM Statistic='+str(LM_statistic))
alpha=0.05
chi2_critical_value=st.chi2.ppf((1.0-alpha), 1)
print('chi2_critical_value='+str(chi2_critical_value))
```

```
BP LM Statistic=18.008298755186722
chi2_critical_value=3.841458820694124
```

Рисунок 3.15 -Перевірка значущості моделі випадкового ефекту

Тестова статистика тесту LM (18,0083) більша, ніж критичне значення Chi-squared = 3,84146 при α=.05, що означає, що випадковий ефект є значущим при альфа 0.05.

3.2 Побудова регресійних моделей для змінної «KAS».

3.2.1 Побудова об’єднаної регресійної моделі (Pooled OLS).

Необхідно визначити залежну та незалежні змінні. Залежна змінна – це KAS, яка показує потік даних з виявлених мережевих атак. Визначаємо вхідні дані моделі:

```
y = 'KAS'
x = df_panel_with_dummies.drop(['MAV', 'KAS', 'IDS', 'Date', 'Country'], axis=1)
```

Рисунок 3.16 – Вхідні змінні моделі

Статистично незначущі фактори було усунуто з моделі та модель побудували знову.

OLS Regression Results

Dep. Variable:	KAS	R-squared:	0.923
Model:	OLS	Adj. R-squared:	0.920
Method:	Least Squares	F-statistic:	357.2
Date:	Sat, 26 Nov 2022	Prob (F-statistic):	0.00
Time:	13:38:14	Log-Likelihood:	-1211.8
No. Observations:	1240	AIC:	2506.
Df Residuals:	1199	BIC:	2716.
Df Model:	40		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	13.0191	0.037	350.683	0.000	12.946	13.092
Date_num	0.0203	0.002	9.787	0.000	0.016	0.024
Afghanistan	-3.6376	0.116	-31.369	0.000	-3.865	-3.410
Armenia	-1.9726	0.116	-17.011	0.000	-2.200	-1.745
Azerbaijan	-1.5515	0.116	-13.379	0.000	-1.779	-1.324
Brazil	3.5276	0.116	30.421	0.000	3.300	3.755
Canada	1.2488	0.116	10.769	0.000	1.021	1.476
Chile	-0.1506	0.116	-1.299	0.194	-0.378	0.077
China	5.0745	0.116	43.760	0.000	4.847	5.302
Colombia	0.6301	0.116	5.434	0.000	0.403	0.858
Costa Rica	-1.5751	0.116	-13.583	0.000	-1.803	-1.348
Cuba	-3.8480	0.116	-33.184	0.000	-4.076	-3.621
Czech Republic	1.3152	0.116	11.342	0.000	1.088	1.543
Egypt	-0.6319	0.116	-5.449	0.000	-0.859	-0.404
France	3.0072	0.116	25.933	0.000	2.780	3.235
Germany	3.1358	0.116	27.041	0.000	2.908	3.363
Hungary	0.2385	0.116	2.056	0.040	0.011	0.466
India	2.7610	0.116	23.810	0.000	2.534	2.989
Indonezia	1.3192	0.116	11.377	0.000	1.092	1.547
Iran	1.1763	0.116	10.144	0.000	0.949	1.404
Italy	2.2894	0.116	19.743	0.000	2.062	2.517
Japan	3.3639	0.116	29.009	0.000	3.136	3.591
Kenya	-0.3822	0.116	-3.296	0.001	-0.610	-0.155
Mexico	1.2818	0.116	11.054	0.000	1.054	1.509
Moldova	-0.3385	0.116	-2.919	0.004	-0.566	-0.111
Paraguay	-1.0199	0.116	-8.795	0.000	-1.247	-0.792
Poland	2.3224	0.116	20.027	0.000	2.095	2.550
Slovakia	-0.5354	0.116	-4.617	0.000	-0.763	-0.308
Somalia	-2.5877	0.116	-22.315	0.000	-2.815	-2.360
South Korea	1.6567	0.116	14.287	0.000	1.429	1.884
Sudan	-2.3620	0.116	-20.369	0.000	-2.590	-2.135
Tanzania	-2.2058	0.116	-19.022	0.000	-2.433	-1.978
Togo	-2.4489	0.116	-21.118	0.000	-2.676	-2.221
Tynisia	-0.5313	0.116	-4.582	0.000	-0.759	-0.304
UK	1.8676	0.116	16.106	0.000	1.640	2.095
Uganda	-0.0887	0.116	-0.765	0.445	-0.316	0.139
Ukraine	1.2904	0.116	11.128	0.000	1.063	1.518
United States	4.4757	0.116	38.597	0.000	4.248	4.703
Venezuala	-0.5113	0.116	-4.409	0.000	-0.739	-0.284
Vietnam	1.9832	0.116	17.102	0.000	1.756	2.211
Zandia	-2.3716	0.116	-20.452	0.000	-2.599	-2.144
Zimbabwe	-2.1955	0.116	-18.933	0.000	-2.423	-1.968

Omnibus:	87.758	Durbin-Watson:	1.297
Prob(Omnibus):	0.000	Jarque-Bera (JB):	383.362
Skew:	0.139	Prob(JB):	5.68e-84
Kurtosis:	5.710	Cond. No.	1.72e+16

Рисунок 3.17 - Об’єднана регресійна модель для змінної KAS

Після побудови об’єднаної регресійної моделі, отримуємо наступне рівняння регресії (3.5):

$$KAS = 13.02 + 0.0203 \cdot Data_{num} - 3.64 \cdot Afganistan - 1.97 \cdot Armenia - 1.55 \cdot Azerbaijan + \cdots - 2.37 \cdot Zandia - 2.2 \cdot Zimbabwe + \epsilon \quad (3.5)$$

Скоригований R-квадрат, який вимірює частку загальної дисперсії в y, яка пояснюється X після врахування ступенів свободи, втрачених через включення змінних регресії, становить 0.920 або близько 92.0 %. Це, безумовно гарний результат.

F - тест для регресії, який вимірює спільну значущість параметрів моделі, дав тестову статистику 357.2 із значенням p = 0.00, що дозволяє зробити висновок, що оцінки коефіцієнтів моделі є спільно значущими при p < 0.001.

Log-правдоподібність моделі становить -1211.8, а показник AIC становить 2506.

Проаналізуємо залишкові похибки :

```
print(pooled_olsr_model_results.resid)
print('Mean value of residual errors='+str(pooled_olsr_model_results.resid.mean()))
```

```
0      -0.350885
1      -0.028983
2      -0.070020
3       0.009218
4      -0.097199
          ...
1235    0.073640
1236   -0.193884
1237   -0.388195
1238   -0.244688
1239   -0.475146
Length: 1240, dtype: float64
Mean value of residual errors=-6.661767911502407e-14
```

Рисунок 3.18 – Залишкові похибки моделі

Це говорить нам про те, що регресійна модель прогнозує KAS із середньою похибкою близько -6.66e-14.

3.2.2 Побудова моделі регресії фіксованих ефектів

Визначаємо залежну та незалежні змінні:

```
y_var_name = 'KAS'
X_var_names = df.drop(['MAV', 'KAS', 'IDS', 'Date', 'Country'], axis=1)
```

Рисунок 3.19 - Визначення вхідних даних моделі

Визначаємо всі країни, які будуть незалежними змінними, які впливають на залежну та будуємо рівняння регресії:

```
lsdv_expr = y_var_name + ' ~ '
i = 0
for X_var_name in X_var_names:
    if i > 0:
        lsdv_expr = lsdv_expr + ' + ' + X_var_name
    else:
        lsdv_expr = lsdv_expr + X_var_name
    i = i + 1
for dummy_name in unit_names[:-1]:
    lsdv_expr = lsdv_expr + ' + ' + dummy_name

print('Regression expression for OLS with dummies=' + lsdv_expr)
```

```
Regression expression for OLS with dummies=KAS ~ Date_num + Germany + Italy + UK + Poland + France + Hungary + Moldova + Slovak
ia + Afghanistan + Indonezia + Japan + China + Vietnam + Armenia + Azerbaijan + Iran + India + Zamdia + Kenya + Zimbabwe + Egyp
t + Sudan + Somalia + Tynisia + Togo + Uganda + Tanzania + Canada + Colombia + Mexico + Cuba + Paraguay + Chile + Brazil
```

Рисунок 3.20 – Побудова рівняння регресії

OLS Regression Results

Dep. Variable:	KAS	R-squared:	0.822
Model:	OLS	Adj. R-squared:	0.817
Method:	Least Squares	F-statistic:	158.6
Date:	Wed, 23 Nov 2022	Prob (F-statistic):	0.00
Time:	14:20:08	Log-Likelihood:	-1728.9
No. Observations:	1240	AIC:	3530.
Df Residuals:	1204	BIC:	3714.
Df Model:	35		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	14.1277	0.088	159.958	0.000	13.954	14.301
Date_num	0.0203	0.003	6.463	0.000	0.014	0.026
Germany	2.0272	0.192	10.554	0.000	1.650	2.404
Italy	1.1808	0.192	6.147	0.000	0.804	1.558
UK	0.7590	0.192	3.952	0.000	0.382	1.136
Poland	1.2138	0.192	6.319	0.000	0.837	1.591
France	1.8986	0.192	9.884	0.000	1.522	2.275
Hungary	-0.8701	0.192	-4.530	0.000	-1.247	-0.493
Moldova	-1.4471	0.192	-7.534	0.000	-1.824	-1.070
Slovakia	-1.6440	0.192	-8.559	0.000	-2.021	-1.267
Afghanistan	-4.7462	0.192	-24.709	0.000	-5.123	-4.369
Indonezia	0.2106	0.192	1.097	0.273	-0.166	0.587
Japan	2.2553	0.192	11.741	0.000	1.878	2.632
China	3.9659	0.192	20.647	0.000	3.589	4.343
Vietnam	0.8746	0.192	4.553	0.000	0.498	1.251
Armenia	-3.0812	0.192	-16.041	0.000	-3.458	-2.704
Azerbaijan	-2.6601	0.192	-13.849	0.000	-3.037	-2.283
Iran	0.0677	0.192	0.352	0.725	-0.309	0.445
India	1.6524	0.192	8.603	0.000	1.276	2.029
Zamdia	-3.4802	0.192	-18.118	0.000	-3.857	-3.103
Kenya	-1.4908	0.192	-7.761	0.000	-1.868	-1.114
Zimbabwe	-3.3041	0.192	-17.202	0.000	-3.681	-2.927
Egypt	-1.7405	0.192	-9.061	0.000	-2.117	-1.364
Sudan	-3.4706	0.192	-18.068	0.000	-3.847	-3.094
Somalia	-3.6962	0.192	-19.243	0.000	-4.073	-3.319
Tynisia	-1.6399	0.192	-8.537	0.000	-2.017	-1.263
Togo	-3.5575	0.192	-18.521	0.000	-3.934	-3.181
Uganda	-1.1973	0.192	-6.233	0.000	-1.574	-0.820
Tanzania	-3.3143	0.192	-17.255	0.000	-3.691	-2.937
Canada	0.1402	0.192	0.730	0.466	-0.237	0.517
Colombia	-0.4784	0.192	-2.491	0.013	-0.855	-0.102
Mexico	0.1732	0.192	0.902	0.367	-0.204	0.550
Cuba	-4.9566	0.192	-25.805	0.000	-5.333	-4.580
Paraguay	-2.1285	0.192	-11.081	0.000	-2.505	-1.752
Chile	-1.2592	0.192	-6.556	0.000	-1.636	-0.882
Brazil	2.4190	0.192	12.594	0.000	2.042	2.796

Omnibus:	150.602	Durbin-Watson:	0.605
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1164.189
Skew:	0.256	Prob(JB):	1.58e-253
Kurtosis:	7.719	Cond. No.	303.

Рисунок 3.24 – Регресія фіксованих змінних для змінної IDS

Рівняння регресії з фіксованими змінними виглядає наступним чином (3.6):

$$KAS = 14.13 + 0.0203 \cdot Data_{num} - 2.03 \cdot Germany - 1.18 \cdot Italy - \\ -1.25 \cdot Chile + \cdots + 2.42 \cdot Brazil + \epsilon \quad (3.6)$$

Далі розглянемо коефіцієнти для цікавих фіктивних змінних, що представляють вплив на конкретну країну. Ми спостерігаємо, що відрізок регресії, який представляє специфічний для країни ефект для України (пропущена змінна), становить 0.822 і є статистично значущим (це означає, що його значення для населення оцінюється як відмінне від нуля), при p-значенні 0,000 .

Скориговане значення R-квадрат дорівнює 0.817, або 81.7% - значення показує дуже хорошу відповідність між незалежними змінними та залежною.

3.2.3 Побудова моделі регресії випадкових ефектів.

Першим кроком для побудови моделі випадкових ефектів, необхідно розрахувати $\sigma^2\_\ \epsilon$ і $\sigma^2\_\ \mu$ - дисперсії компонентів похибки μ і ϵмоделі фіксованих ефектів та об'єднаної моделі та знайти різницю між ними.

```
sigma2_epsilon = lsdv_model_results.ssr/(n*T-(n+k+1))
print('sigma2_epsilon = ' + str(sigma2_epsilon))
```

```
sigma2_epsilon = -31.901267130742646
```

```
sigma2_pooled = pooled_olsr_model_results.ssr/(n*T-(k+1))
print('sigma2_pooled = ' + str(sigma2_pooled))
```

```
sigma2_pooled = -256.29655163768166
```

```
sigma2_u = sigma2_pooled - sigma2_epsilon
print('sigma2_u = ' + str(sigma2_u))
```

```
sigma2_u = -224.39528450693902
```

Рисунок 3.25 – Розрахунок значень дисперсії компонентів похибки моделей

Обчислюємо середні значення y та X для кожної групи (тобто кожної одиниці *i*) на панелі даних таким чином та розраховуємо показник Тета:

```
theta = 1 - math.sqrt(c/(u))
print('theta = ' + str(theta))
```

```
theta = 0.9367805559430118
```

Рисунок 3.26 – Обчислення показника Тета

Тепер будуємо модель випадкових ефектів(рис. 3.27)

OLS Regression Results

Dep. Variable:	KAS	R-squared:	0.662
Model:	OLS	Adj. R-squared:	0.653
Method:	Least Squares	F-statistic:	73.98
Date:	Sat, 26 Nov 2022	Prob (F-statistic):	2.80e-258
Time:	13:41:40	Log-Likelihood:	-21938.
No. Observations:	1240	AIC:	4.394e+04
Df Residuals:	1207	BIC:	4.411e+04
Df Model:	32		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.168e+08	1.12e+07	10.414	0.000	9.48e+07	1.39e+08
Date_num	8.201e+04	3.76e+04	2.183	0.029	8321.468	1.56e+05
UK	-4.902e+07	3.54e+07	-1.384	0.167	-1.19e+08	2.05e+07
Germany	1.259e+08	3.54e+07	3.555	0.000	5.64e+07	1.95e+08
Italy	-9.15e+06	3.54e+07	-0.258	0.796	-7.87e+07	6.04e+07
Poland	-5.396e+06	3.54e+07	-0.152	0.879	-7.49e+07	6.41e+07
France	2.023e+08	3.54e+07	5.711	0.000	1.33e+08	2.72e+08
Hungary	-9.153e+07	3.54e+07	-2.584	0.010	-1.61e+08	-2.2e+07
Moldova	-1.096e+08	3.54e+07	-3.094	0.002	-1.79e+08	-4.01e+07
Slovakia	-1.115e+08	3.54e+07	-3.148	0.002	-1.81e+08	-4.2e+07
Indonezia	-7.33e+07	3.54e+07	-2.069	0.039	-1.43e+08	-3.8e+06
Japan	1.699e+08	3.54e+07	4.795	0.000	1e+08	2.39e+08
China	1.511e+09	3.54e+07	42.647	0.000	1.44e+09	1.58e+09
Vietnam	-3.76e+07	3.54e+07	-1.061	0.289	-1.07e+08	3.19e+07
Armenia	-1.163e+08	3.54e+07	-3.284	0.001	-1.86e+08	-4.68e+07
Azerbaijan	-1.159e+08	3.54e+07	-3.270	0.001	-1.85e+08	-4.64e+07
Iran	-7.528e+07	3.54e+07	-2.125	0.034	-1.45e+08	-5.77e+06
India	5.721e+07	3.54e+07	1.615	0.107	-1.23e+07	1.27e+08
Zamdia	-1.17e+08	3.54e+07	-3.303	0.001	-1.86e+08	-4.75e+07
Kenya	-1.094e+08	3.54e+07	-3.089	0.002	-1.79e+08	-3.99e+07
Zimbabwe	-1.168e+08	3.54e+07	-3.298	0.001	-1.86e+08	-4.73e+07
Egypt	-1.107e+08	3.54e+07	-3.126	0.002	-1.8e+08	-4.12e+07
Sudan	-1.169e+08	3.54e+07	-3.299	0.001	-1.86e+08	-4.74e+07
Somalia	-1.171e+08	3.54e+07	-3.307	0.001	-1.87e+08	-4.76e+07
Tynisia	-1.103e+08	3.54e+07	-3.114	0.002	-1.8e+08	-4.08e+07
Togo	-1.17e+08	3.54e+07	-3.303	0.001	-1.87e+08	-4.75e+07
Canada	-6.731e+07	3.54e+07	-1.900	0.058	-1.37e+08	2.19e+06
Colombia	-9.538e+07	3.54e+07	-2.693	0.007	-1.65e+08	-2.59e+07
Mexico	-5.944e+07	3.54e+07	-1.678	0.094	-1.29e+08	1.01e+07
Cuba	-1.179e+08	3.54e+07	-3.327	0.001	-1.87e+08	-4.84e+07
Chile	-1.084e+08	3.54e+07	-3.061	0.002	-1.78e+08	-3.89e+07
Brazil	2.629e+08	3.54e+07	7.421	0.000	1.93e+08	3.32e+08
Ukraine	-7.687e+07	3.54e+07	-2.170	0.030	-1.46e+08	-7.37e+06

Omnibus:	1386.189	Durbin-Watson:	0.725
Prob(Omnibus):	0.000	Jarque-Bera (JB):	118570.476
Skew:	5.516	Prob(JB):	0.00
Kurtosis:	49.618	Cond. No.	1.92e+03

Рисунок 3.27 – Побудова моделі регресії випадкових ефектів

Рівняння регресії з випадковим ефектом виглядає наступним чином (3.7):

$$KAS = 1.168\text{e}+08 + 8.201\text{e}+04 \cdot Data_{num} - 4.902\text{e}+07 \cdot UK1.+259\text{e} + 08 \cdot Germany + \cdots - 7.687 \cdot Ukraine + \epsilon \quad (3.7)$$

Обчислена дисперсія σ^2_u було оцінено як - -224.4, а σ^2_ϵ було оцінено як - 31.9. Таким чином, частка загальної дисперсії, яка може бути віднесена до випадкового ефекту окремої одиниці, дорівнює (3.8):

$$\frac{-224.4}{-224.7+(-31.9)} = 0.87 \quad (3.8)$$

Це означає, що на 87% присутній випадковий ефект у моделі, але дивлячись на показник скоригованого R, який дорівнює 0.653, можна зробити висновок, що ця модель не є кращою за об'єднану модель та модель фіксованого ефекту.

Тестова статистика тесту LM (18.0083) більша, ніж критичне значення Chi-squared = 3,84146 при α=.05, що означає, що випадковий ефект є значущим при альфа 0.05.

3.3 Побудова регресійних моделей для змінної «IDS»

3.3.1 Побудова об'єднаної регресійної моделі (Pooled OLS).

Необхідно визначити залежну та незалежні змінні.

Залежна змінна – це IDS, яка показує потік даних з виявлених мережевих атак. Визначаємо вхідні дані моделі:

```
y = 'LN_IDS'
x = df_data.drop(['MAV', 'KAS', 'IDS', 'Date', 'Country','LN_IDS'], axis=1)
```

Рисунок 3.28– Вхідні змінні моделі

OLS Regression Results

Dep. Variable:	LN_IDS	R-squared:	0.959
Model:	OLS	Adj. R-squared:	0.958
Method:	Least Squares	F-statistic:	703.9
Date:	Sat, 26 Nov 2022	Prob (F-statistic):	0.00
Time:	12:22:28	Log-Likelihood:	-787.96
No. Observations:	1240	AIC:	1658.
Df Residuals:	1199	BIC:	1868.
Df Model:	40		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Date_num	0.0022	0.001	1.479	0.139	-0.001	0.005
Afghanistan	7.1437	0.087	82.387	0.000	6.974	7.314
Armenia	8.7331	0.087	100.716	0.000	8.563	8.903
Azerbaijan	8.1421	0.087	93.901	0.000	7.972	8.312
Brazil	12.7651	0.087	147.216	0.000	12.595	12.935
Canada	11.2622	0.087	129.884	0.000	11.092	11.432
Chile	11.8809	0.087	137.019	0.000	11.711	12.051
China	13.9827	0.087	161.258	0.000	13.813	14.153
Colombia	11.7506	0.087	135.517	0.000	11.581	11.921
Costa Rica	9.7092	0.087	111.973	0.000	9.539	9.879
Cuba	6.1282	0.087	70.674	0.000	5.958	6.298
Czech Republic	10.2438	0.087	118.138	0.000	10.074	10.414
Egypt	11.3155	0.087	130.499	0.000	11.145	11.486
France	12.6110	0.087	145.440	0.000	12.441	12.781
Germany	12.4079	0.087	143.097	0.000	12.238	12.578
Hungary	9.4115	0.087	108.540	0.000	9.241	9.582
India	12.2508	0.087	141.285	0.000	12.081	12.421
Indonezia	12.7813	0.087	147.403	0.000	12.611	12.951
Iran	11.7913	0.087	135.985	0.000	11.621	11.961
Italy	12.2955	0.087	141.800	0.000	12.125	12.466
Japan	9.6814	0.087	111.653	0.000	9.511	9.851
Kenya	10.5971	0.087	122.213	0.000	10.427	10.767
Mexico	12.9486	0.087	149.332	0.000	12.778	13.119
Moldova	8.9293	0.087	102.980	0.000	8.759	9.099
Paraguay	9.1457	0.087	105.475	0.000	8.976	9.316
Poland	11.6987	0.087	134.918	0.000	11.529	11.869
Slovakia	12.0494	0.087	138.963	0.000	11.879	12.220
Somalia	4.3781	0.087	50.491	0.000	4.208	4.548
South Korea	11.2754	0.087	130.036	0.000	11.105	11.446
Sudan	10.6387	0.087	122.693	0.000	10.469	10.809
Tanzania	9.3409	0.087	107.726	0.000	9.171	9.511
Togo	5.7743	0.087	66.593	0.000	5.604	5.944
Tynisia	10.6471	0.087	122.791	0.000	10.477	10.817
UK	11.0886	0.087	127.882	0.000	10.919	11.259
Uganda	8.6322	0.087	99.553	0.000	8.462	8.802
Ukraine	9.7300	0.087	112.214	0.000	9.560	9.900
United States	13.2521	0.087	152.833	0.000	13.082	13.422
Venezuala	10.2241	0.087	117.912	0.000	10.054	10.394
Vietnam	12.8515	0.087	148.213	0.000	12.681	13.022
Zamdia	6.9104	0.087	79.696	0.000	6.740	7.081
Zimbabwe	8.1591	0.087	94.097	0.000	7.989	8.329

Omnibus:	599.721	Durbin-Watson:	1.077
Prob(Omnibus):	0.000	Jarque-Bera (JB):	14763.177
Skew:	-1.692	Prob(JB):	0.00
Kurtosis:	19.562	Cond. No.	238.

Рисунок 3.29 - Об’єднана регресійна модель для змінної IDS

Після побудови об’єднаної регресійної моделі, отримуємо наступне рівняння регресії (3.9):

$$IDS = 00.22 + 7.14 \cdot Afganistan + 8.73 \cdot Armenia + 6.91 \cdot Zandia + \quad (3.9)$$
$$+8.16 \cdot Zimbabwe\ 6 + \epsilon$$

Скоригований R-квадрат становить 0.958 або 95.8 %. Це, безумовно гарний результат.

F - тест для регресії дав тестову статистику 703.9 із значенням p 0.00, що дозволяє зробити висновок, що оцінки коефіцієнтів моделі є значущими при p < 0.001. оцінки коефіцієнтів моделі є спільно значущими при p < 0,001.

Log-правдоподібність моделі становить -787.96, а показник AIC 1658.

Проаналізуємо залишкові похибки моделі для нормальності, гетероскедастичності та кореляції - трьох властивостей, які впливають на відповідність лінійної моделі.

```
print(pooled_olsr_model_results.resid)
print('Mean value of residual errors='+str(pooled_olsr_model_results.resid.mean()))
```

```
0       0.179944
1       0.178803
2       0.216327
3       0.238730
4       0.165151
          ...
1235   -0.152933
1236   -0.247574
1237   -0.222308
1238   -0.104814
1239   -0.041814
Length: 1240, dtype: float64
Mean value of residual errors=6.778448770025915e-15
```

Рисунок 3.30 – Залишкові похибки моделі

3.3.2 Побудова моделі регресії фіксованих ефектів.

Визначаємо залежну та незалежні змінні:

```
y_var_name = 'LN_IDS'
X_var_names = df.drop(['MAV', 'KAS', 'IDS', 'Date', 'Country', 'Country_id'], axis=1)
```

Рисунок 3.31 - Визначення вхідних даних моделі

Визначаємо всі країни, які будуть незалежними змінними, які впливають на залежну та будуємо рівняння регресії:

```
lsdv_expr = y_var_name + ' ~ '
i = 0
for X_var_name in X_var_names:
    if i > 0:
        lsdv_expr = lsdv_expr + ' + ' + X_var_name
    else:
        lsdv_expr = lsdv_expr + X_var_name
    i = i + 1
for dummy_name in unit_names[:-1]:
    lsdv_expr = lsdv_expr + ' + ' + dummy_name

print('Regression expression for OLS with dummies=' + lsdv_expr)
```

```
Regression expression for OLS with dummies=LN_IDS ~ Date_num + Germany + Italy + UK + Poland + France + Hungary + Moldova + Slo
vakia + Afghanistan + Indonezia + Japan + China + Vietnam + Armenia + Azerbaijan + Iran + India + Zamdia + Kenya + Zimbabwe + E
gypt + Sudan + Somalia + Tynisia + Togo + Uganda + Tanzania + Canada + Colombia + Mexico + Cuba + Paraguay + Chile + Brazil
```

Рисунок 3.32 – Побудова рівняння регресії

OLS Regression Results

Dep. Variable:	LN_IDS	R-squared:	0.914
Model:	OLS	Adj. R-squared:	0.912
Method:	Least Squares	F-statistic:	366.4
Date:	Sat, 26 Nov 2022	Prob (F-statistic):	0.00
Time:	12:38:52	Log-Likelihood:	-1248.3
No. Observations:	1240	AIC:	2569.
Df Residuals:	1204	BIC:	2753.
Df Model:	35		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	10.7391	0.060	179.151	0.000	10.621	10.857
Date_num	0.0022	0.002	1.023	0.307	-0.002	0.006
Germany	1.6688	0.130	12.801	0.000	1.413	1.925
Italy	1.5564	0.130	11.938	0.000	1.301	1.812
UK	0.3496	0.130	2.681	0.007	0.094	0.605
Poland	0.9596	0.130	7.361	0.000	0.704	1.215
France	1.8720	0.130	14.359	0.000	1.616	2.128
Hungary	-1.3276	0.130	-10.183	0.000	-1.583	-1.072
Moldova	-1.8097	0.130	-13.882	0.000	-2.066	-1.554
Slovakia	1.3103	0.130	10.051	0.000	1.055	1.566
Afghanistan	-3.5954	0.130	-27.579	0.000	-3.851	-3.340
Indonezia	2.0422	0.130	15.665	0.000	1.786	2.298
Japan	-1.0577	0.130	-8.113	0.000	-1.313	-0.802
China	3.2436	0.130	24.880	0.000	2.988	3.499
Vietnam	2.1125	0.130	16.204	0.000	1.857	2.368
Armenia	-2.0060	0.130	-15.387	0.000	-2.262	-1.750
Azerbaijan	-2.5970	0.130	-19.920	0.000	-2.853	-2.341
Iran	1.0522	0.130	8.071	0.000	0.796	1.308
India	1.5117	0.130	11.596	0.000	1.256	1.767
Zamdia	-3.8287	0.130	-29.368	0.000	-4.084	-3.573
Kenya	-0.1420	0.130	-1.089	0.276	-0.398	0.114
Zimbabwe	-2.5799	0.130	-19.790	0.000	-2.836	-2.324
Egypt	0.5764	0.130	4.422	0.000	0.321	0.832
Sudan	-0.1004	0.130	-0.770	0.441	-0.356	0.155
Somalia	-6.3610	0.130	-48.793	0.000	-6.617	-6.105
Tynisia	-0.0919	0.130	-0.705	0.481	-0.348	0.164
Togo	-4.9648	0.130	-38.083	0.000	-5.221	-4.709
Uganda	-2.1069	0.130	-16.161	0.000	-2.363	-1.851
Tanzania	-1.3982	0.130	-10.725	0.000	-1.654	-1.142
Canada	0.5231	0.130	4.013	0.000	0.267	0.779
Colombia	1.0116	0.130	7.759	0.000	0.756	1.267
Mexico	2.2095	0.130	16.948	0.000	1.954	2.465
Cuba	-4.6109	0.130	-35.369	0.000	-4.867	-4.355
Paraguay	-1.5934	0.130	-12.222	0.000	-1.849	-1.338
Chile	1.1418	0.130	8.758	0.000	0.886	1.398
Brazil	2.0260	0.130	15.541	0.000	1.770	2.282

Omnibus:	251.203	Durbin-Watson:	0.548
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2683.788
Skew:	0.608	Prob(JB):	0.00
Kurtosis:	10.104	Cond. No.	303.

Рисунок 3.33 – Регресія фіксованих змінних для змінної IDS

Рівняння регресії з фіксованими змінними виглядає наступним чином (3.10):

$$IDS = 10.74 + 00.22 \cdot Germany + 1.53 \cdot Italy + \ldots + 1.14 \cdot Chile + 2.03 \cdot Brazil + \epsilon \quad (3.10)$$

Скориговане значення R-квадрат дорівнює 0.912, або 91.2% - значення показує дуже хорошу відповідність між незалежними змінними та залежною.

F - тест для регресійного аналізу перевіряє, чи всі коефіцієнти моделі є спільно значущими, і, отже, чи відповідність моделі FE краща, ніж у попередній моделі. Статистика F-тесту 366.4 є значною при p < 0.00, що означає, що відповідність моделі справді краща, ніж в об'єднаній регресійній моделі.

3.3.3 Побудова моделі регресії випадкових ефектів.

Першим кроком для побудови моделі випадкових ефектів, необхідно розрахувати $\sigma^2\_\ \epsilon$ і $\sigma^2\_\ \mu$ - дисперсії компонентів похибки μ і ϵ моделі фіксованих ефектів та об'єднаної моделі та знайти різницю між ними.

```
sigma2_epsilon = lsdv_model_results.ssr/(n*T-(n+k+1))
print('sigma2_epsilon = ' + str(sigma2_epsilon))
```

```
sigma2_epsilon = -14.695039254289002
```

```
sigma2_pooled = pooled_olsr_model_results.ssr/(n*T-(k+1))
print('sigma2_pooled = ' + str(sigma2_pooled))
```

```
sigma2_pooled = -129.37891349913338
```

```
sigma2_u = sigma2_pooled - sigma2_epsilon
print('sigma2_u = ' + str(sigma2_u))
```

```
sigma2_u = -114.68387424484438
```

Рис. 3.36 – Розрахунок значень дисперсії компонентів похибки моделей

Обчислюємо середні значення y та X для кожної групи (тобто кожної одиниці *i*) на панелі даних таким чином та розраховуємо показник Тета:

```
theta = 1 - math.sqrt(c/(u))
print('theta = ' + str(theta))
```

```
theta = 0.939969296454986
```

Рис. 3.37 – Обчислення показника Тета

Тепер будуємо модель випадкових ефектів(рис. 3.38):

OLS Regression Results

Dep. Variable:	IDS	R-squared:	0.741
Model:	OLS	Adj. R-squared:	0.734
Method:	Least Squares	F-statistic:	107.7
Date:	Sat, 26 Nov 2022	Prob (F-statistic):	0.00
Time:	13:37:43	Log-Likelihood:	-16342.
No. Observations:	1240	AIC:	3.275e+04
Df Residuals:	1207	BIC:	3.292e+04
Df Model:	32		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.451e+06	1.3e+05	11.194	0.000	1.2e+06	1.71e+06
Date_num	249.9978	412.073	0.607	0.544	-558.462	1058.458
Germany	2.798e+06	4.09e+05	6.835	0.000	1.99e+06	3.6e+06
Italy	2.332e+06	4.09e+05	5.698	0.000	1.53e+06	3.14e+06
UK	-3.164e+05	4.09e+05	-0.773	0.440	-1.12e+06	4.87e+05
Poland	7.927e+05	4.09e+05	1.937	0.053	-1.04e+04	1.6e+06
France	3.732e+06	4.09e+05	9.117	0.000	2.93e+06	4.53e+06
Hungary	-1.241e+06	4.09e+05	-3.032	0.002	-2.04e+06	-4.38e+05
Moldova	-1.318e+06	4.09e+05	-3.219	0.001	-2.12e+06	-5.14e+05
Slovakia	4.456e+06	4.09e+05	10.886	0.000	3.65e+06	5.26e+06
Indonezia	4.802e+06	4.09e+05	11.733	0.000	4e+06	5.61e+06
Japan	-1.174e+06	4.09e+05	-2.868	0.004	-1.98e+06	-3.71e+05
China	1.913e+07	4.09e+05	46.746	0.000	1.83e+07	1.99e+07
Vietnam	5.699e+06	4.09e+05	13.924	0.000	4.9e+06	6.5e+06
Armenia	-1.341e+06	4.09e+05	-3.275	0.001	-2.14e+06	-5.38e+05
Azerbaijan	-1.383e+06	4.09e+05	-3.379	0.001	-2.19e+06	-5.8e+05
Iran	8.704e+05	4.09e+05	2.126	0.034	6.73e+04	1.67e+06
India	2.23e+06	4.09e+05	5.449	0.000	1.43e+06	3.03e+06
Zamdia	-1.436e+06	4.09e+05	-3.508	0.000	-2.24e+06	-6.33e+05
Kenya	-5.993e+05	4.09e+05	-1.464	0.143	-1.4e+06	2.04e+05
Zimbabwe	-1.392e+06	4.09e+05	-3.402	0.001	-2.2e+06	-5.89e+05
Egypt	-6192.2912	4.09e+05	-0.015	0.988	-8.09e+05	7.97e+05
Sudan	-7.24e+05	4.09e+05	-1.769	0.077	-1.53e+06	7.9e+04
Somalia	-1.452e+06	4.09e+05	-3.548	0.000	-2.26e+06	-6.49e+05
Tynisia	-7.084e+05	4.09e+05	-1.731	0.084	-1.51e+06	9.47e+04
Togo	-1.448e+06	4.09e+05	-3.537	0.000	-2.25e+06	-6.45e+05
Canada	-1.019e+05	4.09e+05	-0.249	0.803	-9.05e+05	7.01e+05
Colombia	1.092e+06	4.09e+05	2.668	0.008	2.89e+05	1.89e+06
Mexico	6.769e+06	4.09e+05	16.538	0.000	5.97e+06	7.57e+06
Cuba	-1.446e+06	4.09e+05	-3.532	0.000	-2.25e+06	-6.43e+05
Chile	1.065e+06	4.09e+05	2.602	0.009	2.62e+05	1.87e+06
Brazil	4.64e+06	4.09e+05	11.336	0.000	3.84e+06	5.44e+06
Ukraine	-1.161e+06	4.09e+05	-2.836	0.005	-1.96e+06	-3.58e+05

Omnibus:	1540.323	Durbin-Watson:	0.882
Prob(Omnibus):	0.000	Jarque-Bera (JB):	307346.453
Skew:	6.282	Prob(JB):	0.00
Kurtosis:	79.097	Cond. No.	2.02e+03

Рисунок. 3.38 – Побудова моделі регресії випадкових ефектів

Рівняння регресії з випадковим ефектом виглядає наступним чином (3.11):

$$IDS = 1.451\text{e} + 06 + 249.998 \cdot Data_{num} + 2.798t + 06 \cdot Germany + 3.322t + 06 \cdot Italy + \ldots + 4.64t + 06 \cdot Chile - 1.16t + 06 \cdot Ukraine + \epsilon \quad (3.11)$$

Обчислена дисперсія σ^2_u було оцінено як - -114.68, а σ^2_ϵ було оцінено як - 14.70. Таким чином, частка загальної дисперсії, яка може бути віднесена до випадкового ефекту окремої одиниці, дорівнює (3.12):

$$\frac{-114.68}{-114.68+(-14.70)} = 0{,}89 \quad (3.12)$$

Це означає, що на 89% присутній випадковий ефект у моделі, але дивлячись на показник скоригованого R, який дорівнює 0.734, можна зробити висновок, що ця модель не є кращою за об’єднану модель та модель фіксованого ефекту.

3.4 Оцінка отриманих результатів.

Оцінемо отримані результати побудованих регресій для кожної залежної змінної.

Таблиця 3.1 – Порівняння результатів побудованих моделей для MAV

Показник	Pooled OLS Regression Model	The Fixed Effects Regression Model	The Random Effects Regression Model
Скоригований R^2	0.769	0.702	0.640
Log-Likelihood	-1785.7	-1946.1	-12222
AIC	3653	3964	24510
MRE	5.804e-15	4.410e-13	1.017e-10

Скоригований R-квадрат об’єднаної моделі (Pooled OLS) дорівнює 76.9% значно кращий порівняно з моделями фіксованого та випадкового ефекту.

Pooled OLS також забезпечує невелике збільшення логарифмічної ймовірності до -1785.7 порівняно з іншими моделями, я також показник AIC теж показав кращий результат. Середні похибка залишків теж являються найменшою для об’єднаної

Можна зробити висновок, що об’єднана модель є найкращою для опису залежної змінної MAV.

Таблиця 3.2 – Порівняння результатів побудованих моделей для KAS

Показник	Pooled OLS Regression Model	The Fixed Effects Regression Model	The Random Effects Regression Model
Скоригований R^2	0.920	0.817	0.653
Log-Likelihood	-1277.8	-1728.9	-21938
AIC	2506	3530	43940
MRE	-6.66e-14	8.52e-13	3.92e-08

Скоригований R-квадрат об’єднаної моделі (Pooled OLS) дорівнює 92.0% значно кращий порівняно з моделями фіксованого та випадкового ефекту.

Pooled OLS також забезпечує невелике збільшення логарифмічної ймовірності до -1277.8 порівняно з іншими моделями, я також показник AIC теж показав кращий результат. Середні похибка залишків теж являються найменшою для об’єднаної

Можна зробити висновок, що об’єднана модель є найкращою для опису залежної змінної KAS.

Таблиця 3.3 – Порівняння результатів побудованих моделей для IDS

Показник	Pooled OLS Regression Model	The Fixed Effects Regression Model	The Random Effects Regression Model
Скоригований R^2	0.958	0.912	0.734
Log-Likelihood	-787.96	-1248.3	-16342
AIC	1658	2569	32750
MRE	6.77e-15	6.51e-13	1.41e-08

Скоригований R-квадрат об’єднаної моделі (Pooled OLS) дорівнює 95.8% значно кращий порівняно з моделями фіксованого та випадкового ефекту.

Pooled OLS також забезпечує невелике збільшення логарифмічної ймовірності до -787.96 порівняно з іншими моделями, я також показник AIC теж показав кращий результат. Середні похибка залишків теж являються найменшою для об’єднаної

Можна зробити висновок, що об’єднана модель є найкращою для опису залежної змінної IDS.

3.5 Прогнозування трендів кібератак

3.5.1 Прогнозування трендів кібератаки виду MAV

Визначивши найкращу модель для моделювання виду кібератаки MAV, спрогнозуємо тренд за допомогою об’єднаної моделі (Pooled model) та LSTM моделі.

Перш ніж побудувати прогноз необхідно поділити базу даних на дві частини - тестову та тренувальну. Головна проблема та складність цього процесу полягає у тому, як правильно поділити панельні дані на дві частини. Розподіл відбувається для кожної країни, а потім об’єднуємо тестову та тренувальну частини для кожної країни (див.рис.3.39) [31].

```
def train_test_split(data):
    size=int(len(data)*0.8)
    # for train data will be collected from each country's data which index is from 0-size (80%)
    x_train =data.drop(['MAV'], axis=1).iloc[0:size]
    # for test data will be collected from each country's  data which index is from size to the end (20%)
    x_test = data.drop(['MAV'], axis=1).iloc[size:]
    y_train=data['MAV'].iloc[0:size]
    y_test=data['MAV'].iloc[size:]
    return x_train, x_test,y_train,y_test
```

```
country=list(set(dt.Countries))
# loop each station and collect train and test data
X_train=[]
X_test=[]
Y_train=[]
Y_test=[]
for i in range(0,len(country)):
    data=dt[dt['Countries']==country[i]]
    x_train, x_test,y_train,y_test=train_test_split(data)
    X_train.append(x_train)
    X_test.append(x_test)
    Y_train.append(y_train)
    Y_test.append(y_test)
```

Рисунок 3.39 – Розподіл бази даних та тренувальну та тестову частини

Після розподілу бази даних на тестову та тренувальну частину, ідентифікуємо тренувальні та тестові залежні та незалежні змінні (див. рис. 3.40) [36]:

```
from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder()
#combine x train and y train as train data
train_data=pd.DataFrame()
train_data[X_train.columns]=X_train
train_data[Y_train.columns]=Y_train
train_data['Countries']= encoder.fit_transform(train_data['Countries'])
#combine x test and y test as test data
test_data=pd.DataFrame()
test_data[X_test.columns]=X_test
test_data[Y_test.columns]=Y_test
test_data['Countries']= encoder.fit_transform(test_data['Countries'])
# using the function to obtian reshaped x_train,x_test,y_train,y_test
x_train,x_test,y_train,y_test=reshape_data(train_data,test_data)
```

Рисунок 3.40 – Ідентифікація тренувальних та тестових змінних

Всі дані підготовлені до побудови прогнозної моделі на основі об’єднаної моделі та LSTM моделі [37].

Спочатку побудуємо об’єднану прогнозу модель з використанням тестового та тренувального розподілу бази даних.

```
y_pred = model.predict(X_test)
df_results = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
df_results
```

	Actual	Predicted
780	1.609438	1.174482
817	8.043021	7.272133
363	9.110631	8.806825
308	6.848005	6.625916
1205	7.132498	7.771581
...	...	...
609	8.336390	8.479036
332	5.587249	4.932872
1088	4.828314	4.211598
1137	3.367296	6.138822
149	7.920810	7.571089

Рисунок 3.41 – Результати побудови прогнозної моделі на основі Pooled regression

Одним із способів оцінити, наскільки добре регресійна модель відповідає набору даних, є обчислення середньої квадратичної похибки кореня (RMSE), яка є метрикою, яка повідомляє нам середню відстань між прогнозованими значеннями від моделі та фактичними значеннями в наборі даних [38].

Чим нижче RMSE, тим краще дана модель здатна «підігнати» набір даних (3.13).

$$RMSE = \frac{\sqrt{\sum(P-A)^2}}{n} \tag{3.13}$$

де,

P – прогнозне значення;

A – значення спостереження;

n – розмір вибірки.

Тому розраховуємо середню квадратичну похибку кореня та коефіцієнт детермінації для прогнозної моделі (див. рис. 3.42):

```
from sklearn.metrics import r2_score, mean_squared_error
RMSE = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)
print('RMSE:', RMSE, 'R2:',r2)
```

```
RMSE: 1.1298507283393997 R2: 0.7140165385759949
```

Рисунок 3.42 – Розрахунок показників для прогнозної моделі для змінної MAV

Середню квадратичну похибку кореня становить 1.12985 та коефіцієнт детермінації дорівнює 0.71, що являється досить гарним результатом.

Тепер необхідно побудувати нову прогнозну модель - LSTM модель для порівняння та обрати кращу модель для прогнозування явища кібератак.

LSTM модель буде будуватися за допомогою Sequential() задачі (див.рис. 3.41).

Всі дані підготовлені до побудови LSTM моделі. LSTM модель буде будуватися за допомогою Sequential() задачі (див. рис. 3.43) [39].

Sequential (Класифікація послідовностей) - це задача прогностичного моделювання, де є певна послідовність входів у просторі або часі, і завдання полягає в тому, щоб передбачити категорію для послідовності.

```
model = Sequential()
model.add(LSTM(60, activation='sigmoid',input_shape=(x_train.shape[1], x_train.shape[2])))
model.add(Dense(1))
model.compile(loss='mae', optimizer='adam')
# fit network
history = model.fit(x_train, y_train, epochs=1000, batch_size=64, verbose=0, shuffle=False)
```

```
# make a prediction
y_test_pre=model.predict(x_test)
```

```
9/9 [==============================] - 0s 875us/step
```

```
# make a prediction
y_test_pre=model.predict(x_test)
y_test_pre.shape,y_test.shape
```

```
9/9 [==============================] - 0s 750us/step
((279, 1), (279,))
```

Рисунок 3.43 – Побудова прогнозної LSTM моделі

Після побудови прогнозної моделі отримуємо наступні результати (див.рис.3.44).

```
pa=pd.DataFrame()
pa['Data']=X_test.reset_index().Data.iloc[1:-1]
pa['Prediction']=[i[0] for i in y_test_pre][1:]
pa['Actual Values']=y_test[:-1]
pa.head()
```

	Data	Prediction	Actual Values
1	2022-09-08	5.344671	4.553877
2	2022-09-09	5.325020	4.418841
3	2022-09-10	4.153235	2.772589
4	2022-09-11	5.082784	3.912023
5	2022-09-12	5.840456	4.867534

Рисунок 3.44 – Прогнозні дані для змінної MAV

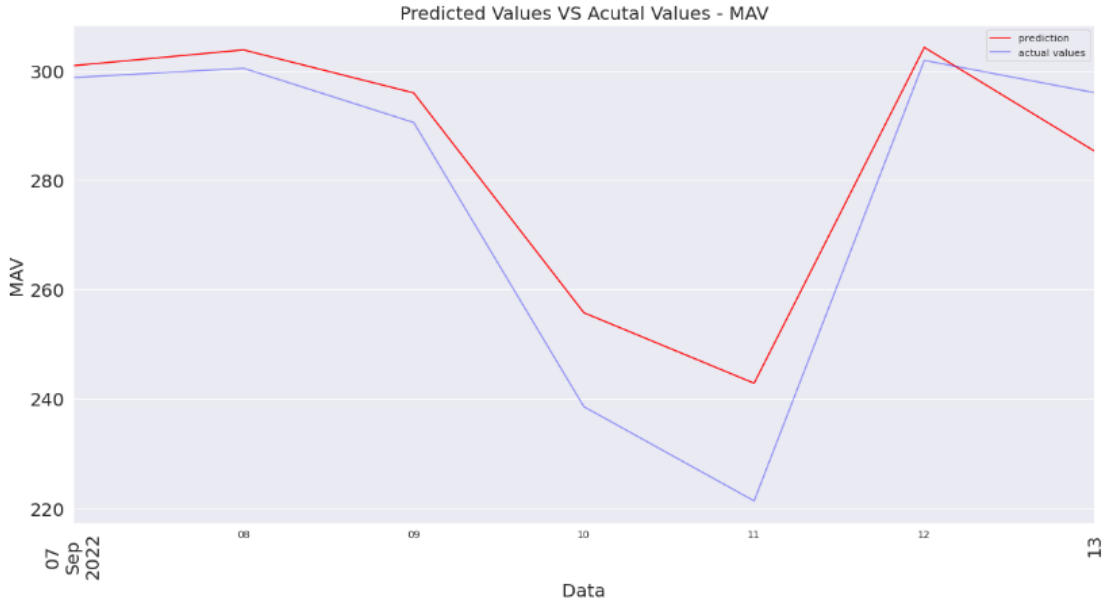


Рисунок 3.45 – Результати побудови прогнозної LSTM моделі

```
print(RMSE(y_test[:-1],[i[0] for i in y_test_pre][1:]))
```

```
0.5821494070053101
```

```
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
print('R2 Score: ', r2_score(y_test, y_test_pre))
```

```
R2 Score:  0.5151603123958282
```

Рисунок 3.46 – Розрахунок показників для прогнозної моделі для змінної MAV

Середню квадратичну похибку кореня становить 0.58 та коефіцієнт детермінації дорівнює 0.51, що являється теж досить гарним результатом. Але спираючись на той факт, що коефіцієнт детермінації не є повністю надійним показником для порівняння моделей, використаємо середню квадратичну

похибку, яка повідомляє нам середню відстань між прогнозованими значеннями від моделі та фактичними значеннями в наборі даних.

Тому кращою прогнозною моделлю для змінної MAV є LSTM модель. Спробуємо побудувати прогнозну LSTM модель для кожної країни (див. рис. 3.47) [40].

```
def normalization_train_test_split(country):
    scaler = PowerTransformer(method='yeo-johnson', standardize=True)
    scaled = scaler.fit_transform(country.drop(columns=['Country','Date']))
    # create dateframe fir scaled data
    scaled_df=pd.DataFrame(data=scaled,columns=country.drop(columns=['Country','Date']).columns)
    scaled_df['MAV']=list(country.MAV)
    X_train, X_test,Y_train,Y_test=train_test_split(scaled_df)
    #combine x train and y train as train data
    train_data=pd.DataFrame()
    train_data[X_train.columns]=X_train
    train_data['MAV']=Y_train
    #combine x test and y test as test data
    test_data=pd.DataFrame()
    test_data[X_test.columns]=X_test
    test_data['MAV']=Y_test

    # using the function to obtian reshaped x_train,x_test,y_train,y_test
    x_train,x_test,y_train,y_test=reshape_data(train_data,test_data)
    return x_train, x_test,y_train,y_test
```

```
# Loop through top 10 countries' data
#
for i in range(len(top_10_country_names)):
    # obtian one country's data
    country=df_data[df_data.Country==top_10_country_names[i]]
    # train test split, nomalization and reshape the data
    x_train, x_test,y_train,y_test=normalization_train_test_split(country)
    # model
    model = Sequential()
    model.add(LSTM(60, activation='sigmoid',input_shape=(x_train.shape[1], x_train.shape[2])))
    model.add(Dense(1))
    model.compile(loss='mae', optimizer='adamax')
    # fit network
    history = model.fit(x_train, y_train, epochs=2000, batch_size=128,  verbose=0, shuffle=False)
    # make a prediction
    y_test_pre=model.predict(x_test)
    #RMSE
    rmse=RMSE(y_test[:-1],[i[0] for i in y_test_pre][1:])
    print('{} - RMSE: {}'.format(top_10_country_names[i],rmse))
    #create new dataframe for plot
    pa=pd.DataFrame()
    pa['Date']=list(country.Date.iloc[int(len(country)*0.8):])[1:-1]
    pa['Prediction']=[i[0] for i in y_test_pre][1:]
    pa['Actual Values']=list(y_test[:-1])

    plt.figure(figsize=(20,10))
    pa.groupby('Date')['Prediction'].sum().plot(kind='line',label='prediction',color='red',alpha=1)
    pa.groupby("Date")['Actual Values'].sum().plot(kind='line',label='actual values',color='blue',alpha=0.4)
    plt.xticks(rotation=90,size=20)
    plt.yticks(size=20)

    plt.ylabel('MAV',fontsize=20)
    plt.xlabel('Date',fontsize=20)
    plt.title('Predicted Values VS Acutal Values - MAV in {}'.format(top_10_country_names[i]),fontsize=20)
    plt.legend()
```

Рисунок 3.47 – Побудова прогнозної моделі для кожної країни

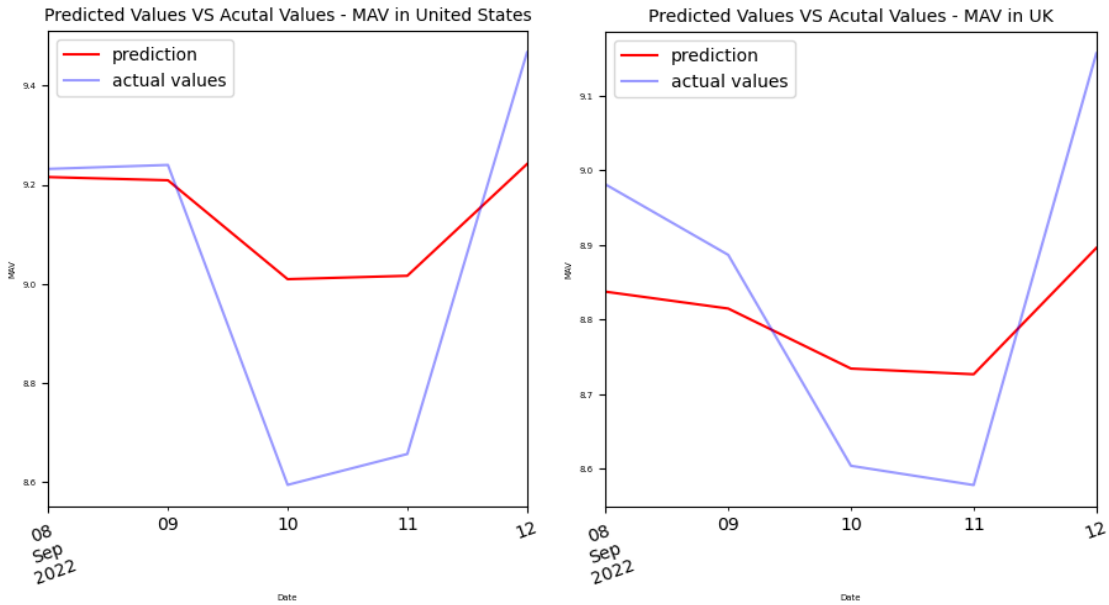


Рисунок 3.48 – Прогноз значення MAV для США та Велокобританії

Отримані гарні результати для країн. Прикладом прогнозу для країн виступають США та Великобританія за значенням середньої квадратичної похибки 0.266 та 0.163 відповідно.

3.5.2 Прогнозування трендів кібератаки виду KAS

Визначивши найкращу модель для моделювання виду кібератаки KAS, спрогнозуємо тренд за допомогою об’єднаної моделі (Pooled model) та LSTM моделі.

Перш ніж побудувати прогноз необхідно поділити базу даних на дві частини - тестову та тренувальну (див. рис. 3.39).

Після розподілу бази даних на тестову та тренувальну частину, ідентифікуємо тренувальні та тестові залежні та незалежні змінні (див. рис. 3.40):

Спочатку побудуємо об’єднану прогнозу модель з використанням тестового та тренувального розподілу бази даних.

```
y_pred = model.predict(X_test)
df_results = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
df_results
```

	Actual	Predicted
999	14.939141	13.788484
651	9.952278	12.723177
1023	13.138237	14.310919
250	12.398757	12.584448
860	11.362103	11.019525
...	...	...
631	11.532728	10.876397
473	15.882500	15.127464
916	11.497812	11.197719
760	10.373491	11.069807
173	16.437204	16.313346

Рисунок 3.49 – Результати побудови прогнозної моделі на основі Pooled regression

Одним із способів оцінити, наскільки добре регресійна модель відповідає набору даних, є обчислення середньої квадратичної похибки кореня (RMSE

Тому розраховуємо середню квадратичну похибку кореня та коефіцієнт детермінації для прогнозної моделі [див. рис. 3.50]:

```
from sklearn.metrics import r2_score, mean_squared_error
RMSE = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)
print(RMSE, r2)
```

```
0.6147591605646721 0.9297993080835679
```

Рисунок 3.50 – Розрахунок показників для прогнозної моделі для змінної KAS

Середню квадратичну похибку кореня становить 0.6148 та коефіцієнт детермінації дорівнює 0.9298, що являється досить гарним результатом.

Тепер необхідно побудувати нову прогнозну модель - LSTM модель для порівняння та обрати кращу модель для прогнозування явища кібератак.

Після побудови прогнозної моделі отримуємо наступні результати (див.рис.3.51).

	Date	Prediction	Actual Values
1	2022-09-08	16.024281	15.841156
2	2022-09-09	16.052122	15.875068
3	2022-09-10	15.683013	15.471450
4	2022-09-11	15.695789	15.488205
5	2022-09-12	15.915629	15.735956

Рисунок 3.51 – Прогнозні дані для змінної KAS

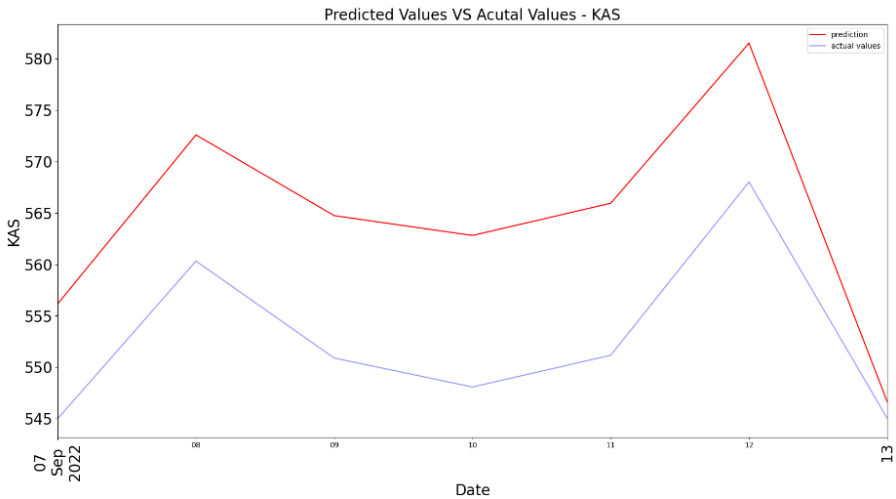


Рисунок 3.52 – Результати побудови прогнозної LSTM моделі

```
print(RMSE(y_test[:-1],[i[0] for i in y_test_pre][1:]))
```

```
0.505295611130907
```

```
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
print('R2 Score: ', r2_score(y_test, y_test_pre))
```

```
R2 Score:  0.7116407077885306
```

Рисунок 3.53 – Розрахунок показників для прогнозної моделі для змінної KAS

Середню квадратичну похибку кореня становить 0.51 та коефіцієнт детермінації дорівнює 0.71, що являється теж досить гарним результатом. Але спираючись на той факт, що коефіцієнт детермінації не є повністю надійним показником для порівняння моделей, використаємо середню квадратичну похибку, яка повідомляє нам середню відстань між прогнозованими значеннями від моделі та фактичними значеннями в наборі даних.

Тому кращою прогнозною моделлю для змінної KAS є LSTM модель. Спробуємо побудувати прогнозну LSTM модель для кожної країни (див. рис. 3.54).

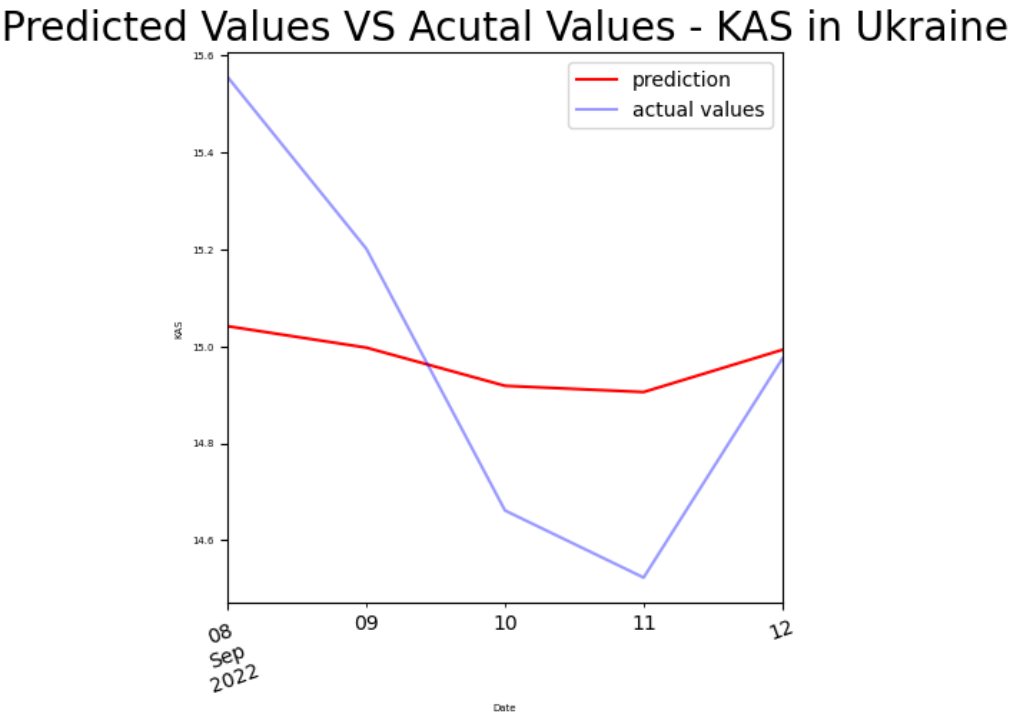


Рисунок 3.54 – Прогноз значення KAS для України

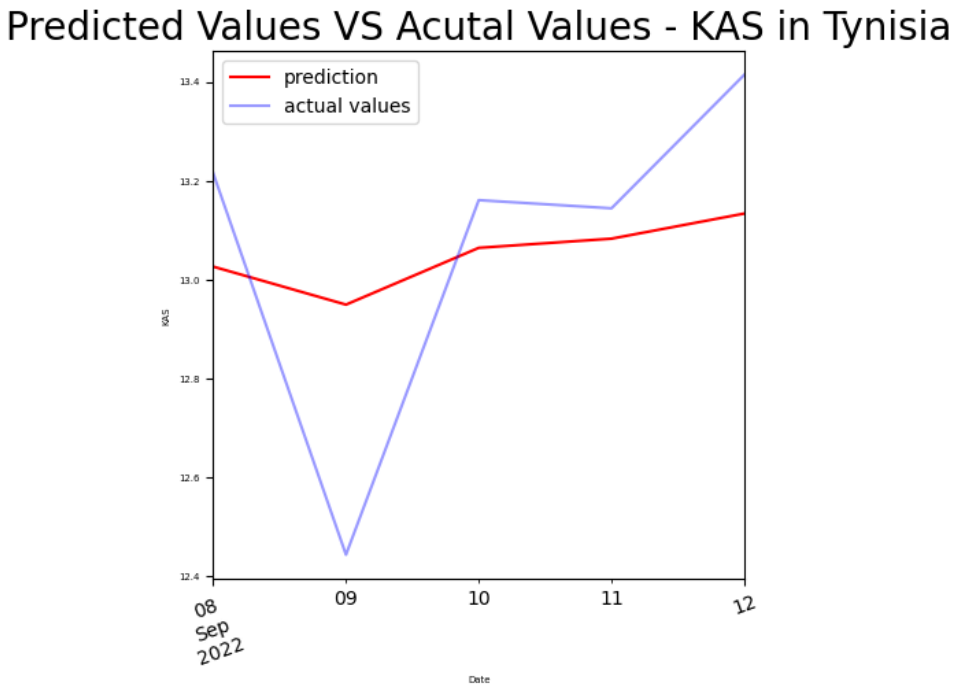


Рисунок 3.55 – Прогноз значення KAS для Тунісу

Отримані гарні результати для країн. Прикладом прогнозу для країн виступають Україна та Туніс за значенням середньої квадратичної похибки 0.329 та 0.277 відповідно.

3.5.3 Прогнозування трендів кібератаки виду IDS

Визначивши найкращу модель для моделювання виду кібератаки IDS, спрогнозуємо тренд за допомогою об’єднаної моделі (Pooled model) та LSTM моделі.

Перш ніж побудувати прогноз необхідно поділити базу даних на дві частини - тестову та тренувальну (див. рис. 3.39).

Після розподілу бази даних на тестову та тренувальну частину, ідентифікуємо тренувальні та тестові залежні та незалежні змінні (див. рис. 3.40):

Спочатку побудуємо об’єднану прогнозу модель з використанням тестового та тренувального розподілу бази даних.

```
y_pred = model.predict(X_test)
df_results = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
df_results
```

	Actual	Predicted
729	11.349959	11.356542
925	9.655667	9.319099
1070	10.403202	10.244445
249	9.120963	12.228848
961	13.244149	13.253300
...	...	...
880	9.081597	8.644902
99	11.076465	11.119290
332	7.352441	7.202662
400	9.927790	9.709871
418	11.359040	11.312384

Рисунок 3.56 – Результати побудови прогнозної моделі на основі Pooled regression

Одним із способів оцінити, наскільки добре регресійна модель відповідає набору даних, є обчислення середньої квадратичної похибки кореня (RMSE

Тому розраховуємо середню квадратичну похибку кореня та коефіцієнт детермінації для прогнозної моделі (див. рис. 3.56):

```
from sklearn.metrics import r2_score, mean_squared_error
RMSE = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)
print(RMSE, r2)
```

```
0.5360467594061045 0.9429578957759858
```

Рисунок 3.57 – Розрахунок показників для прогнозної моделі для змінної IDS

Середню квадратичну похибку кореня становить 0.536 та коефіцієнт детермінації дорівнює 0.943, що являється досить гарним результатом.

Тепер необхідно побудувати нову прогнозну модель - LSTM модель для порівняння та обрати кращу модель для прогнозування явища кібератак.

Після побудови прогнозної моделі отримуємо наступні результати (див.рис.3.57).

	Date	Prediction	Actual Values
1	2022-09-08	9.170294	9.013839
2	2022-09-09	8.823923	8.629629
3	2022-09-10	8.770832	8.567126
4	2022-09-11	9.114446	8.941807
5	2022-09-12	9.222954	9.060099

Рисунок 3.58 – Прогнозні дані для змінної IDS

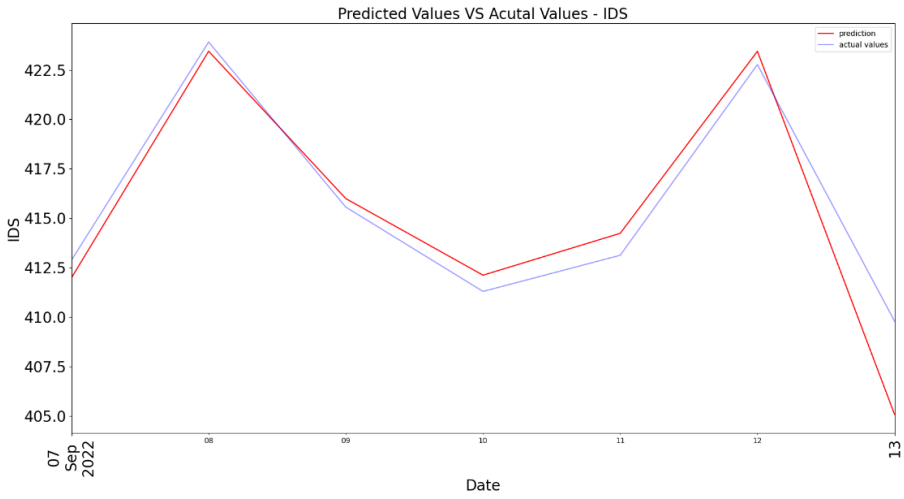


Рисунок 3.59 – Результати побудови прогнозної LSTM моделі

```
print(RMSE(y_test[:-1],[i[0] for i in y_test_pre][1:]))
```

```
0.4595955566642357
```

```
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
print('R2 Score: ', r2_score(y_test, y_test_pre))
```

```
R2 Score:  0.7257339291988435
```

Рисунок 3.60 – Розрахунок показників для прогнозної моделі для змінної IDS

Середню квадратичну похибку кореня становить 0.459 та коефіцієнт детермінації дорівнює 0.726, що являється теж досить гарним результатом.

Але спираючись на той факт, що коефіцієнт детермінації не є повністю надійним показником для порівняння моделей, використаємо середню квадратичну похибку, яка повідомляє нам середню відстань між

прогнозованими значеннями від моделі та фактичними значеннями в наборі даних. Тому кращою прогнозною моделлю для змінної KAS є LSTM модель.

Спробуємо побудувати прогнозну LSTM модель для кожної країни (див. рис. 3.61).

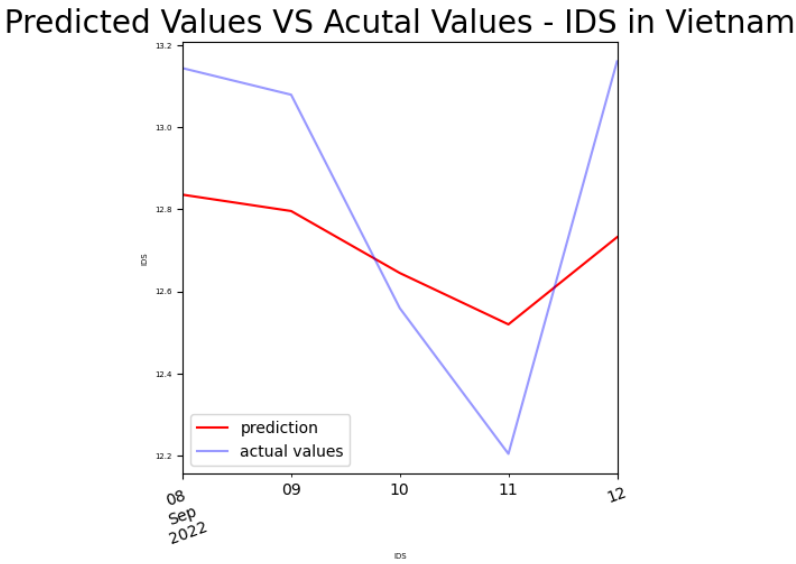


Рисунок 3.61 – Прогноз значення KAS для В’єтнаму

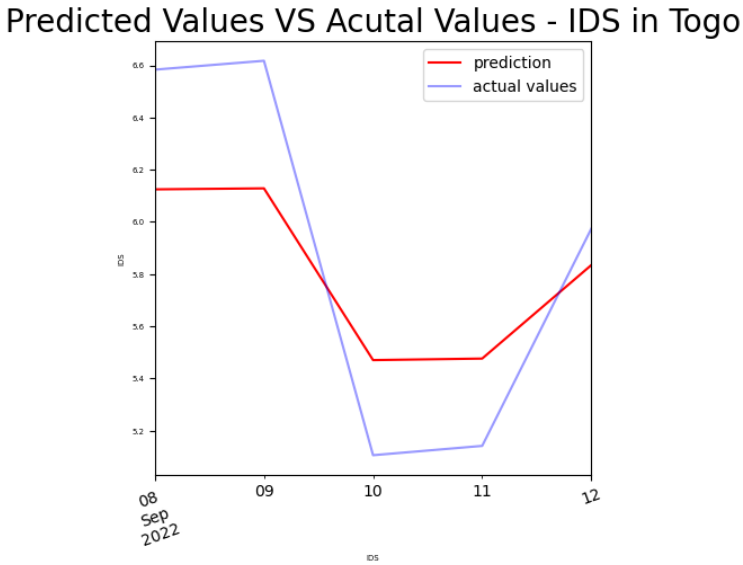


Рисунок 3.62 – Прогноз значення KAS для Того

Отримані гарні результати для країн. Прикладом прогнозу для країн виступають В’єтнам та Того за значенням середньої квадратичної похибки 0.304 та 0.377 відповідно.

ВИСНОВКИ

Проблема кібератак на сьогоднішній день є досить актуальним явищем, що пояснюється впливом різних факторів. Тому на практиці існує потреба не тільки у моделюванні трендів, але й у попередженні їх настання. Це можливо здійснити тільки із використанням сучасних інформаційних технологій та математичних методів.

У даному дослідженні було висвітлено поняття та сутність кібератак та проаналізовано сучасні тренди та методи їх попередження.

Було здійснено аналіз, моделювання та прогнозування трендів кібератак за допомогою побудови математичних моделей та регресії для панельних даних, а саме об’єднаної моделі, моделі фіксованих ефектів, випадкових ефектів та LSTM моделі.

Відповідні розрахунки було проведено із використанням сучасної мови програмування Python.

Вважаємо, що побудована об’єднана модель буде одним із найкращих методів, що дозволяє змоделювати тренди кібератак та продемонструвала гарні результати скоригованого коефіцієнту детермінації, залишкових похибок моделі та параметру Акайка (AIC).

Також було побудовано прогнозну модель на основі об’єднаної та LSTM модель. На останньому етапі було проведено порівняння побудованих прогнозних моделей для кожної незалежної змінної, в результаті чого найкращі результати продемонструвала LSTM, незважаючи на гірший результат скоригованого коефіцієнту детермінації, середня квадратна похибка кореня показала кращий результат, що означає кращу здатність «підігнати» дані набору даних для прогнозування. Щоб мати постійне уявлення про ймовірні кібератакb, результати повинні регулярно доповнюватися, оновлюватися для використання їх у реальних умовах, що дозволить вчасно реагувати на злочинні дії та попереджати їх виникнення.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Що таке кібератака? . [Електронний ресурс]. – Режим доступу до ресурсу:https://www.techtarget.com/searchsecurity/definition/cyber-attack.

2. What Is a Cyber Threat? Definition, Types, Hunting, Best Practices, and Examples. [Електронний ресурс]. – Режим доступу до ресурсу: https://www.spiceworks.com/it-security/vulnerability management/articles/what-is-cyber-threat/.

3. Кібератака. [Електронний ресурс]. – Режим доступу до ресурсу: https://vue.gov.ua/%D0%9A%D1%96%D0%B1%D0%B5%D1%80%D0%B0%D1%82%D0%B0%D0%BA%D0%B0.

4. Сучасні тренди кібербезпекової політики: висновки для України. Аналітична записка. [Електронний ресурс]. – Режим доступу до ресурсу: https://niss.gov.ua/doslidzhennya/nacionalna-bezpeka/suchasni-trendi-kiberbezpekovoi-politiki-visnovki-dlya-ukraini.

5. 166 Cybersecurity Statistics and Trends. [Електронний ресурс]. – Режим доступу до ресурсу: https://www.varonis.com/blog/cybersecurity-statistics#trends.

6. The Global Risk Report. [Електронний ресурс]. – Режим доступу до ресурсу:https://www3.weforum.org/docs/WEF_Global_Risk_Report_2020.pdf.

7. Reports largest single day virus spike. [Електронний ресурс]. – Режим доступу до ресурсу: https://abcnews.go.com/Health/wireStory/latest-india-reports-largest-single-day-virus-spike-70826542.

8. 2022 Must-Know Cyber Attack Statistics and Trendshttps. [Електронний ресурс]. – Режим доступу до ресурсу: www.embroker.com/blog/cyber-attack-statistics/.

9. Microsoft Digital Defense Report shares new insights on nation-state attacks. [Електронний ресурс]. – Режим доступу до ресурсу:

https://www.microsoft.com/en-us/security/blog/2021/10/25/microsoft-digital-defense-report-shares-new-insights-on-nation-state-attacks/.

10. Кібератаки в Україні 2022. [Електронний ресурс]. – Режим доступу до ресурсу: https://www.ukrinform.ua/rubric-technology/3584942-majze-polovinu-kiberatak-sbu-viavlae-u-rezimi-realnogo-casu.html.

11. Blochchain. [Електронний ресурс]. – Режим доступу до ресурсу: https://www.it.ua/knowledge-base/technology-innovation/blockchain#:.

12. Що таке патч? (Визначення виправлення / виправлення). [Електронний ресурс]. – Режим доступу до ресурсу: https://uk.go-travels.com/91430-what-is-a-patch-2625960-6579554.

13. 5 Cybersecurity Policy Trends. [Електронний ресурс]. – Режим доступу до ресурсу: https://sopa.tulane.edu/blog/5-cybersecurity-policy-trends.

14. CYBERTHREAT REAL-TIME MAP. [Електронний ресурс]. Режим доступу до ресурсу: https://cybermap.kaspersky.com/stats#country=27&type=RMW&period=m.

15. Introduction to Pandas and NumPy. [Електронний ресурс]. – Режим доступу до ресурсу: https://www.codecademy.com/article/introduction-to-numpy-and-pandas.

16. Reading an excel file using Python. . [Електронний ресурс]. – Режим доступу до ресурсу: https://www.geeksforgeeks.org/reading-excel-file-using-python/.

17. Statsmodels.tsa.seasonal.seasonal_decompose. [Електронний ресурс]. Режим доступу до ресурсу: statsmodels.tsa.seasonal.seasonal_decompose — statsmodels.

18. Нормальний розподіл. [Електронний ресурс]. Режим доступу до ресурсу: https://pidru4niki.com/19240701/statistika/normalniy_rozpodil.

19. Normalization. [Електронний ресурс]. Режим доступу до ресурсу: https://developers.google.com/machine-learning/data-prep/transform/normalization.

20. What is Jarque Bera test in ML python. [Електронний ресурс]. Режим доступу до ресурсу: https://www.statology.org/jarque-bera-test-python/#:~:text=How%20to%20Perform%20a%20Jarque-Bera%20Test%20in%20Python,skewness%20and%20kurtosis%20that%20matches%20a%20normal%20distribution.

21. What Is Panel Data? (With Uses, Advantages and an Example). [Електронний ресурс]. Режим доступу до ресурсу: https://www.indeed.com/career-advice/career-development/panel-data.

22. МОДЕЛЮВАННЯ НА ОСНОВІ ПАНЕЛЬНИХ ДАНИХ. [Електронний ресурс]. Режим доступу до ресурсу: https://posibniki.com.ua/post-modeluvannya-na-osnovi-panelnih-danih

23. The Pooled OLS Regression Model For Panel Data Sets. [Електронний ресурс]. Режим доступу до ресурсу: https://timeseriesreasoning.com/contents/pooled-ols-regression-models-for-panel-data-sets/.

24. The Fixed Effects Regression Model For Panel Data Sets. [Електронний ресурс]. Режим доступу до ресурсу: https://timeseriesreasoning.com/contents/the-fixed-effects-regression-model-for-panel-data-sets/ю

25. The Random Effects Regression Model for Panel Data Sets. [Електронний ресурс]. Режим доступу до ресурсу: https://timeseriesreasoning.com/contents/the-random-effects-regression-model-for-panel-data-sets/.

26. LSTM. [Електронний ресурс]. Режим доступу до ресурсу: https://habr.com/ru/company/wunderfund/blog/331310/.

27. Time Series - LSTM Model. [Електронний ресурс]. Режим доступу до ресурсу: https://www.tutorialspoint.com/time_series/time_series_lstm_model.htm

28. LSTMs Explained: A Complete, Technically Accurate, Conceptual Guide with Keras. [Електронний ресурс]. Режим доступу до ресурсу:

https://medium.com/analytics-vidhya/lstms-explained-a-complete-technically-accurate-conceptual-guide-with-keras-2a650327e8f2.

29. Statsmodels. [Електронний ресурс]. Режим доступу до ресурсу: https://www.statsmodels.org/stable/index.html.

30. Using The F Statistic. [Електронний ресурс]. Режим доступу до ресурсу: F Statistic / F Value: Definition and How to Run an F-Test (statisticshowto.com).

31. How to Interpret Log-Likelihood Values (With Examples). [Електронний ресурс]. Режим доступу до ресурсу: How to Interpret Log-Likelihood Values (With Examples) - Statology.

32. Akaike Information Criterion | When & How to Use It (Example). [Електронний ресурс]. Режим доступу до ресурсу: https://www.scribbr.com/statistics/akaike-information-criterion/.

33. Residual Standard Deviation/Error: Guide for Beginners. [Електронний ресурс]. Режим доступу до ресурсу: https://quantifyinghealth.com/residual-standard-deviation-error/.

34. How to Calculate Standardized Residuals in Python. [Електронний ресурс]. Режим доступу до ресурсу: https://www.statology.org/standardized-residuals-python/

35. The Breusch-Pagan Test: Definition & Example. [Електронний ресурс]. Режим доступу до ресурсу: https://www.statology.org/breusch-pagan-test/.

36. Train-Test Split for Evaluating Machine Learning Algorithms. [Електронний ресурс]. Режим доступу до ресурсу: https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/.

37. LabelEncoder. [Електронний ресурс]. Режим доступу до ресурсу: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html.

38. RMSE: Root Mean Square Error. [Електронний ресурс]. Режим доступу до ресурсу: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/.

39. Sequence Classification with LSTM Recurrent Neural Networks in Python with Keras. [Електронний ресурс]. Режим доступу до ресурсу: https://machinelearningmastery.com/sequence-classification-lstm-recurrent-neural-networks-python-keras/.

40. How to make predictions from LSTM and plot it using python? [Електронний ресурс]. Режим доступу до ресурсу: https://stackoverflow.com/questions/64750530/how-to-make-predictions-from-lstm-and-plot-it-using-python.

ДОДАТКИ

ДОДАТОК А

SUMMARY

Kobzenko V.V. Modeling and forecasting of cyberattack trends. - Qualification master's work. Sumy State University, Sumy, 2022 The essence and types of cyberattacks are investigated, statistical data and current trends in cybersecurity policy are considered. The methods of panel data research are analyzed, as a result, regressions for all independent variables are built and the relevant one is selected. The main purpose of the study is to build models for forecasting cyberattack trends.
Keywords: cyber attack, cyber war, modeling, forecasting, fixed effects, random effects, LSTM model.

АНОТАЦІЯ

Кобзенко В.В. Моделювання та прогнозування трендів кібератак. - Кваліфікаційна магістерська робота. Сумський державний університет, Суми, 2022 р. У роботі досліджено сутність та види кібератак, розглянуто статистичні дані та сучасні тренди кібербезпекової політики. Проаналізовано методи дослідження панельних даних, в результаті чого, побудовано регресії для всіх незалежних змінних та обрано релевантну. Основною метою дослідження є побудова моделей прогнозування трендів кібератак.
Ключові слова: кібератака, кібервійна, моделювання, прогнозування, фіксовані ефекти, випадкові ефекти, LSTM модель.

ДОДАТОК Б

```
#create log-transformed data
logKAS = np.log(df["KAS"])

#define grid of plots
fig, axs = plt.subplots(nrows=1, ncols=2)

#create histograms
axs[0].hist(df["KAS"], edgecolor='black')
axs[1].hist(logKAS, edgecolor='black')

#add title to each histogram
axs[0].set_title('Original Data')
axs[1].set_title('Log-Transformed Data')
```

```
Text(0.5, 1.0, 'Log-Transformed Data')
```

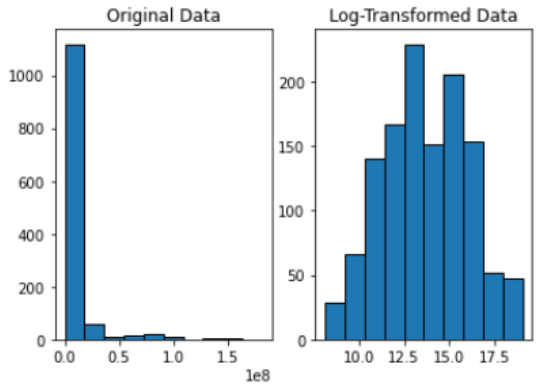


Рис. Б1 – Трансформування даних з x на log(x)

```
#create log-transformed data
logIDS = np.log(df["IDS"])

#define grid of plots
fig, axs = plt.subplots(nrows=1, ncols=2)

#create histograms
axs[0].hist(df["IDS"], edgecolor='black')
axs[1].hist(logIDS, edgecolor='black')

#add title to each histogram
axs[0].set_title('Original Data')
axs[1].set_title('Log-Transformed Data')
```

```
Text(0.5, 1.0, 'Log-Transformed Data')
```

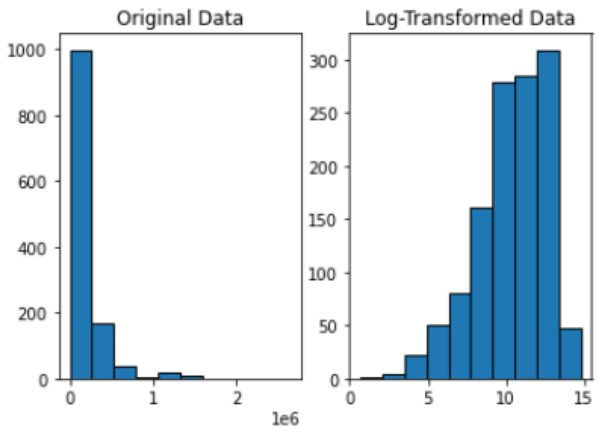


Рис. Б2 – Трансформування даних з x на log(x)

```
print(lsdv_model_results.resid)
print('Mean value of residual errors='+str(lsdv_model_results.resid.mean()))
```

```
0      -1.659387
1      -1.079064
2      -0.969684
3      -0.120892
4      -0.392670
          ...
1235    0.829994
1236   -1.006582
1237   -1.348459
1238    0.597081
1239    0.442660
Length: 1240, dtype: float64
Mean value of residual errors=4.410210548010421e-13
```

Рис. Б3 –Розрахунок залишкових похибок для моделі фіксованих ефектів для змінної MAV

```
: print(re_model_results.resid)
print('Mean value of residual errors='+str(re_model_results.resid.mean()))
```

```
0        -344.765675
1        -268.235060
2        -261.704446
3          45.826169
4        -107.643216
             ...
1235    13397.481925
1236   -16459.987460
1237   -18109.456845
1238     6280.073770
1239     2280.604385
Length: 1240, dtype: float64
Mean value of residual errors=-1.0169617692759681e-10
```

Рис. Б4 –Розрахунок залишкових похибок для моделі випадкових ефектів для змінної MAV

```
print(lsdv_model_results.resid)
print('Mean value of residual errors='+str(lsdv_model_results.resid.mean()))
```

```
0      -0.169107
1       0.152794
2       0.111757
3       0.190995
4       0.084579
          ...
1235    0.073640
1236   -0.193884
1237   -0.388195
1238   -0.244688
1239   -0.475146
Length: 1240, dtype: float64
Mean value of residual errors=8.521255385894913e-13
```

Рис. Б5 –Розрахунок залишкових похибок для моделі фіксованих ефектів для змінної KAS

```
#Concatenate the unit names column to the Dataframe containing the residuals from the Pooled OLSR model
df_pooled_olsr_resid_with_unitnames = pd.concat([df_data[unit_col_name],pooled_olsr_model_results.resid], axis=1)

df_pooled_olsr_resid_group_means = df_pooled_olsr_resid_with_unitnames.groupby(unit_col_name).mean()

ssr_grouped_means=(df_pooled_olsr_resid_group_means[0]**2).sum()

ssr_pooled_olsr=pooled_olsr_model_results.ssr


LM_statistic = (n*T)/(2*(T-1))*math.pow(((T*T*ssr_grouped_means)/ssr_pooled_olsr - 1),2)

print('BP LM Statistic='+str(LM_statistic))
```

```
BP LM Statistic=18.008298755186722
```

```
alpha=0.05
chi2_critical_value=st.chi2.ppf((1.0-alpha), 1)
print('chi2_critical_value='+str(chi2_critical_value))
```

```
chi2_critical_value=3.841458820694124
```

Рисунок Б6-Перевірка значущості моделі випадкового ефекту для змінної KAS

```
print(re_model_results.resid)
print('Mean value of residual errors='+str(re_model_results.resid.mean()))
```

```
0      -2.014812e+05
1       1.970110e+05
2       8.100335e+04
3       1.689956e+05
4      -6.101206e+04
           ...
1235    3.608206e+06
1236   -2.737802e+06
1237   -6.386810e+06
1238   -3.127817e+06
1239   -7.400325e+06
Length: 1240, dtype: float64
Mean value of residual errors=3.922370172316028e-08
```

Рис. Б7 –Розрахунок залишкових похибок для моделі випадкових ефектів для змінної KAS

```
print(lsdv_model_results.resid)
print('Mean value of residual errors='+str(lsdv_model_results.resid.mean()))
```

```
0      -0.829117
1      -0.830258
2      -0.792734
3      -0.770331
4      -0.843910
          ...
1235   -0.152933
1236   -0.247574
1237   -0.222308
1238   -0.104814
1239   -0.041814
Length: 1240, dtype: float64
Mean value of residual errors=6.515333075918944e-13
```

Рисунок Б8-Перевірка значущості моделі випадкового ефекту для змінної IDS