Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Програмної інженерії

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

другий (магістерський)
(рівень вищої освіти)

Дослідження ціноутворення криптовалюти та ефективність використання машинного навчання для пошуку тренду вартості

Виконав:
студент 2 курсу групи ІПЗм-20-2

Бизкровний О. М.
(прізвище, ініціали)

Спеціальність	121 – Інженерія програмного забезпечення
Тип програми	Освітньо-наукова
Керівник	проф. Смеляков К. С.
	(посада, прізвище, ініціали)

Допускається до захисту
Зав. Кафедри З.В. Дудар

2022 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Програмної інженерії

Рівень вищої освіти другий (магістерський)

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

Тип програми освітньо-наукова програма

Освітня програма Інженерія програмного забезпечення

ЗАТВЕРДЖУЮ:

Зав. кафедри _____
(підпис)
«__»_____2022 р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студента Бизкровному Олександр Миколайовичу
(прізвище, ім'я, по батькові)

1. Тема роботи «Дослідження ціноутворення криптовалюти та ефективність використання машинного навчання для пошуку тренду вартості»

затверджена наказом університету від «___» _____ 2022 р. № ___

2. Термін подання студентом роботи до екзаменаційної комісії «__» _____ 202__ р.

3. Вихідні дані до роботи *не підготовані дані для регресійного аналізу, розподілене обчислення, AWS, Java, IntelijIDEA, Apache Spark, PostgreSQL.*

4. Перелік питань, що потрібно опрацювати в роботі *аналіз вже створених систем прогнозу вартості криптовалюти, виявлення інформативних факторів, що описують середовище криптовалюти та можуть бути використаними у алгоритмах машинного навчання, дослідження алгоритмів машинного навчання, покращення або створення алгоритму дослідження коливань вартості криптовалюти.*

КАЛЕНДАРНИЙ ПЛАН

№	Назви етапів роботи	Терміни виконання етапів роботи	Примітка
1.	Аналіз предметної галузі		виконано
2.	Дослідження предметної області		виконано
3.	Дослідження аналогічних систем		виконано
4.	Постановка задачі		виконано
5.	Планування та проведення експерименту		виконано
6.	Проектування програмної системи на базі результатів експерименту		виконано
7.	Підготовка пояснювальної записки		виконано
8.	Нормоконтроль, рецензування		виконано
9.	Занесення диплома в електронний архів		виконано
10.	Попередній захист		виконано
11.	Допуск до захисту у зав. кафедри		виконано

Дата видачі завдання 25 січня 2021 р.

Студент _____
(підпис)

Керівник роботи _____ проф. Смеляков К. С.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Кваліфікаційна робота магістра містить: 72 сторінок, 15 рисунків, 15 таблиць.

ДЕРЕВА РІШЕНЬ, КРИПТОВАЛЮТА, МЕТРИКИ, НЕЛІНІЙНИЙ РЕГРЕСІЙНИЙ АНАЛІЗ, ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ, ФУНДАМЕНТАЛЬНИЙ АНАЛІЗ КРИПТОВАЛЮТИ.

Об’єктом дослідження є закономірності ціноутворення криптовалюти та можливість застосування алгоритмів нелінійного регресійного аналізу та алгоритмів передбачення часових рядів для створення системи передбачення вартості криптовалюти. Метою роботи є створення або удосконалення існуючих підходів до створення прогнозу вартості криптовалюти у короткочасній перспективі. Методи дослідження базуються на нелінійному регресійному аналізі факторів, що можуть впливати на вартість криптовалюти. Використані технології: Java, Python та Apache Spark, EC2. Результати роботи можуть бути успішно використані для створення сервісу побудови прогнозів вартості криптовалюти.

DECISION TREES, CRYPTOCURRENCY, METRICS, NONLINEAR REGRESSION ANALYSIS, TIMESERIES FORECASTING, FUNDAMENTAL ANALYSIS OF CRYPTOCURRENCY.

The object of research is the laws of cryptocurrency pricing and the possibility of using non-linear regression analysis algorithms and time series prediction algorithms to create a system for predicting the value of cryptocurrency. The aim of the work is to create or improve existing approaches to creating a forecast of the value of cryptocurrency in the short term. The research methods are based on nonlinear regression analysis of factors that may affect the value of cryptocurrency. Technologies used: Java, Python and Apache Spark, EC2. The results of the work can be successfully used to create a service for building forecasts of the value of cryptocurrency.

Я, Бизкровний Олександр Миколайович, студент групи ІПЗм-20-2, здобувач вищої освіти на другому (магістерському) рівні кафедри «Програмна інженерія», заявляю: моя кваліфікаційна робота на тему «Дослідження ціноутворення криптовалюти та ефективність використання машинного навчання для пошуку тренду вартості», що буде представлена в екзаменаційну комісію для публічного захисту, виконана самостійно, в ній не міститься елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElArKhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання. Я ознайомлений з діючим положенням «Про протидію академічному плагіату ХНУРЕ», згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

ЗМІСТ

ВСТУП

Криптовалюта стає все більш актуальною галуззю розвитку фінансової індустрії. Цифрові гроші стають більш доступними з року в рік, так як процес їх легалізації не стоїть на місці, а також рухається у напрямку використання криптовалюти, як засобу електронного розрахунку та збереження активів.

Популярність криптовалюти обумовлена в її головному принципу – децентралізації, коли жоден орган влади, окрім безпосередніх холдерів мережі блокчейну, не знає звідки прийшли гроші та куди пішли. Та навіть ті, хто записує нові блоки блокчейну, не знають звідки і куди прямують гроші. Такий механізм дозволяє уникати зайвих податків, сторони транзакції не знають нічого про один одного – тобто повна анонімність транзакції, та не має зайвих питань з податкової інспекції, щодо походження коштів. Все це звісно гарно, проте наведені вище пункти також породжують і темну сторону криптовалюти.

Вона поширена серед чорного ринку, терористів, індустрії наркотиків та обігу зброї. Може бути використана для оминання санкцій, що накладені на державу для оплати певних послуг. Також широко використовується в DarkNet мережі, як засіб оплати послуг або товарів, через те що криптовалюта – це анонімний спосіб розрахунку. Криптовалюта, як і готівка – певного роду папірці, які можна комусь передати,проте ми не знаємо того кому ми передаємо та чи тій людині, що необхідно.

Окрім всіх озвучених фактів, криптовалюта – це нестабільний інвестиційний актив, що має іншу природу ціноутворення, ніж цінні папери та валюта в цілому.

Ця робота націлена на дослідження природи ціноутворення криптовалюти та автоматизації процесу виконання аналізу ціноутворення криптовалюти.

Методами дослідження для досягнення мети кваліфікаційної роботи було використання нелінійних моделей регресійного аналізу факторів, що є метриками середовища криптовалюти, а також використання алгоритмів для визначення тренду кожного з факторів криптовалюти для побудови прогнозу і подальшого використання створеними моделями регресійного аналізу.

Криптовалюта та Bitcoin, в цілому, демонструє цінність крайні роки, що

підтверджується 14 мільйонами коінів в обігу. Інвестори, що бачуть майбутні можливості цієї технології створюють більшу частину капіталізації і цей процес більш за все буде продовжуватись до их пір, поки ринок остаточно не прийме криптовалюту та не стабілізує її вартість.

Технологія блокчейн, що лежить в основі криптовалюти, може бути використана і в інших галузях фінансів. Покупка та утримання акцій, облігацій та інших фінансових інструментів може бути виконана також через блокчейн. Зараз є можливість у продажу NFT за яким користувачі найчастіше продають предмети мистецтва. Тобто технологія блокчейну все більше і більше входить в звичайне життя людей.

Окрім фінансового сектору технологія блокчейн також має свої відгуки і в сфері страхування: домогамає автоматизувати функції страхування шляхом автоматичної перевірки покриття страховки між компаніями та страховими особами. Також блокчейн не оминув і медицинську сферу, де блокчейн виконує роль способу передачі даних про пацієнтів, медичні записи, рецепти ліків. Блокчейн вніс зміни і до сфери нерухомості, де з’явилась можливість у зв’язку між продавцев та покупцем напряму без посередників, що в значній мірі зменшує витрати покупців.

Зараз існує велика кількість моделей, алгоритмів та технологій, що забезпечують безпеку від шахраїв, покращують якість передбачення вартості різних інструментів інвестицій, покращують ефективність майнінгу, допомагають формувати портфель інвестицій. Проте такого роду алгоритми часто є не вичерпними, тобто втрачають з поля зору деякі фактори, що можуть в значній мірі мати вплив на вартість.

З огляду на це, виникає необхідність у створенні більш глибокого аналізу впливу різних факторів на ціноутворення криптовалюти, для того аби побудувати систему, що матиме змогу у передбаченні вартості криптовалюти.

1 АНАЛІТИЧНИЙ ОГЛЯД

1.1 Загальні відомості

Криптовалюта – ресурс, що з’явився нещодавно (відносно інших варіантів фінансів) та набирає обертів. Нажаль, все, що нове, не завжди сприймається відкрито. Криптовалюту це також не оминуло і вона у більшості країн не признана законом, як фінансовий ресурс. Така ситуація існує через те, що не має певних важелів, або вони знаходяться у початковому етапі для регулювання цього ресурсу.

Криптовалюта вважається платіжним засобом, яким зазвичай користуються кримінальні суб’єкти суспільства, аби деперсоналізувати себе у якихось нелегальних оборотах коштів. Проте, те ж саме можна сказати і про готівку, бо люди, які її віддають та приймають, не завжди бувають знайомі одна з одною, не завжди готівка є легальною та не завжди можна прослідкувати куди звідки та скільки готівки переходить від одного суб’єкта до іншого.

Криптовалюта створена на технології блокчейну, що забезпечує анонімність транзакцій. Якщо розглядати Bitcoin, то кожен блок ланцюга зберігає в собі хеш відправника, хеш отримувача і інформацію транзакції. Більш ніякої інформації щодо того, хто саме є отримувачем, нічого про те чи отримав адресат валюту і тд.

Окрім цього криптовалюта досить безпечна до взлому, тому що кожен наступний блок залежить від хешу попереднього і різні частини ланцюга блокчейну зберігаються на багатьох комп’ютерах мережі, що унеможливлює підміну частини блокчейну з метою перезапису або підміни ланцюга. Загалом, проблема перезапису частини ланцюга носить назву «Проблема 51%», де йдеться про те, що якщо виникне така потужність комп’ютерів, яка зможе обігнати 51% комп’ютерів мережі за продуктивністю, то в такому випадку є велика загроза виконання підміни блокчейну. Для того, аби уникнути цієї проблеми зі збільшенням кількості комп’ютерів в мережі, збільшується складність алгоритму для створення нового блоку в блокчейні. Таким чином, навіть якщо колись буде комп’ютер, що зможе виграти у 51% комп’ютерів мережі, то буде штучно збільшена складність алгоритму вираховування нового блоку блокчейну.

Існує безліч криптовалют та безліч ідей та технологій, які вони реалізовують. Завдяки тому, що є можливість у використанні криптовалюти, збільшується на неї попит і відповідно її вартість.

Криптовалюта вже давно стала дуже небезпечним для інвестування активом, проте з дуже великою прибутковістю. Вона популярна серед трейдерів, так як спекуляція цим активом – це те, що дає гарний прибуток. Проте, аби цей інструмент зміг бути прибутковим, а не навпаки, необхідно вміти його використовувати.

Якщо переглянути те, що є наразі на біржах – це акції компаній, валюта, дорогоцінні метали, облігації, а також криптовалюта, яка є програмним забезпеченням зі специфічним способом використання, що дозволяє включати її на валютний ринок і здійснювати торгівлю. У роботах [1-3] представлено сучасний огляд криптовалютних систем, моделей та алгоритмів. Зокрема, у роботі [1] проведено порівняльний аналіз алгоритмів майнінгу, у [2, 3] сформульовано основні проблеми та можливості, а також аналіз основних методів штучного інтелекту, які використовуються для вирішення найважливіших проблеми криптовалюти, пов’язані з прогнозуванням цін, ризиками, загрозами кібербезпеці та рядом інших. Основна ідея криптовалюти типу «coin» — можливість виконувати анонімні транзакції [4]. У роботі [5] показано особливості моделі децентралізованої конфіденційної платіжної системи. Крім того (основне обмеження щодо типу криптовалюти), існують інші типи криптовалют, але це дослідження орієнтоване лише на дослідження факторів, які впливають на тип криптовалюти «монета» [6-8]. Усі криптовалюти належать до ПЗ, і їх ціна є складним аналізом багатьох факторів [9]. У глобальному масштабі є дві частини, які містять власні фактори: сам продукт (криптовалюта як продукт і їх механізм) і торговий ринок. Ніщо не живе поза середовищем, і криптовалюта не є винятком. Зв'язок між вибором факторів, моделей та алгоритмів функціонування екосистеми криптовалюти блокчейн описано в роботах [10-12].

Якщо звернути увагу на те яким чином обираються кандидати для інвестування у не криптовалютний сектор, то можна виділити декілька етапів відбору, серед яких є аналіз компанії, аналіз галузі, розвиток індустрії та інше.

Криптовалюта ж має іншу природу, аніж звичні компанії, нерухомість, дорогоцінні метали, так як це перш за все програмне забезпечення. Виникає питання яким чином аналізувати програмне забезпечення, або технологію, що створена ним. Існує низка фундаментальних метрик, що використовуються для аналізу криптовалюти:

а) on-chain метрики:

1) кількість транзакцій. Кількість транзакцій, що відбуваються в мережі. Слід зауважити, що до цього показника слід ставитися з обережністю, тому що не має впевненості, що транзакції не відбуваються між гаманцями однієї особи, щоб збільшити активність у мережі;

2) об’єм валюти, що продається за певним проміжком часу;

3) активні адреси. Кількість активних адрес або загальна кількість адрес відправників та отримувачів в кожній транзакції. Окрім цього ця метрика ще визначає кількість унікальних адрес за певний проміжок часу;

4) fees paid. Визначення попиту на блок криптовалюти. Якщо провести аналогію з традиційними біржами, то це чим прибуткова ціна на ордер, тим скоріше його викуплять. Схожим чином регулюється попит на блок криптовалюти. З плином часу збільшується вартість добування криптовалюти, а прибуток з кожного блоку, що надається мережею зменшується (це явище носить назву network halving), тому виникає необхідність у збільшенні надбавки за підтвердження транзакції, створювачем транзакції. Якщо надбавку за підтвердження транзакції не збільшувати, то майнери будуть працювати у збиток;

5) hash rate. При використанні алгоритмів майнингу, що базуються на POW принципах, то при збільшенні попиту на майнинг криптовалюти, необхідно збільшувати складність хешу для того аби уникнути проблеми 51%. Зменшення складності хешу сигналізує про те, що інтерес до цієї криптовалюти є невеликим. (інтерес втрачається, або ще не досягнув великих розмірів). При використанні алгоритмів майнингу, що базуються на POS принципах, то кількість монет, що є на руках у майнерів і є оцінкою попиту на криптовалюту;

б) проектні метрики:

1) white paper. Документ, що дає відповіді на наступні питання:

- які технології використовуються;
- які варіанти використання мережі;
- шлях розвитку, нові функції;
- що саме розповсюджує мережа; (монети чи токени)

2) команда. Дослідження команди, яка працює над продуктом. Чи приймала команда участь у проектах, що досягли успіхів;

3) конкуренти. White paper повинна надати точну ідею використання криптовалюти для розуміння хто є її конкурентом;

4) токеномія та початкова дистриб'юція. Існує декілька варіантів використання криптовалюти: коіни та токени. Використання токенів для вирішення проблеми предметної області токенами не завжди є правильним та вигідним рішенням. Необхідно звернути увагу яким чином надається можливість у розповсюдженні та використанні токенів. Початкове розповсюдження криптовалюти є гарним способом вийти на ринок. Існує декілька способів для досягнення цієї мети ICO та IEO, що означають відповідно Initial Coin Offering та Initial Exchange Offering. Перший із способів розповсюдження є більш вразливим, тому що коіни розповсюджує команда. В іншому ж випадку коіни розповсюджує біржа, що вступає в певне партнерство з командою. Той факт, що біржа пропонує певні коїни для покупки, є певною гарантією для користувачів, що криптовалюта не є мильною кулею та має перспективи розвитку. В цілому, необхідно звернути увагу яка кількість коінів доступна для майнингу та маркету, а яка залишається команді розробників. Якщо команда розробників має велику частину всього об'єму, то виникає ризик, що вони зможуть керувати значною частиною ринку даної криптовалюти та робити хвилювання курсу;

в) фінансові метрики:

1) ринкова капіталізація. Дана метрика вказує на загальну вартість мережі, тобто це добуток однієї одиниці на загальну кількість криптовалюти доступної на ринку;

2) ліквідність та об'єм. Ліквідність вказує на те як швидко та легко

криптовалюта може бути продана. Trading volume – це метрика, що допомагає визначити ліквідність криптовалюти. Вираховується за допомогою визначення кількості проданої валюти за день;

3) stock to flow модель. Метрика, яка вказує на те з якою швидкістю добувається ресурс. Вираховується за формулою: Загальна кількість ресурсу поділена на добуту кількість ресурсу за рік. Чим більше це відношення, тим більш придатний ресурс для довгострокового інвестування;

4) максимальне споживання. Кількість криптовалюти, що може бути видобута. Кількості криптовалюти більше цієї цифри не може бути створено або видобуто;

5) циркулююча кількість валюти. Кількість криптовалюти, що публічно доступна та знаходиться на ринку.

Окрім даних метрик існує велика кількість додаткових, що також можуть братись до аналізу криптовалюти.

Криптовалюта — це звичайне програмне забезпечення, тому загальні правила, які стосуються будь-якого продукту в цій області, також можуть вплинути на криптовалюту. Приклади таких функцій описані в [13-15]: у кожного товару є конкуренти; кожен продукт повинен надавати певні особливості, щоб вижити; кожен товар залежить від купівельної спроможності потенційних клієнтів тощо.

Загалом, у всіх криптовалютах є валідатори транзакцій, роль яких відіграють майнери, що видобувають кожен з наступних блоків блокчейну. У результаті виникають дві ролі, потреби яких необхідно задовольнити. Якщо користувачі не матимуть можливості використовувати крипто-монети, то криптовалюта може померти [16]. Інша частина цього: якщо майнерам не буде вистачати прибутку для покриття всіх витрат, транзакції будуть схвалені з величезною затримкою, що призведе до зниження популярності криптовалюти, і, як наслідок, ця причина може бути основною причиною криптовалюти піти з ринку [17, 18].

2 АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ АБО АЛГОРИТМІВ

Існує безліч методів та алгоритмів у передбаченні вартості криптовалюти або виявленні факторів, що мають вплив на її утворення. Деякі з них включають більше факторів та є більш точними, а деякі взагалі не відповідають принципам утворення вартості криптовалюти. Буде розглянуто наступні методи аналізу та передбачень:

– моделювання кількості активних користувачів криптовалюти (Bitcoin) базуючись на законі Metcalfe`s [19];

– аналіз факторів, що впливають на вартість Bitcoin.

Перший метод аналізу заснований на наступному твердженні: «Вартість мережі пропорційна квадрату кількості користувачів у ній». Ця модель була запропонована Timothy Peterson в основі якої є Gompertz функція замість логістичної функції. На рисунку 1 зображено передбачення збільшення кількості активних користувачів.

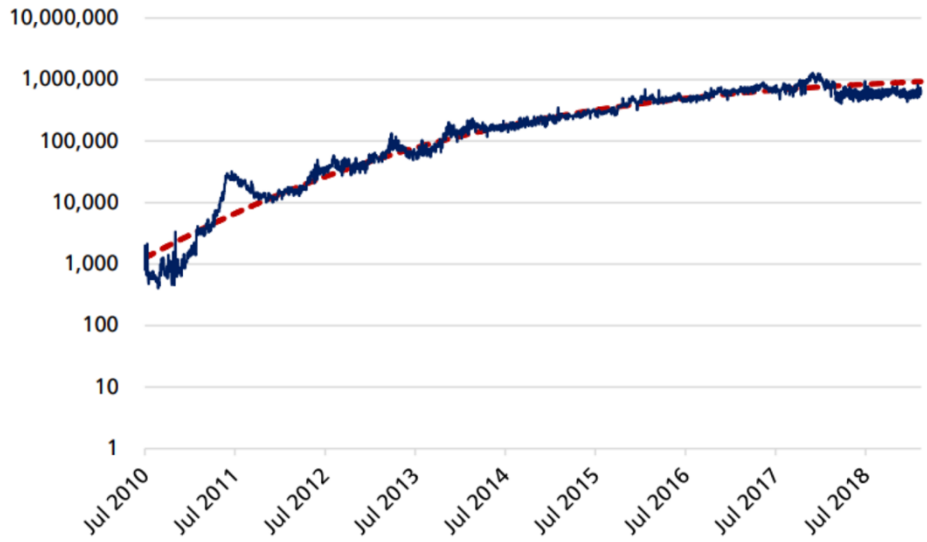

Рисунок 1 – Передбачення створеної моделі [19]

Функція для розрахунків виглядає наступним чином:

$$(f)_t = ae^{-e^{b-ct}},$$

де a є асимптотою, оскільки $\lim_{e \to infinity} ae^{-e^{b-ct}} = ae^0 = a$; b використовується для переміщення по осі X; c – відповідає за показчик росту (вісь Y); e – число Ейлера.

Дана функція використовується для навчання моделі на числі активних користувачів. Окремо цей метод аналізу від інших методів аналізу використовуватись не може, так як існує безліч факторів, які мають вплив на вартість криптовалюти, окрім прогнозування лише кількості активних користувачів та залежності вартості від неї.

Наступний метод націлений на виявлення факторів, що мають найбільший вплив на вартість Bitcoin. Для того, аби дослідити це, було використано Autoregressive Distributed Lag та Generalized Autoregressive Conditional Heteroscedasticity підходи [20]. Для більш точного виявлення факторів, що можуть впливати на вартість Bitcoin було обрано періоди, до різкого проявлення попиту на Bitcoin та після (2017 рік), щоб виявити чи такий же самий вплив мають фактори з яких складається модель та чи може з часом вплив змінюватись факторів.

Якщо звернутись до питання вибору даних для формування моделі, то залежною змінною моделі є курс обміну Bitcoin та USD. Окрім цього для уникнення проблеми автокореляції, денні дані було модифіковано у середні за тиждень, а також, якщо певних змінних не вистачало, то такий рядок видалявся. Дані для моделі було отримано з різних джерел (див. табл 1).

Таблиця 1 – Опис змінних, що приймають участь у створенні моделей

Змінна	Опис	Джерело
BTC	Курс обміну на біржі	Quandl
Hashrate	Складність алгоритму для створення нового блоку.	Quandl
Volume	Загальний обсяг видобутой та використаной валюти	Quandl

Кінець таблиці 1

Змінна	Опис	Джерело
S&P 500	Індекс, що показує 500 найбільших компаній США	Thomson Reuters Eikon
Gold	Індекс вартості золота	Thomson Reuters Eikon
Нафта	Спотова ціна за барель	Thomson Reuters Eikon
Vix	Вимір волатильності маркету S&P500.	Thomson Reuters Eikon
Google	Тижнева статистика за словом Bitcoin	Google Trend

Змінна S&P 500 – це гарний індикатор, який показує стан економіки США, а VIX використовується для визначення впевненості ринку щодо направлення руху індексу S&P 500 у наступні 30 днів.

Нафта та золото вважаються одними із найважливіших ресурсів, зміна вартості яких впливає на весь світ, а якщо ці ресурси мають вплив на світ, то і мають вплив на, наприклад, вартість добування криптовалюти, шляхом зміни вартості електроенергії або вартості апаратного забезпечення. Отже, ці фактори також можуть бути включені в модель. Для визначення залучення уваги аудиторії Bitcoin була використана статистика Google Trends.

Щодо алгоритмів, що були використані для проведення експерименту, то ARDL використовується для визначення короткострокових або довгострокових зв’язків між групами змінних включаючи лаги для залежних та незалежних змінних.

GARCH – це авторегресивна умовно гетероскедатична модель, що використовується для опису і моделювання часових рядів, коли є підстави вважати, що в кожному відрізку часу, дисперсія часового ряду залежить від різних параметрів і не є сталою.

Для експерименту було утворено три моделі з різним набором змінних для визначення відсутності якого з факторів є найбільш впливовим на передбачення:

1. $\Delta lnBTC_t = \alpha + \beta 1 \Delta lnBTC_{t-1} + \beta 2 \Delta lnVolume_t + \beta 3 \Delta lnSP500_t + \sum_{p=1}^{n} \beta 4 \Delta lnOil_{t-p} + \sum_{p=1}^{n} \beta 5 \Delta lnGold_{t-p} + \beta 6 \Delta lnVIX_t + \sum_{p=2}^{n} \beta 7 \Delta lnGoogle_{t-p} + Trend + \varepsilon_t.$
2. $\Delta lnBTC_t = \alpha + \beta 1 \Delta lnBTC_{t-1} + \beta 2 \Delta lnVolume_t + \beta 3 \Delta lnSP500_t + \sum_{p=2}^{n} \beta 5 \Delta lnGoogle_{t-p} + Trend + \varepsilon_t.$
3. $\Delta lnBTC_t = \alpha + \beta 1 \Delta lnBTC_{t-1} + \beta 2 \Delta lnHashrate_t + \beta 3 \Delta lnVolume_t + \beta 4 \Delta lnSP500_t + \sum_{p=1}^{n} \beta 4 \Delta lnOil_{t-p} + \sum_{p=1}^{n} \beta 5 \Delta lnGold_{t-p} + \beta 6 \Delta lnVIX_t + \sum_{p=2}^{n} \beta 7 \Delta lnGoogle_{t-p} + Trend + \varepsilon_t$

Наступні таблиці відображають результати експерименту, де розглядались дані у наступних часових проміжках:

– перший тиждень 2013 року – 7 тиждень 2018 року;
– перший тиждень 2013 року – 25 тиждень 2016 року;
– перший тиждень 2017 року – 7 тиждень 2018 року.

Таблиця 2 – Результати ARDL та GARCH моделей для першого набору змінних.

	ARDL			GARCH		
Період/Змінні	1	2	3	1	2	3
$\Delta lnBTC_{t-1}$	0.19 (2.23) **	0.222 (2.14) **	0.206 (1.04)	0.225 (5.43) ***	0.329 (6.44) ***	0.293 (4.98) ***
$\Delta lnVolume_t$	−0.042 (1.4)	−0.027 (0.79)	−0.134 (2.33) **	−0.046 (2.61) ***	0.022 (1.26)	-0.15 (0.62)
$\Delta lnSP500_t$	1.772 (2.16) **	2.55 (2.69) ***	−1.707 (1.04)	1.038 (1.59)	1.272 (1.85) *	1.318 (1.90) *
$\Delta lnOil_t$	−0.072 (0.5)	−0.075 (0.47)	−0.141 (0.40)	−0.001 (0.00)	0.021 (0.17)	0.005 (0.04)
$\Delta lnOil_{t-1}$	0.142 (0.95)	0.147 (0.87)	0.341 (0.77)	0.023 (0.18)	0.027 (0.22)	0.005 (0.04)

Кінець таблиці 2

	ARDL		GARCH			
Період/Змінні	1	2	3	1	2	3
$\Delta lnGold_{t-1}$	−0.415 (0.62)	−0.384 (0.49)	−0.337 (0.35)	0.049 (0.18)	0.068 (0.24)	0.048 (0.17)
$\Delta lnVIX_t$	0.029 (0.34)	0.126 (1.34)	−0.279 (1.92) *	0.008 (0.12)	−0.039 (0.56)	−0.186 (0.93)
$\Delta lnGoogle_t$	0.109 (3.60) ***	0.102 (2.84) ***	0.140 (3.16) ***	0.045 (2.85) ***	0.030 (1.61)	0.022 (1.18)
$\Delta lnGoogle_{t-1}$	0.105 (3.46) ***	0.093 (2.58) **	0.176 (4.04) ***	0.088 (4.27) ***	0.081 (4.34) ***	0.076 (3.86) ***
$\Delta lnGoogle_{t-2}$	0.082 (2.18) **	0.088 (1.98) **	0.022 (0.39)	0.053 (2.76) ***	0.062 (3.35) ***	0.057 (3.00) ***
Adjusted R2	0.29	0.23	0.54			
Observations	264	205	56	264	205	56

Як показано в таблиці 2, лаг Bitcoin, здається, має значний позитивний вплив на ціну біткойна на рівні 5%. Якщо його прибутковість минулого тижня вища на 1%, то, за оцінками, прибуток цього тижня буде вищим на 0,19%.

Відмінність S&P 500 на рівні 5% і має позитивний сигнал, коли S&P500 зростає на 1%, ціна Bitcoin зростає на 1,77%. На відміну від цього, VIX, Oil, Gold і Volume, не мають значного впливу на ціну за прогнозований період. Різниця у змінній Google Trends та її відставання є значними на рівні 1%. Термінові ефекти показують, що коли тенденції Google зростуть на 1%, очікується, що ціна зросте на 0,11%. Враховуючи різницю Google, короткостроковий вплив пошуку Google на ціну біткойна становить 0,22%. Крім того, якщо включити другу різницю Google, яке є значним на рівні 5%, загальний короткостроковий вплив пошуку Google на

ціну біткойна становить 0,30%. Нарешті, довгостроковий вплив тенденцій Google на ціну біткойна становить 0,37%.

У моделі GARCH лаг Bitcoin має майже ідентичний ефект, як і в моделі ARDL, і є значним на рівні 1%. Google і його два лаги є значними на рівні 1%, що майже так само, як і в моделі ARDL, хоча коефіцієнт як для першої різниці, так і для двох лагів зменшився. Крім того, S&P 500 визнано незначним, тоді як у моделі ARDL він був визнаний значним. Подібно до моделі ARDL, VIX, нафта та золото незначні. Обсяг є значним на рівні 1%, що не узгоджується з моделлю ARDL. Ефекти ARCH позитивні та значущі на рівні 1%, що вказує на те, що значні зміни у дисперсії два тижні тому матимуть вплив приблизно на 56,2% на волатильність на наступному тижні. Ефекти GARCH значні на рівні 5%. Це значення вказує на те, що 31,5% волатильності минулого тижня впливають на волатильність цього тижня. Сума ефектів ARCH і GARCH становить приблизно 87,7%, що показує збереження всієї нестабільності та стрімких змін минулого тижня, а також вплив, який він має на цей тиждень.

Таблиця 3 – Результати ARDL та GARCH моделей для другого набору змінних.

	ARDL			GARCH		
Період/Змінні	1	2	3	1	2	3
$\Delta lnBTC_{t-1}$	0.187 (2.05) **	0.226 (2.05) **	0.065 (0.46)	0.215 (5.24) ***	0.318 (6.05) ***	0.293 (5.01) ***
$\Delta lnSP500_t$	1.411 (3.45) ***	1.364 (2.99) ***	1.59 (1.62)	0.926 (2.76) ***	0.873 (2.62) ***	0.779 (2.27) **
$\Delta lnGoogle_t$	0.105 (3.50) ***	0.099 (2.79) ***	0.100 (1.82) *	0.033 (2.43) **	0.023 (1.5)	0.09 (0.99)

Кінець таблиці 3

	ARDL			GARCH		
Період/Змінні	1	2	3	1	2	3
$\Delta lnGoogle_{t-2}$	0.077 (2.01) **	0.084 (1.88) *	0.061 (1.06)	0.047 (2.63) ***	0.06 (3.35) ***	0.055 (2.84) ***
$Adjusted^2$	0.29	0.23	0.50			
Observations	264	205	56	264	205	56

Як показано в таблиці 3, зв'язок між лагом і ціною Bitcoin майже такий самий, як і в моделі 1, і є значним на рівні 5%. Якщо ціна біткойна минулого тижня зросла на 1%, це означає зростання ціни цього тижня на 0,19%. Більше того, S&P 500, схоже, має значний вплив на ціну біткойна, подібно до моделі 1. Ця змінна є значною на рівні 1%. Коли S&P 500 зросте на 1%, біткойн, за оцінками, зросте на 1,41%. Тенденції Google мають той самий рівень значимості, що й у моделі 1, і майже рівні коефіцієнти.

Короткостроковий ефект від пошуків у Google становить 0,11%, а загальний короткостроковий ефект становить 0,21%. Загальний довгостроковий ефект становить 0,34%.

У моделі GARCH лаг Bitcoin має майже ідентичний ефект, як і в моделі ARDL, і є значним на рівні 1%. Індекс S&P 500 також є значущим у моделі GARCH, як і в моделі ARDL, але з дещо нижчим коефіцієнтом. Google і його два лаги є значними на рівнях 5% і 1% відповідно, що майже узгоджується з модель ARDL. Однак коефіцієнт як для першої різниці, так і для лагу зменшився. Ефекти ARCH позитивні та значущі на рівні 1% і впливають приблизно на 58,1% на волатильність наступного тижня. Ефекти GARCH позитивні та значущі на рівні 1%. Близько третини (32,4%) волатильності минулого тижня впливає на волатильність цього тижня. Сума ефектів ARCH і GARCH становить приблизно 90,5%.

Таблиця 4 – Результати ARDL та GARCH моделей для третього набору змінних.

	ARDL			GARCH		
Період/Змінні	1	2	3	1	2	3
$\Delta lnBTC_{t-1}$	0.19 (2.25) **	0.222 (2.10) **	0.206 (1.66)	0.225 (5.28) ***	0.329 (6.28) ***	0.258 (3.89) ***
$\Delta lnHashrate_t$	0.067 (0.96)	0.22 (0.26)	0.274 (2.51) ***	0.031 (0.50)	−0.005 (0.08)	−0.039 (0.57)
$\Delta lnVolume_t$	−0.041 (1.36)	−0.027 (0.78)	−0.139 (2.51) **	0.04 (2.64) ***	−0.022 (1.27)	−0.139 (0.34)
$\Delta lnSP500_t$	1.725 (2.11) **	2.532 (2.68) ***	−1.952 (1.33)	1.023 (1.55)	1.274 (1.86) *	1.472 (1.98) **
$\Delta lnOil_t$	−0.069 (2.11) **	−0.075 (0.47)	−0.073 (0.22)	0.008 (0.06)	0.02 (0.16)	−0.012 (0.10)
$\Delta lnOil_{t-1}$	0.137 (0.92)	0.145 (0.85)	0.324 (0.77)	0.019 (0.15)	0.028 (0.22)	0.067 (0.53)
$\Delta lnGold_t$	0.53 (1.01)	0.537 (0.90)	0.918 (0.87)	−0.03 (0.13)	−0.004 (0.02)	−0.061 (0.24)
$\Delta lnGold_{t-1}$	−0.404 (0.60)	−0.38 (0.49)	−0.393 (0.41)	0.04 (0.15)	0.068 (0.23)	0.020 (0.08)

Кінець таблиці 4

	ARDL			GARCH		
Період/Змінні	1	2	3	1	2	3
$\Delta lnGoogle_t$	0.108 (3.58) ***	0.102 (2.84) ***	0.118 (2.53) **	0.046 (2.83) ***	0.03 (1.57)	0.021 (1.09)
$\Delta lnGoogle_{t-1}$	0.104 (3.41) ***	0.093 (2.56) **	0.165 (3.96) ***	0.088 (4.68) ***	0.081 (4.32) ***	0.076 (3.41) ***
$\Delta lnGoogle_{t-2}$	0.081 (2.17) **	0.088 (1.98) **	0.014 (−0.25)	0.056 (2.87) ***	0.062 (3.35) ***	0.055 (2.87) ***
$Adjusted^2$	0.29	0.23	0.54			
Observations	264	205	56	264	205	56

Модель, представлена в таблиці 4, включає змінну Hashrate, але в іншому схожа на модель 1. Властивості, які відображаються змінними, та їхні результати також подібні до результатів моделі 1. Однак перша відмінність Hashrate має позитивний знак у всіх розрахункові періоди, але значний лише в третьому періоді, з 2017 року першого тижня до 2018 сьомого тижня.

У результаті виконаних досліджень було встановлено певні залежності між факторами, які мають вплив на формування вартості криптовалюти (Bitcoin), проте це дослідження не слід вважати вичерпним. Воно лише вказує на напрямок того, які фактори необхідно враховувати при прогнозуванні ціни.

Дане дослідження показало, що обсяги пошуку в Google відіграють не аби яку роль у інформуванні виникання тренду. А також, що hashrate не може використовуватись як змінна, що впливає на вартість, тому що вона залежить навпаки від попиту на криптовалюту. Окрім цього, дослідження показало, що коли економічна ситуація стабільна та виникає тенденція до позитивного росту ринку, то інвестори більш відкриті для інвестування в ризикові активи.

Інша проаналізована робота [21] використовує алгоритми машинного навчання GRU, LSTM та bi-LSTM, моделі котрих навчені на даних, змінні яких наведені далі: ціна відкриття, найвища ціна, найнижча ціна, ціна закриття, дата.

Фактори, що були відібрані в цій роботі є недостатньо інформативними, на мою думку, тому що вони не визначають першопричини певного значення ціни на певну дату. Іншими словами, фактори не є описовими. Незважаючи на це припущення, процес перевірки моделі показує наступні результати (Рисунок 2, Рисунок 3).

Рисунок 2 – Фактична та прогнозована ціна BTC за допомогою LSTM [21]

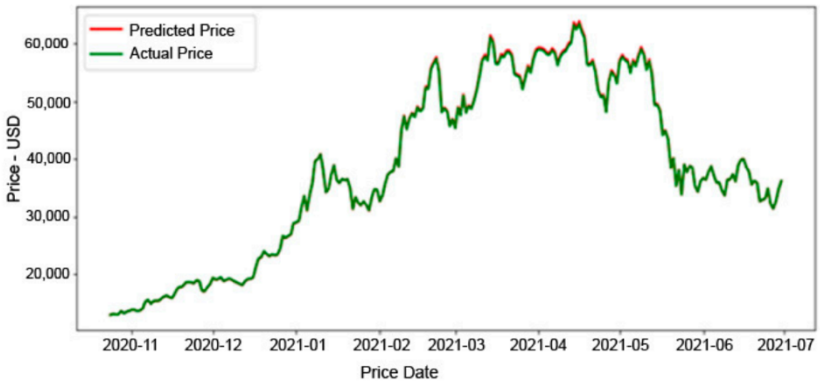

Рисунок 3 – Фактична та прогнозована ціна BTC за моделлю GRU [21]

У розглянутій роботі немає опису того, яким саме чином моделі були навчені та перевірені. Та не зважаючи на отримані результати, існує гіпотеза [22], що історичні значення цін ніколи не можна використовувати для прогнозування цін у майбутньому.

Інша робота [23] використовує технічні показники криптовалюти для формування набору данных для прогнозування ціни, використовуючи FFNN, LSTM та BNN алгоритми для створення моделей: SMA, EMA, Momentum (MOM), MACD, RSI. Загальну модель, що використовувалась для експерименту, описано нижче (Рисунок 4).

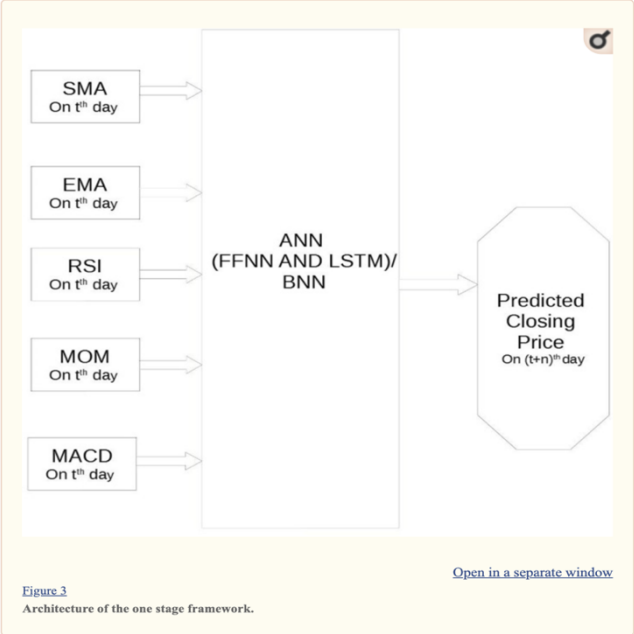

Рисунок 4 – Архітектура моделі [23]

Технічні показники також не можуть бути вичерпним джерелом даних для навчання ML моделі. В цілому, більш менш повний аналіз включає в себе окрім технічного аналізу ще й фундаментальний, який в свою чергу включає: статистику мережі криптовалюти, статистику з бірж та брокерів, світовий економічний стан тощо. Однією з основних цілей даного дослідження є визначення інформативних факторів, які можуть бути використані для прогнозування вартості криптовалюти, а також визначення ефективності використання нелінійних регресійних моделей машинного навчання [24] та виявлення найбільш підходящого алгоритму для вирішення зазначеної вище задачі.

3 ПОСТАНОВКА ЗАДАЧІ

3.1 Загальна мета

Головної метою даного дослідження є створення або покращення алгоритмів аналізу та моделей машинного навчання для визначення тренду вартості криптовалюти. Для досягнення цієї мети необхідно дослідити доцільність використання окреслених вище факторів, що можуть мати вплив на тренд вартості. Окрім цього, вже досліджені роботи є підґрунтям для подальших досліджень. Також необхідно визначити шляхи передбачення короткострокового тренду для наведених факторів для побудови прогнозу.

Для виконання задач поставлених для цього досліду, необхідно використовувати алгоритми регресійного аналізу, щоб з’ясувати чи мають фактори певну залежність один від одного.

Необхідно звернути увагу на те, що політичні, техногенні та інші типи катастроф, що можуть виникнути у світі, не використовуються як чинники чи фактори в моделях.

3.2 Опис даних для створення регресійних моделей

Дані досліджуваної проблеми мають характер часових рядів, тому що фактори та метрики світової економіки оновлюються у часових проміжках: від секунди до дня. Тренувальний набір даних включає лише загальноекономічні фактори світу та дані досліджуваних криптовалют. Виключенням є дані, що розподіляються між країнами, так як ціна на криптовалюту глобальна та не має локальної вартості.

Також, необхідно зазначити, що фактори іноді не співпадають за датами, так як криптобіржі цілодобово без вихідних, на відміну від звичайних бірж, що мають п’ятиденний робочий тиждень. То ж тут необхідно провести підготовку даних. Наприклад, використовувати дані п’ятого дня зі звичайних бірж протягом шостого та сьомого дня на криптобіржах.

Окрім цього факту, ще виникає питання у неповноцінності даних або відсутності певних значень у рядках. Такі рядки можуть видалятись повністю або

пропущені значення можуть бути відновлені шляхом визначення поточного значення за допомогою сусідніх значень. До таких методів можна віднести:

– визначення простої (середньоарифметична) змінної середньої;

$$\widetilde{Y_t} = \frac{\sum_{i=t-p}^{r+t} Y_i}{2p+1}, p < t < n-p$$

– визначення зваженої (середньозваженої) змінної середньої;

$$\widetilde{Y_t} = \frac{\sum_{i=t-p}^{r+t} p_i Y_i}{\sum_{i=t-p}^{r+t} P_i}, p < t < n-p$$

– експоненційне згладжування.

$$\widetilde{Y_t} = \frac{\sum_{i=t}^{t} p_i Y_i}{\sum_{i=1}^{t} P_i}$$

Якщо набір даних для навчання має великі розміри, то можна використовувати відновлювання значень, проте якщо набір даних має невеликі розміри, то відновлення даних є не дуже гарним способом боротьби з пропущеними значеннями, так як таким чином можна створити не правильні зв’язки між значеннями факторів, що в свою чергу веде до не коректного прогнозування.

Окрім алгоритмів згладжування необхідно виконати масштабування даних – rescale, що допоможе зберігати дані у проміжку між 0 та 1, а отже мати змогу використовувати інші дані, якщо їх джерело буде змінено і матиме інший масштаб.

Тренувальний набір даних не є у вільному доступі в Інтернеті, так як складається з різних частих у CSV форматі, що комбінуються використовуючи інструменти Java та Apache Spark.

Інструмент Apache Spark надає можливості у очистці та перетворенні даних у потрібний формат. Наприклад, механізм MinMaxScaller допомагає систематизувати весь набір даних певної колонки до загального вигляду у рамках від 0 до 1. Також можна використати механізм, що групує всі колонки до однієї, утворюючи вектор, який можна використовувати для тренування моделей.

4 ПЛАНУВАННЯ ЕКСПЕРЕМЕНТУ З РЕГРЕСІЙНИМИ МОДЕЛЯМИ

4.1 Виявлення факторів впливу

Для того аби виявити фактори впливу, необхідно звернутись до природи досліджуваного ресурсу. Даний ресурс – є програмним продуктом, що задовольняє потреби користувачів у різних сферах діяльності. Перша криптовалюта (Bitcoin) використовується як метод проведення анонімних транзакцій [25]. Такого роду криптовалюти відносяться до класу Defi: Децентралізовані фінанси [26]. Окрім цього, криптовалюта може використовуватись і в інших варіантах.

На рисунку 5 представлені ключові ролі та чинники екосистеми будь-якої криптовалюти.

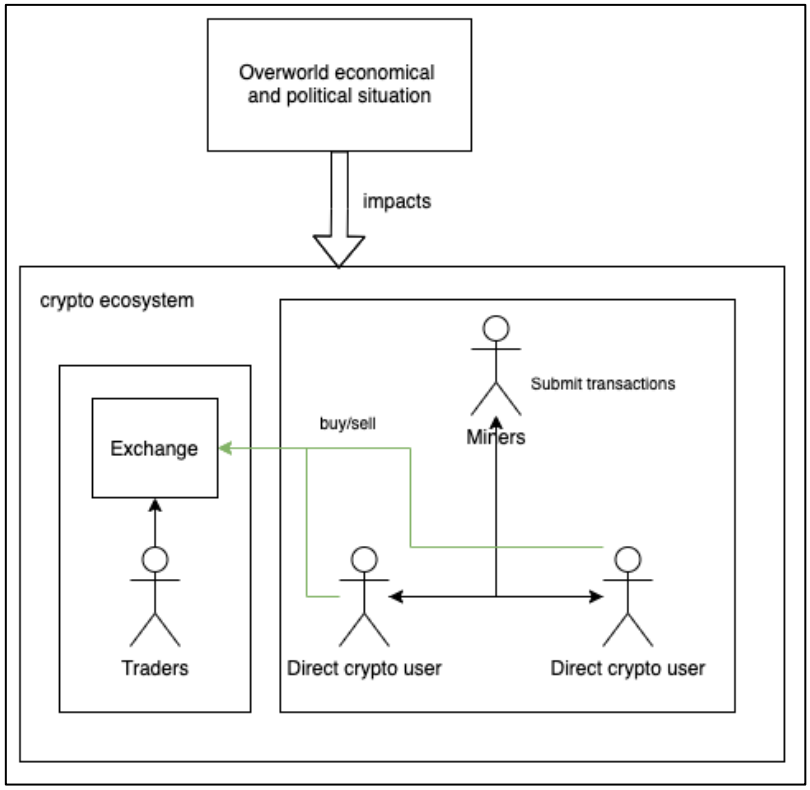

Рисунок 5 – Схема інфраструктури екосистеми криптовалюти

Наведена діаграма відображає основних акторів, котрі існують в екосистемі криптовалюти. Кожна криптовалюта має своїх користувачів, тих, хто підтверджує транзакції, та тих, хто торгує нею. Зовнішні фактори мають вплив на екосистему криптовалюти в цілому, а отже на всіх акторів. Якщо підвести підсумки, то

зовнішні та внутрішні фактори екосистеми криптовалюти мають вплив на кожного з акторів і тому формують вартість.

На рисунку 6 зображена верхнорівнева діаграма, яка описує залежності кожного актора і на що, перш за все, необхідно звернути увагу у аналізі криптоактиву.

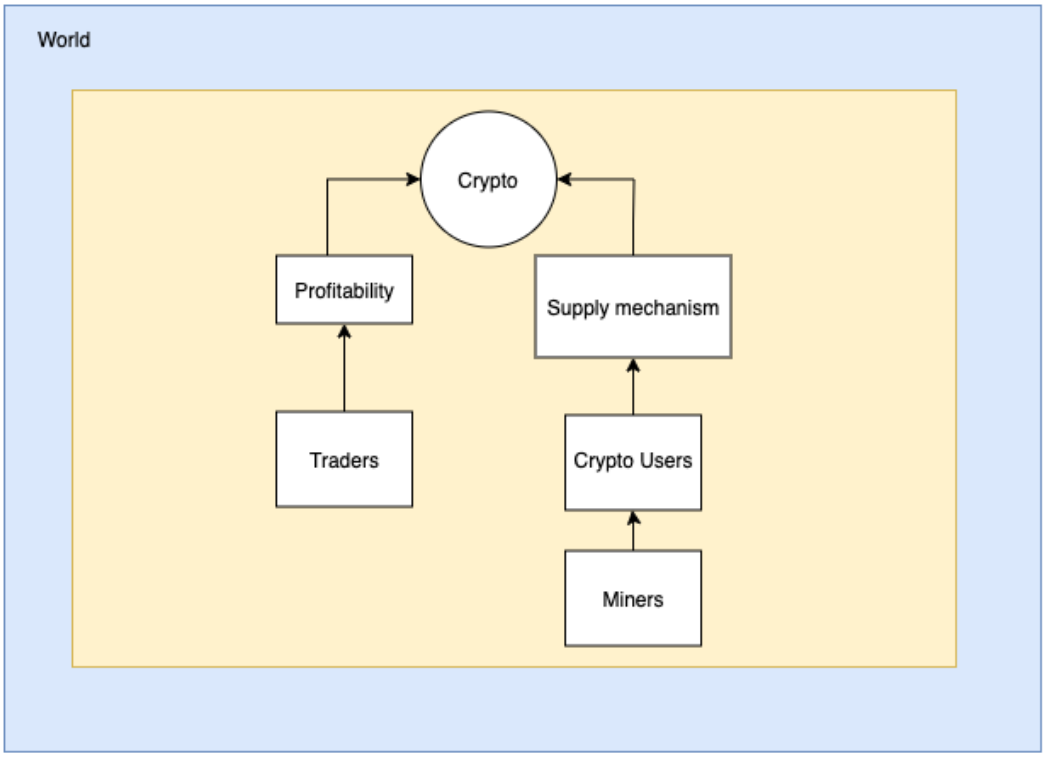

Рисунок 6 – Високорівнева діаграма залежності ролей

Необхідно розглянути кожну роль детальніше, для того, аби окреслити фактори впливу. Ключове поняття для будь-якого ресурсу в плані визначення його вартості – це попит. Користувачі криптовалюти повинні мати змогу реалізувати якимось чином отриману криптовалюту. Саме спосіб реалізації має вплив на попит. Так як, якщо не має шляхів до реалізації отриманого ресурсу, то не має сенсу його добувати, купляти. Способи використання криптовалюти, що були закладені розробниками, закони, що регулюють використання криптовалюти, спекулятивний інтерес трейдерів, конкуренція – є одними із факторів впливу на попит в цілому. Інвестиційним активом криптовалюта поки не може бути у зв’язку з високою мінливістю. Також, необхідно додати, що ті криптовалюти, які створені на хвилі

ажіотажу, котрий існує навкруги якогось явища (наприклад серіал «Гра в кальмара» [27]) не розглядаються для аналізу, так як заздалегідь відомо, що це валюта одного дня, купивши котру, є великий шанс не мати змоги її продати. То ж наступні фактори можна віднести для аналізу того чи є можливості у використанні валюти:

а) світова економічна ситуація. Показниками цього фактору можуть бути метрики, що відображають економічні індекси частин світу:

1) Stoxx 600;
2) S&P 500;
3) Wisdomtree ICBCCS S&P China 500;
4) WisdomTree India Earnings Fund;
5) Topix 500;
6) iShares MSCI South Korea ETF.

Чим більш стабільна економічна ситуація в світі, тим більш позитивно вона впливає на виникання та використання криптовалют. Також необхідно приділити увагу тому, де знаходиться найбільша мережа криптоактиву. В залежності від місцезнаходження найбільшої кількості пулів мережі, треба надавати перевагу індексу інфраструктурної стабільності тієї країни.

Щодо факторів, які впливають на використання криптовалюти, наступні метрики з фундаментального аналізу криптовалюти можуть бути використані:

– white paper. Документ, що дає відповіді на наступні питання: Які технології використовуються? Які варіанти використання мережі? Шлях розвитку, нові функції? Що саме розповсюджує мережа: коїни чи токени;

– дослідження команди, яка працює над продуктом. Чи приймала команда участь у проектах, що досягли успіхів. Якщо в команді є відомі розробники, то шанс на успіх більший;

– дослідження конкурентів є важливим аспектом, тому що аналізована криптовалюта повинна надавати певні переваги у її використанні перед іншими;

– дослідження варіанту розповсюдження криптовалюти: коїни та токени.

Використання токенів для вирішення проблеми предметної області не завжди є правильним та вигідним рішенням. Необхідно звернути увагу яким чином надається можливість у розповсюдженні та використанні токенів;

До зовнішніх факторів також можна віднести тренди пошуку в пошукових системах та ажіотаж навколо криптовалюти у засобах масової інформації. Цей чинник також має вплив на трейдерів, які можуть робити необдумані вчинки і таким чином колихати маркет та відповідно ціну.

Майнер – актор, який залежить від декількох факторів: попит на криптовалюту та вартість добування блоків. Якщо не буде достатнього попиту на криптовалюту аби відшкодувати затрати на апаратне забезпечення або розробники не будуть стимулювати майнерів додатковими надбавками за підтвердження блоків, то така криптовалюта приречена на крах. Наступні фактори впливають на вартість видобування криптовалюти за умови, якщо на неї є попит [28]:

а) вартість електроенергії в певній країні;
б) вартість інтернету в певній країні;
в) вартість апаратного забезпечення;
г) економічний стан виробників апаратного забезпечення:
    1) NASDAQ: NVDA;
    2) NASDAQ: AMD;
    3) ASUSTeK Computer Inc;
    4) NASDAQ: INTC.

Загалом, попит на всю індустрію криптовалюти спричиняє підвищення цін на апаратне забезпечення.

Трейдер, або актор, який не використовує криптовалюту у своїх потребах, а лише спекулює нею, не впливає на криптовалюту напряму, а лише шляхом законів ринку. Трейдери можуть створити хвильоподібні рухи вартості скуповуючи або продаючи активи, створювати фіктивний ажіотаж навколо валюти. Проте, якщо криптовалюта не користується великим попитом і на її курсі не має хвиль вартості – така валюта є не дуже вигідною для спекуляцій, а тому фактор впливу трейдерів

там не великий. До факторів, що відносяться до утворення ринку криптовалюти можна віднести:

– кількість криптовалюти, що залишається у розробників. Якщо ця частка велика, то вони можуть значно колихати ринок у майбутньому;

– «кити» або трейдери, які значним чином впливають на ринок, скуповуючи або продаючи велику кількість активу;

– розповсюдження криптовалюти поміж біржами. Чим більше розповсюдження, тим більший обіг валюти;

– кореляція з іншими криптовалютами. Так як якщо одна з криптовалют, що забезпечує певний сектор своїми послугами втрачає свої позиції через якусь вразливість, то ціна на всі криптовалюти тієї ж індустрії також може піти до низу. Або навпаки, коли один з учасників ринку втрачає свої позиції, то конкуренти відбирають частину споживачів і від цього їх вартість росте;

– створення hard fork може спричинити падіння ціни криптовалюти через те, що частина користувачів підуть з неї до fork. Проте це не завжди так.

4.2 Визначення вхідних даних для формування моделей

Для того аби визначити якого роду будуть використовуватись алгоритми для моделей аналізу, необхідно визначитись з вхідними даними для тренувань моделей. Загалом, всі фактори можна розподілити на загальноекономічні та крипто специфічні.

Наступні фактори тісно пов’язані із середовищем криптовалюти та принципами її роботи. Наведені нижче показники є маркерами різних аспектів всередині криптовалюти протягом її життєвого циклу.

Обсяг притоку – загальна сума (у доларах або токенах), що надходить до гаманців біржових депозитів. Різкі стрибки вверх значень цього фактору, як правило, збігаються з періодами високої волатильності, а іноді й передують їм. Такий сигнал потенційно можна інтерпретувати як ознаку того, що криптовалюта набирає тенденцію до продажу на централізованих біржах.

Обсяг відтоку – загальна сума (у доларах або токенах), що покидає гаманці біржі. Обсяг відтоку часто зростає після значного падіння криптовалюти. Потенційно це можна інтерпретувати як те, що користувачі вирішують тримати свою криптовалюту поза централізованими біржами або сигнал до того, що криптовалюта буде купуватись у найближчий час.

Ціна ETH – залежна змінна в регресійній моделі, що означає вартість досліджуваної криптовалюти -- Ethereum.

Кореляція ETH – BTC – коефіцієнт, що використовується для відображення кореляції між цінами найбільших криптовалют. Таке порівняння необхідно для порівняння того як зміна вартості однієї валюти впливає на іншу. Цю думку можна перенести на звичайний продукт: якщо якийсь товар втрачає довіру покупця, то інші товари в цій сфері також можуть втратити його.

Кількість великих транзакцій – індикатор показує транзакції, за якими було перераховано суму понад 100 000 доларів США.

Обсяг транзакцій у доларах США – фактор, що вказує на загальну суму в доларах, передану транзакціями за певний день. Цей показник може дати уявлення про можливі зміни на ринку криптовалют, якщо величезні обсяги валюти передаються між адресами.

Кількість транзакцій – індикатор показує активність транзакцій у мережах блокчейну певної криптовалюти. Загалом, може показувати загальну поведінку ринку. Якщо кількість транзакцій збільшується, то зростає популярність галузі або конкретної криптовалюти.

Дохід майнерів – показник, що може вказувати на загальну активність майнерів і те, скільки вони заробляють. Величезний дохід майнерів може означати підвищену потребу в них, а отже і у підтвердженні транзакцій.

Відтік коштів від майнерів – індикатор може вказувати те, коли майнери продають свої заощадження в криптовалюті в біржу. Великий обсяг продажів може сигналізувати про вихід з майнингу певної криптовалюти, а це, в свою чергу, про падіння попиту на криптовалюту, що продають.

Прибуток майнерів – метрика описує винагороду майнера. Якщо винагорода низька, то криптовалюта може застрягти з довгою затримкою підтвердження транзакції. Крім того, якщо більша частина винагороди складається з комісії, яку сплачує користувач, то це може сигналізувати про скорочення популярності криптовалюти.

Середня комісія за транзакцію – фактор, що може вказувати на збільшення попиту на певну криптовалюту. У випадку, коли в черзі знаходиться величезна кількість транзакцій на підтвердження, клієнти починають платити додаткову комісію, щоб майнер з більшою вірогідністю вибрав їх транзакцію для обробки першою.

Середній обсяг транзакції – обсяг транзакцій може вказувати як на торгівлю криптовалютою, так і на неспекулятивну діяльність з нею, тобто пряме її використання. Подібно до обсягу торгів, що спостерігається на біржах, обсяг транзакцій може бути корисним для виявлення різких коливань вартості.

Активність GitHub — це декілька факторів, які мають в своєму переліку:

– існуючі проблеми;
– вирішені проблеми;
– кількість спостерігачів в репозиторіях криптовалюти;
– кількість форків;
– кількість відкритих та закритих запитів на зміну коду.

Всі перелічені фактори можуть вказувати на активність розробки криптовалюти.

Тренди в пошукових системах – показник, що вказує на те, як часто криптовалюта стає в центрі уваги. Підвищена увага може свідчити про майбутні рухи вартості.

Аналіз настроїв повідомлень публічних тематичних товариств в Telegram – індикатор допомагає зрозуміти настрої трейдерів у відношенні до певної криптовалюти. У випадку з біткойнами позитивні настрої в Telegram кілька разів передували руху ціни, як це було в грудні 2019, квітні та червні 2020 року. У той же час відсоток повідомлень, які сприймаються як негативні, має тенденцію зростати під час зменшення вартості. А також, загальна кількість повідомлень

свідчить про рівень активності в тематичних групових чатах. Це не обов'язково відображається на руху ринку криптовалюти, але варто відзначити, що її коливання є показником загального залучення спільноти до обговорювань, що може свідчити про якісь події у близькому майбутньому.

Аналіз настроїв повідомлень публічних тематичних товариств в Twitter – фактор, що є описує настрої учасників крипто ринку. Ці показники можуть бути індикаторами коливань вартості, як це було у випадку з Ethereum в травні та серпні 2021. Іншими словами, виникає більше позитивних настроїв серед учасників, коли ціни зростають, і негативних, коли ціни падають.

Окрім факторів, що мають відношення безпосередньо до криптовалюти, є ті, які не є прямими метриками середовища криптовалюти, але все-таки можуть мати вплив. Насамперед – це глобальні економічні фактори, які демонструють глобальну економічну ситуацію в світі, яка може вплинути на попит на криптовалюту.

S&P 1200 – фактор, що є метрикою глобальної економічної ситуації у світі на основі фінансових індексів 1200 найбільших компаній світу. Цей фактор був включений як індекс купівельної здатності інвестора. Чим краща світова економічна ситуація, тим більше інвесторів можуть витрачати кошти на покупку таких нестабільних інвестицій, як криптовалюта.

Dow Jones Global – фактор, що окреслює економічний стан промислових компаній. Цей фактор можна використовувати як S&P 1200.

4.3 Валідація моделей машинного навчання

Коректність роботи регресійних моделей може визначатись за допомогою вимірювання похибки визначеного моделлю значення від тестового значення. Механіка роботи регресійних алгоритмів відрізняється від алгоритмів класифікації, де можуть бути "False positive" та "False negative" значення.

Ці метрики вибудовують певні математичні очікування, побудовані, загалом, на знаходженні середніх показників помилки серед отриманих результатів.

Наступні метрики використовуються для визначення коректності моделей.

Таблиця 5 – Метрики визначення коректності роботи регресійних моделей

Назва	Формула
Mean Absolute Error	$MAE = \frac{1}{N} \cdot \sum_{t=1}^{N} \lvert Y(t) - \hat{Y}(t) \rvert$
Mean square error	$MSE = \frac{1}{N} \cdot \sum_{t=1}^{N} \lvert (Y(t) - \hat{Y}(t))^2 \rvert$
Root mean square error	$RMSE = \sqrt{MSE}$
Explained Variance	$VAR = 1 - \frac{VAR(y - \widehat{y})}{VAR(y)}$

Для того, аби мати змогу використати наведені вище формули, необхідно правильним чином підготувати дані для тренування та тестування. Якщо дані, що не є часовими рядами можна поділити у певній пропорції та перемішати для створення тренувального та тестувального набору даних, то дані часових рядів таким чином обробляти не можна. Для коректного тренування та валідації необхідно використовувати підхід крос валідації, коли весь набір даних поділяється на ітерації та використовується для тренування та тестування.

4.4 Моделі машинного навчання для регресійного аналізу вартості

Існує безліч алгоритмів машинного навчання для розв’язання проблем регресії. Перш за все, вибір типу регресійної моделі повинен ґрунтуватися на вимогах та специфікації даних, на основі яких модель буде навчатися та перевірятися. Так як вартість криптовалюти змінюється кожного дня, то вихідні дані мають характер є часових рядів. Тип регресійної моделі є нелінійним, так як:

– логістична регресія має лише два значення (1, 0) для залежної змінної. Досліджувана проблема відноситься до визначення вартості, що змінюється з часом, тому даний тип алгоритмів не може бути використаний.

– зв'язки між змінними даних, які були описані вище, не є лінійними, оскільки збільшення однієї змінної в кілька разів в купі зі зменшенням іншої, може вплинути на залежну змінну непередбачуваним чином. Це означає, що

використання алгоритмів лінійної регресії може призвести до неправильного визначення зв’язків у даних.

– поліноміальна регресія моделює нелінійний набір даних за допомогою лінійної моделі. Ця модель, подібно до множинної лінійної регресії, має кілька незалежних змінних, проте використовує нелінійну криву.

Найбільш підходящим типом моделі регресійного аналізу є алгоритми нелінійної регресії. Для проведення експерименту визначення залежності факторів між собою будуть використовуватись наступні типи нелінійних алгоритмів:

– Decision Tree;
– Random Forest;
– Gradient Boosted trees.

Головною ідеєю моделі Дерева рішень є розділення набору даних на менші фрагменти.

Алгоритми Decision Tree мають гіперпараметри для налаштування моделі. Один з них – глибина дерева. Необхідно мати оптимальну величину глибини дерева для створення правильної моделі. Величина глибини дерева підбирається експерементальним шляхом. Якщо модель має гарні показники на тренувальних даних, а на тестувальних – ні, то спостерігається перенавчання моделі.

Алгоритм Random forest також є широко використовуваним алгоритмом для не лінійної регресії в машинному навчанні. Decision Tree алгоритм має структурні відмінності від Random Forest. Наприклад, для того аби опрацювати масив даних, Random Forest використовує низку Decision Tree, на відміну від самого Decision Tree, де є лише одне дерево. Random Forest алгоритм створює усереднену відповідь всіх Decision tree, що знаходяться всередині структури. Random Forest алгоритм є потужним через те , що велика кількість відносно некорельованих моделей (дерев), що працюють як комітет, перевершить будь-яку з окремих складових моделей (на даний момент – дерев).

Низька кореляція між моделями є ключовим пунктом успішності. Подібно до того, як інвестиції з низькою кореляцією (наприклад, акції та облігації) об’єднуються, щоб утворити портфель, який перевищує суму його частин,

некорельовані моделі можуть створювати сукупні прогнози, які є більш точними, ніж будь-які окремі прогнози. Причина цього чудового ефекту полягає в тому, що дерева захищають одне одного від їхніх індивідуальних помилок (поки вони не помиляються постійно в одному напрямку). Хоча деякі дерева можуть бути неправильними, багато інших будуть правими, тому як група дерева можуть рухатися в правильному напрямку. Рисунок 7 ілюструє роботу Random Forest алгоритму.

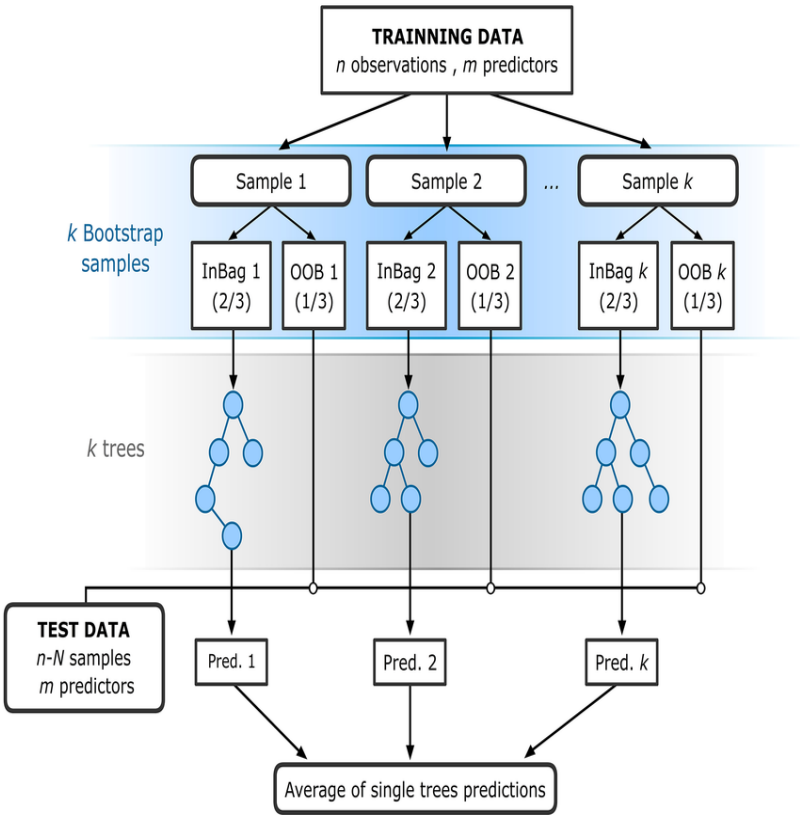

Рисунок 7 – Принцип роботи random forest алгоритму [29]

Також, необхідно розглянути Gradient Boosted Regression Trees алгоритм, що є одним із найефективніших алгоритмів для створення прогнозів. Цей алгоритм найчастіше використовують на практиці в машинному навчанні. GBRT модель працює за алгоритмом, який поєднує всі висновки та результати послідовності базових моделей. Базовою моделлю є Decision Tree. На відміну від Random Forest моделі, яка обчислює кожне дерево окремо, GBRT модель використовує вивід одного дерева, як вхідні дані для іншого. Таким чином виникає, так званий, градієнтний бустинг.

5 ПРОВЕДЕННЯ ЕКСПЕРИМЕНТУ З РЕГРЕСІЙНИМИ МОДЕЛЯМИ

5.1 Загальні відомості

Все дослідження протягом кваліфікаційної роботи можна розділити на два експерименти, де перший – це дослідження того, який з алгоритмів машинного навчання найкраще підходить для вирішення задачі нелінійної регресії та того переліку даних, що були зазначені вище. Друге дослідження – це експеримент пошуку найбільш підходящого алгоритму для прогнозування факторів, що використовуються в регресійній моделі, як вхідні дані, та застосування в цілому моделі та алгоритмів передбачення факторів.

5.2 Планування експерименту дослідження алгоритмів нелінійної регресії

Будь-яке дослідження вимагає процессинг вхідних даних. Для даного експерименту вхідні дані були завантажені з різних джерел у форматі CSV та сформовані у директорії ієрархічної структури для більш зручного доступу до них. Кожен із вихідних файлів має часову точку, коли певна подія відбувалась, таким чином є можливість у формуванні єдиного набору даних з вихідних файлів. Для виконання цього завдання використовується інструмент Apache Spark.

Apache Spark – це інструмент для паралельної обробки великої кількості даних. Цей інструмент надає високорівневе API для наступних мов програмування: Java, Python, Scala, R. Apache Spark складається з великої кількості доповнень та пакетів: Spark SQL, MLlib, GraphX та інші.

Spark.mllib надає функціонал для використання Decision tree для логістичної та багатокласової класифікації. Окрім класифікації, доступна також можливість у використанні алгоритму і для регресійного аналізу.

Random Forest алгоритм використовує в собі набір Decision Tree. Також цей алгоритм вважається одним із найефективніших для регресії та класифікації, бо поєднує в собі простоту та об’єктивність завдяки набору відповідей дерев. Як і Decision Tree алгоритм Random forest також підтримує обробку категоріальних даних. Spark.mllib використовує існуючу імплементацію Decision Trees для реалізації Random Forest.

Gradient Boosted Trees – це алгоритм, що в поєднує в собі Decision Tree також, проте навідміну від Random Forest вихідні дані в моделі отримуються шляхом передачі вихідних даних на вхід до інших дерев. Таким чином підвищується точність алгоритму.

Після того як набір даних створений та поділений на тренувальний та тестувальний, необхідно створити моделі регресійного аналізу за допомогою вказаних алгоритмів та за допомогою метрик визначити котрий із алгоритмів є найбільш вдалим для описаних даних та проблеми.

5.3 Тренування моделей нелінійного регресійного аналізу

Дані, що були обрані для тренування регресійної моделі мають період у один рік, так як таких період надає найбільш актуальну поведінку криптовалюти на ринку у її життєвому циклі.

Вхідні дані мають новий запис кожного дня, тобто загалом, набір даних включає в себе 365 прикладів того, яким чином певні індикатори впливають на вартість Ethereum. Невеликий набір даних не може бути вичерпним для створення точної моделі, проте підтвердження чи спростування цієї тези нададуть результати експерименту.

Із апаратного забезпечення та програмного забезпечення використовується 2,6 GHz Quad-Core Intel Core i7, 16 GB RAM 2133 MHz LPDDR3 та MacOS Monterey.

Для того аби зробити вимірювання, протягом тренування моделей, для порівняння, необхідно використовувати інструменти бенчмаркінгу. Одним із таких інструментів є JMH Benchmark. JMH Benchmark – це бібліотека для вимірювання швидкодії у JVM, вона була створена як частина OpenJDK проекту. Ця бібліотека надає можливість у створенні бенчмарків, на які оптимізація віртуальної машини не має впливу. Загалом, існує декілька режимів роботи бібліотеки (Таблиця 6).

Таблиця 6 – Режими роботи JMH Benchmark

Назва	Опис
Throughput	Вимірювання кількості операцій виконаних за одиницю часу.
Average Time	Вимірюється середній час для виконання методу для бенчмарку.
Sample Time	Вимірюється загальний час виконання методу, для якого будується бенчмарк. Цей режим роботи включає в себе також мінімальний та максимальний час виконання.
Single Shot Time	Вимірюється час для виконання методу, для якого будується бенчмарк. Використовується під час «холодного запуску»

В рамках експерименту необхідно визначити час, що витрачається на навчання моделей визначених вище алгоритмів. В якості одиниці часу було обрано мілісекунди. Результати бенчмарку наведені у таблиці 7.

Таблиця 7 – Середній час тренування моделей

Час, мілісекунди	Алгоритм
5570	Decision Tree
1400	Random Forest
5960	Gradient Boosted Decision Trees

Експеримент вказує на те, що найшвидший алгоритм для вказаної кількості даних є Random Forest.

5.4 Результати валідації створених моделей регресійного аналізу

Наступні таблиці та діаграми зображують результати, що були отримані протягом експерименту.

Результати валідації Decision Tree моделі показані у наступній таблиці (табл. 8).

Таблиця 8 – Результати валідації моделі Decision Tree

Метрика	Значення
Root Mean Squared Error (RMSE)	285.94868249
Mean absolute error (MAE)	233.59134551
Mean squared error (MSE)	81766.649019
Explained Variance	77061.375871

Наступна таблиця (табл. 9) містить значення, що були створені за допомогою натренованої моделі Decision Tree алгоритму.

Таблиця 9 – Результати прогнозу вартості криптовалюти

Прогноз вартості	Дійсна вартість
4537.324	4346.08
4216.365234	4342.58
4730.384277	4283.6
4340.763672	4059.81
3970.181885	3848.18
3970.181885	3883.93
4030.908936	3960.15
4294.453612	3869.35
4269.73291	4076.1

Результати валідації для створеної моделі для Random Forest алгоритму наведені у наступній таблиці.

Таблиця 10 – Результати валідації моделі Random Forest

Метрика	Значення
Root Mean Squared Error (RMSE)	238.085738
Mean absolute error (MAE)	158.3889740
Mean square error (MSE)	56684.8187

Кінець таблиці 10

Метрика	Значення
Explained Variance	31666.02249

Наступна таблиця (табл. 11) містить результати прогнозування вартості Ethereum за допомогою створеної моделі алгоритмом Random Forest.

Таблиця 11 – Результати прогнозування за допомогою Random Forest

Прогноз вартості	Дійсна вартість
4274.41063283	4346.08
4176.54161135	4342.58
4242.195490373	4283.6
4186.625616152	4059.81
4049.135305026	3848.18
3940.053935855	3883.93
3898.710454863	3960.15
3635.556222742	3869.35
4306.760494009	4283.6
4207.974993752	4082.56
4142.932871340	4057.3
4154.965331289	4079.46

Наступна таблиця (табл. 12) містить результати валідації моделі Gradient Boosting Trees.

Таблиця 12 – Результати прогнозування за допомогою моделі Gradient Boosted Trees

Метрика	Значення
Root Mean Squared Error (RMSE)	261.4591287
Mean absolute error (MAE)	223.36118231

Кінець таблиці 12

Mean square error (MSE)	68360.875985
Explained Variance	41659.384746

Наступна таблиця (табл. 13) відображає результати прогнозування вартості за допомогою створеної моделі Gradient Boosted Trees алгоритму. Спрогнозована вартість Ethereum за допомогою цієї моделі є найбільш точною.

Таблиця 13 – Результати прогнозування за допомогою моделі Gradient Boosted Trees

Прогноз вартості	Дійсна вартість
4374.544273549	4346.08
4374.58427354	4342.58
4374.74427354	4283.6
4281.1311509146	4059.81
4037.4451253475	3848.18
4037.609813992	3883.93
4037.622749360	3960.15
4038.030281828	3869.35
4280.553806327	4076.1
4280.3500029089	4082.56
4280.70529148001	4057.3
4651.65643435048	4079.46

5.5 Представлення результатів валідації моделей регресійного аналізу

Для того аби створити загальну порівняльну статистику отриманих результатів від створених моделей нелінійної регресії, необхідно створити порівняльні графіки за обраними метриками, які були використані у попередніх

розділах, а саме: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Square Error (MSE), Explained Variance.

Також, створені графіки будуть використовуватись для створення загального висновку, щодо того який саме алгоритм найкраще використовувати для досліджуваної проблеми та наведеної специфіки вхідних даних. Кожна із діаграм відображає залежність величини отриманої від метрики та часу, що був витрачений для тренування моделі.

Рисунок 8 репрезентує залежність часу від RMSE метрики, що в свою чергу означає розбіжність прогнозованого та справжнього значення в квадраті і корінь квадратний звідти.

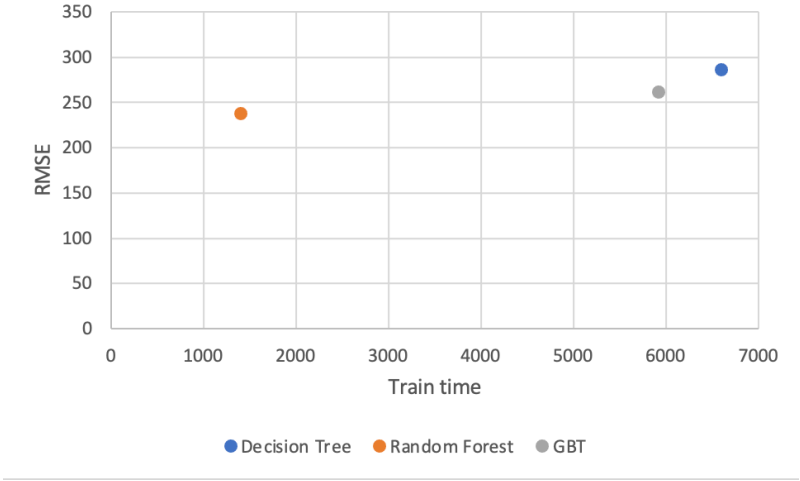

Рисунок 8 – Результати порівняння алгоритмів за RMSE

Найменша величина помилки спостерігається у Random Forest алгоритму, проте окрім цього час навчання моделі є також найменшим. Такого роду результат отриманий завдяки тому, що Random Forest алгоритм створює усереднену відповідь всіх Decision tree, що знаходяться всередині структури. Random Forest алгоритм є потужним через те , що велика кількість відносно некорельованих моделей (дерев), що працюють як комітет, перевершить будь-яку з окремих складових моделей (на даний момент – дерев).

Наступний графік (рис.9) є відображенням залежності між MAE та часом для досліджуваних алгоритмів. Показник MAE є метрикою величини різниці між

спрогнозованим та справжнім значенням величини. MAE приймає середнє значення абсолютних помилок для групи прогнозів і спостережень як вимірювання величини помилок для всієї групи. MAE також можна назвати функцією втрат L1.

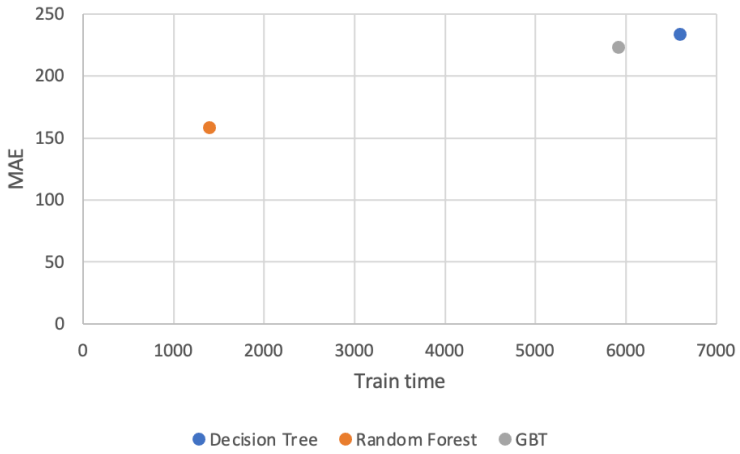

Рисунок 9 – Порівняння отриманих результатів за метрикою MAE

Результати на представленому графіку є такими ж самими, як і на попередньому, де Random Forest алгоритм має найбільшу точність, Gradient Boosted Tree алгоритм на другому місті, та Decision Tree на третьому місці.

Наступний графік (рис. 10) відображує залежність величини метрики MSE від часу навчання. Random forest є переможцем і тут.

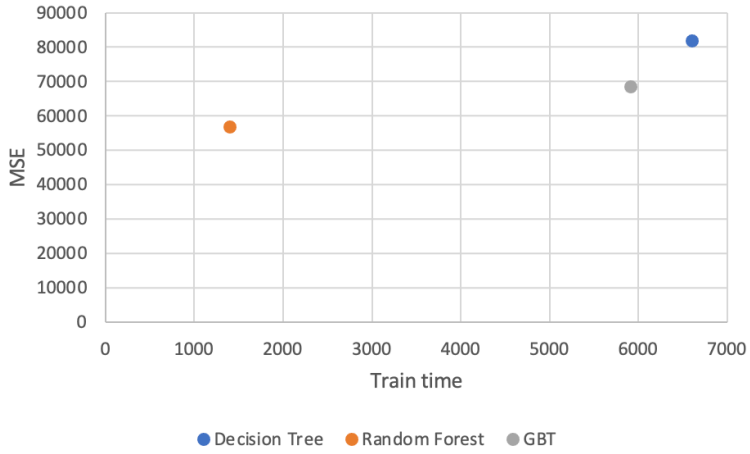

Рисунок 10 – Порівняння отриманих результатів за метрикою MSE

Наступний графік (рис. 11) зображує результати за Explained Variance, який використовується для вимірювання невідповідності між досліджуваними та

актуальними даними. Іншими словами, це частина від загальної кількості варіантів, що представлена факторами, або не представлена через похибку.

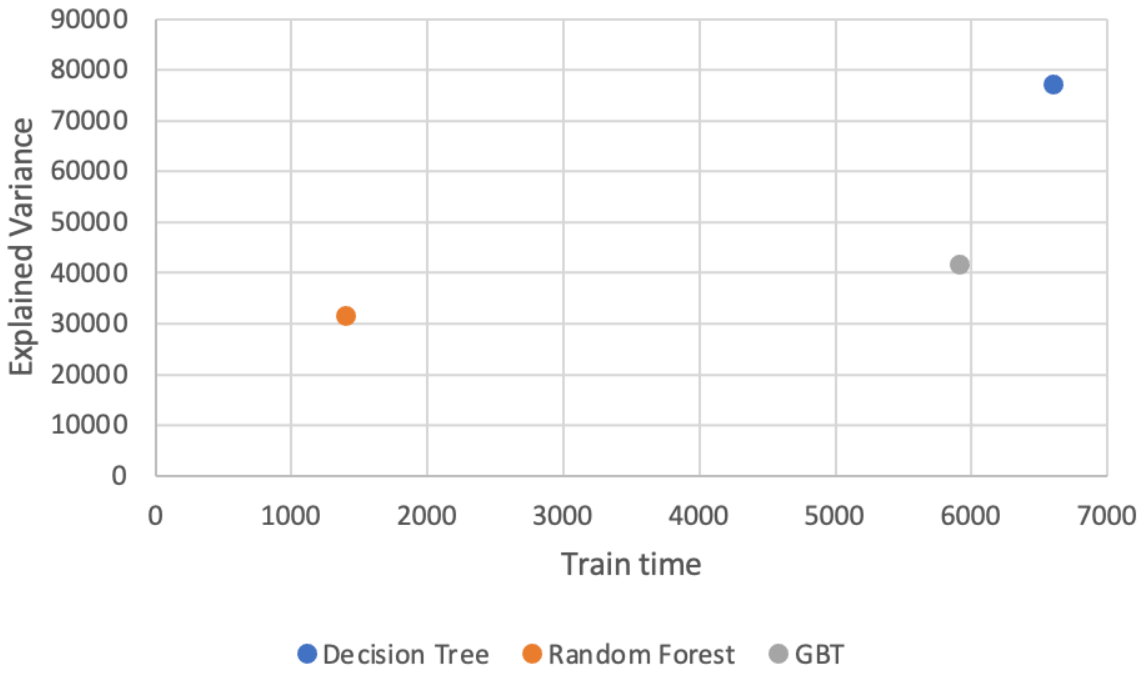

Рисунок 11 – Порівняння результатів алгоритмів за допомогою Explained Variance метрики

Чим більша величина метрики, тим більша сила асоціації. Це факт також означає, що є можливість у створенні кращого передбачення, якщо більше значення для цієї метрики прослідковується.

За результатами практичного експерименту виявлення найбільш відповідного алгоритма для створення моделі нелінійного регресійного аналізу були виявлені наступні факти:

– Decision Tree алгоритм не є оптимальним для використання у регресії для представлених даних та даної проблеми.

– Random Forest алгоритм є найбільш точним у прогнозуванні, окрім цього час тренування є найменшим.

Для того аби зробити моделі алгоритмів Decision Tree та Gradient Boosted Tree більш точними, можна застосувати наступні рекомендації:

а) модель алгоритму Decision tree може бути надлишково тренованою, що

характеризується незадовільною здатністю до обробки нових даних, коли в той же час працездатність моделі на тренувальних даних навпаки вказує на досить гарні результати. Такі явища можуть спостерігатись якщо не має точно вказаної граничної кількості Leaf для дерева. В такому випадку метрика MSE на тренувальних даних покаже значення, що дорівнює 0, так як для кожного окремого спостереження або запису буде створений свій Leaf. Для того аби не допустити перетренування моделі, необхідно визначити обмеження щодо розміру дерева та застосувати Tree pruning.

б) для налаштування моделі Gradient Boosting Tree алгоритму можна використовувати параметри, що зосереджені на налаштуваннях дерева в цілому та параметрах бустингу. Сам по собі алгоритм Gradient Boosting Tree є стійким до перенавчання зі збільшенням кількості дерев в моделі, проте застосування великої кількості дерев для певного learning rate може спровокувати перенавчання моделі. Якщо зменшити learning rate та збільшити кількість дерев, то час для навчання моделі збільшиться. Ці принципи необхідно пам’ятати під час створення оптимізацій для моделі. В якості оптимізації моделі можна застосувати такі рекомендації:

1) необхідно обирати високий learning rate. Зазвичай, 0.1 працює але краще обирати значення між 0.05 до 0.2, так як саме такі значення працюють для різного роду задач.

2) необхідно обрати оптимальну кількість trees для обраного вище learning rate. Зазвичай це значення знаходиться між 40-70. Це значення повинне бути також оптимальним для вашої системи за швидкодією.

5.6 Висновки щодо проведеного експерименту нелінійної регресії факторів

Даний експеримент був проведений з метою визначення факторів, що мають вплив на вартість криптовалюти. Виконані роботи за цим спрямуванням надали уяву про те, що саме вже виконано, та окреслили ту область, що ще не була достатньо досліджена. Також, було виявлено, що виконані роботи є не вичерпними та було запропонована власна думка та підхід до розв’язання задачі. Даний підхід

базується на тому, що існують певні закономірності та зв’язки між екосистемою криптовалюти та загальноекономічними світовими метриками.

Наступний крок у експерименті був етапом визначення типу алгоритму для регресійного аналізу. Як результат, був обране сімейство нелінійних алгоритмів, а саме Trees. Також цей етап включає визначення факторів, що повинні бути досліджені.

Виконання експерименту було наступним кроком, під час якого було обрано Apache Spark, як інструмент для створення набору даних із вихідних даних та створення моделей регрсійного аналізу. Для визначення кількості часу, що був витрачений для навчання моделей, використовувався інструмент JMH.

За результами валідації моделей було визначено, що Random Forest алгоритм є найбільш точним, проте окрім цього час навчання є найменшим серед інших алгоритмів. Gradient Boosted Tree алгоритм знаходиться на другому місці. Decision Tree є алгоритмом з найбільшим часом навчання, окрім цього точність також залишає бажати кращого.

Наступні дослідження можуть бути зосереджені на виявленні додаткових факторів, що можуть використовуватись в моделі та насамперед у створенні прогнозів вартості за допомогою вже досліджених алгоритмів та моделей або інших.

Окрім цього, можливим розвитком ідеї є обчислення та внесення до розрахунків факторів, що відображать стан інфраструктурних частин екосистеми криптовалюти: перебої в електроенергії в місцях накопичення найбільших майнінг пулів, або перебоїв мережі Інтернет чи політичні зміни в країнах.

6 ПРОВЕДЕННЯ ЕКСПЕРИМЕНТУ ПРОГНОЗУВАННЯ ВАРТОСТІ КРИПТОВАЛЮТИ

6.1 Визначення оптимального алгоритму для прогнозування факторів.

Для визначення вартості криптовалюти, використовуючи регресійну модель з визначеними експерементальним шляхом факторами, необхідне створення прогнозу кожного з факторів для подальшого виконання прогнозу. Для цієї мети можна використати створений компанією Facebook інструмент Prophet для передбачення часових рядів, які мають не лінійний тренд, можуть мати сезонність (річну, тижневу та (або) всередині дня).

Окрім цього, інструмент дозволяє додавати до моделі ефект різних подій, що мають вплив на дані. Серед цих подій можуть бути як національні свята, які вже вбудовані в інструмент, так і, наприклад, створення хард-форку криптовалюти, що сприяє збільшенню покупок криптовалюти. Ці події можуть мати як різний мультиплікатор ефекту, так і різний проміжок часу протягом якого цей ефект діє. Наприклад подія оновлення протоколу Ethereum відома задовго до його введення. Приклад налаштувань подій на рисунку 12.

```r
# R
library(dplyr)
playoffs <- data_frame(
  holiday = 'playoff',
  ds = as.Date(c('2008-01-13', '2009-01-03', '2010-01-16',
                 '2010-01-24', '2010-02-07', '2011-01-08',
                 '2013-01-12', '2014-01-12', '2014-01-19',
                 '2014-02-02', '2015-01-11', '2016-01-17',
                 '2016-01-24', '2016-02-07')),
  lower_window = 0,
  upper_window = 1
)
superbowls <- data_frame(
  holiday = 'superbowl',
  ds = as.Date(c('2010-02-07', '2014-02-02', '2016-02-07')),
  lower_window = 0,
  upper_window = 1
)
holidays <- bind_rows(playoffs, superbowls)
```

Рисунок 12 – Приклад налаштувань подій моделі [30]

Таким чином, є можливість у створенні додаткових умов визначення переломних моментів тренду.

Задля покращення точності моделі, є можливість у визначенні «Changepoints». Ця характеристика допомагає інструменту вказати де саме відбувається зміна тренду. На рисунку 13 зображено автоматичне розпізнавання зміни тренду.

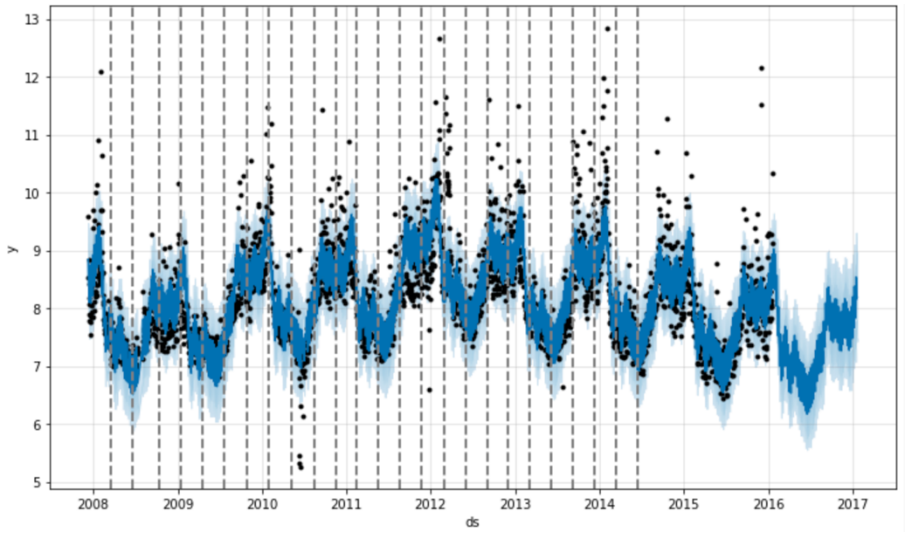

Рисунок 13 – Автоматичне розпізнавання changepoints даних [31]

Проте автоматичний режим не завжди може бути точним, і задля покращення точності моделі, необхідно визначити в ручному режимі changepoints (рисунок 14).

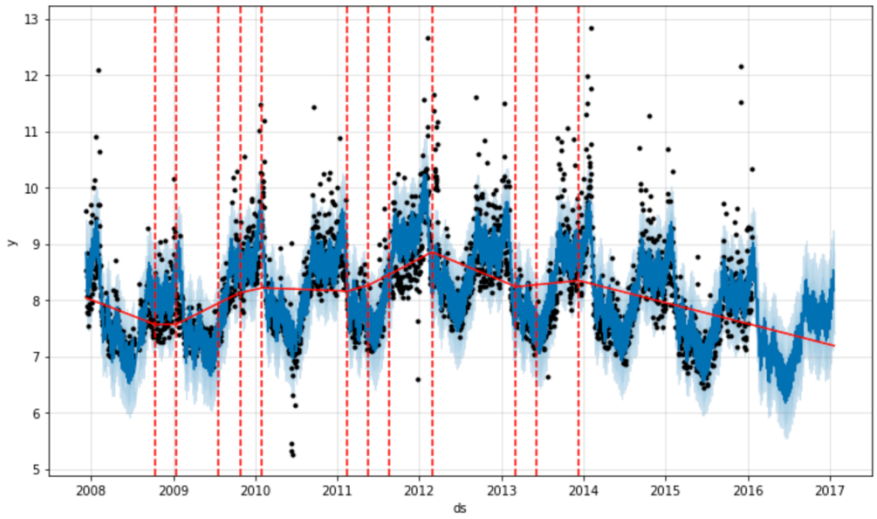

Рисунок 14 – Визначення changepoints в ручному режимі [31]

Таким чином, користувач визначив де, на його думку, після аналізу створених в автоматичному режимі changepoints, необхідно їх задати, щоб навчити модель правильному тренду.

Prophet також гарно справляється і з пропущеними даними. Проте у цьому дослідженні буде використовуватись послідовувач Prophet – NeuralProphet, який побудований на AR-Net [32] та PyTorch [33].

Порівняння точності обох моделей на різного розміру наборах даних надав інформацію який алгоритм краще використовувати для досліджуваної проблеми (рисунок 15).

n_training_days	prophet_RMSE	neural_RMSE	prophet_MAPE	neural_MAPE
730	234378.761025	70271.728217	0.295893	0.068367
910	50849.885531	47513.701719	0.046069	0.043358
1090	55759.466652	53768.762780	0.055645	0.053843
1270	51161.887917	66381.028717	0.057436	0.077146
1450	63721.356010	113188.359929	0.071921	0.140630
1630	61874.272860	64740.967981	0.057310	0.065376
1810	55379.708338	59813.661440	0.046136	0.061819

Рисунок 15 – Порівняльна статистика Prophet та NeuralProphet [34]

За результатами дослідження було виявлено, що NeuralProphet краще використовувати на малих об’ємах даних, а Prophet на великих. Для досліджуваного випадку, краще підходить NeuralProphet.

6.2 Прогнозування факторів

Для виконання прогнозування факторів необхідно підготувати дані для використання їх у алгоритмі NeuralProphet, який приймає їх у форматі [ds, y], де ds – часова мітка, y – величина, яку необхідно спрогнозувати.

Метою виконання прогнозування є створення набору даних для регресійної моделі, що складається з:

– час;
– очікувана вартість криптовалюти;
– спрогнозовані величини факторів на зазначений час.

Окрім цього, кожна згенерована величина кожного фактору є значенням one-step ahead прогнозу, що означає створення прогнозу лише на одну часову точку вперед. Кожна наступна точка буде створюватись за тим же алгоритмом.

Для створення спрогнозованого набору даних для валідації регресійної моделі було обрано відрізок часу 01.12.2021 – 31.12.2021, так як дані факторів за цей місяць вже відомі, то є можливість у створенні прогнозованого набору даних за наведений місяць.

6.3 Визначення алгоритму регресійного аналізу для прогнозування вартості

Виконане раніше дослідження надало вектор розвитку. Для прогнозування вартості криптовалюти необхідно використовувати нелінійні алгоритми регресії. Для того аби визначити найбільш підходящий алгоритм для його використання в купі зі спрогнозованими даними факторів, необхідно виконати порівняльний аналіз помилок моделей, використовуючи наступні метрики: MAE, RMSE, MSE. Для експерименту було відібрано наступні алгоритми:

- decision tree;
- random forest;
- gradient boosted trees;
- xgboost.

Навчання кожної з моделей відбувалось на даних за попередній рік від періоду тестування: 1.01.2021 - 30.11.2021. Для кожної із моделей були підібрані параметри, що надають найменшу помилку. Decision Tree: depth – 30. Random Forest: depth – 5. Gradient Boosted Trees: depth – 5, subsample rate – 0.4, iteration – 10. XGBoost: eta – 0.3, depth – 6, rounds number – 7, workers number – 2.

Наступна таблиця (табл. 14) показує результати експерименту та виявлені величини помилок створених моделей.

Таблиця 14 – Порівняльна таблиця помилок регресії на спрогнозованих даних

Назва	Decision Tree	Random Forest	GBT	XGBoost
RMSE	560.40725	355.62862	623.47651	295.41653
MSE	314056.28882	126471.71769	388722.9612	87270.93094
MAE	507.24214	313.06250	582.93858	198.03023

Наведена вище таблиця вказує на те, що модель на базі XGBoost алгоритму є найбільш продуктивною у порівнянні з іншими моделями.

Також, необхідно з’ясувати наскільки точність моделей зменшилась у порівнянні з реальними даними для тестування (див. рис. 15).

Таблиця 15 – Порівняльна таблиця результатів валідації моделей на спрогнозованих даних та реальних даних за період 1.12.2021 – 31.12.2021

	Алгоритм	RMSE	MSE	MAE
Дійсні дані	DecisionTree	285.94868249	81766.649019	233.59134551
	Random Forest	238.085738	56684.8187	158.3889740
	GBT	261.4591287	68360.875985	223.36118231
Спрогнозовані дані	Decision Tree	560.40725	314056.28882	507.24214
	Random Forest	355.62862	126471.71769	313.06250
	GBT	623.47651	388722.9612	582.93858
	XGBoost	295.41653	87270.93094	198.03023

Порівняльна таблиця вказує на те, що використання спрогнозованих на один день вперед факторів у регресійній моделі, що була натренована на дійсних даних спричиняє більшу помилку, ніж помилка при використанні дійсних даних.

Проте при порівнянні двух моделей, що мають найкращі результати серед тестованих: Random Forest та XGBoost, можна виявити, що розбіжність у відхиленні не є значною: 158 у Random Forest та 198 у XGBoost.

ВИСНОВКИ

Ця робота є підґрунтям до створення більш досконалих систем для прогнозування часових рядів, якими є вартість криптовалюти. Також дана робота зосередила увагу на природі утворення вартості криптовалюти та на чинниках, що впливають на неї. Кожен елемент світу не може існувати у вакуумі.

Криптовалюта – перш за все програмне забезпечення, яке використовується нестандартним шляхом. Та послуги, що криптовалюта виконує мають змінну вартість – в залежності від попиту, а попит створюється також завдяки певним регуляторам: політика, закони, розповсюдженість. Проте окрім цього сервіс цього програмного забезпечення має змогу приймати участь у торгах, що залучає ще одну сторону до утворення ціни – трейдерів. Цей підхід монетизації є інакшим на відміну від стандартних підписок або купівлі ліцензії програмного забезпечення.

На цю роботу надихнула велика кількість статей в Інтернеті, що розповідають як легко передбачити вартість криптовалюти, надаючи користувачам хибні матеріали щодо цього. Багато авторитетних фінансистів стверджують, що передбачити вартість будь-якої валюти, розглядаючи лише її вартість у відношенні до часу, є абсурдною задачею, так як лише вартість, як дані, не відображають дійсних подій, що відбуваються лаштунками. Вартість – це набір факторів, які регулюються попитом.

За результатами дослідження можна сказати, що існує залежність між факторами, які були обрані для регресійних моделей, проте це не вичерпна їх кількість. Є необхідність у додатковому додаванні або створенні ансамблів моделей, що аналізують технічні показники та різні інфраструктурні фактори: політичні регулятори, проблеми електромережі або мережі Інтернет.

Також проведений експеримент вказує на те, що створення прогнозу на один день вперед є більш-менш успішним з огляду на помилку, що створюют моделі регресії. На мою думку, довготривалого прогнозу важко досягти наведеними мною методами. Для цього необхідні інші інструменти.

ПЕРЕЛІК ПОСИЛАНЬ

[1] U. Mukhopadhyay, A. Skjellum, O. Hambolu, J. Oakley, L. Yu and R. Brooks, "A brief survey of Cryptocurrency systems," 2016 14th Annual Conference on Privacy, Security and Trust (PST), 2016, pp. 745-752, doi: 10.1109/PST.2016.7906988.

[2] K. Smelyakov, A. Chupryna, D. Sandrkin and M. Kolisnyk, "Search by Image Engine for Big Data Warehouse," 2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), 2020, pp. 1-4, doi: 10.1109/eStream50540.2020.9108782.

[3] K. Smelyakov, D. Karachevtsev, D. Kulemza, Y. Samoilenko, O. Patlan and A. Chupryna, "Effectiveness of Preprocessing Algorithms for Natural Language Processing Applications," 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T), 2020, pp. 187-191, doi: 10.1109/PICST51311.2020.9467919.

[4] F. Béres, I. A. Seres, A. A. Benczúr and M. Quintyne-Collins, "Blockchain is Watching You: Profiling and Deanonymizing Ethereum Users," 2021 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPS), 2021, pp. 69-78, doi: 10.1109/DAPPS52256.2021.00013.

[5] Yu Chen, Xuecheng Ma, Cong Tang and Man Ho Au, "Pgc: Pretty good decentralized confidential payment system with auditability", Cryptology ePrint Archive Report 2019/319, 2019, [online] Available: https://eprint.iacr.org/2019/319.

[6] Understanding The Different Types of Cryptocurrency. URL: https://www.sofi.com/learn/content/understanding-the-different-types-of-cryptocurrency.

[7] The 10 Most Popular Cryptocurrencies, and What You Should Know About Each Before You Invest. URL: https://time.com/nextadvisor/investing/cryptocurrency/types-of-cryptocurrency.

[8] P. Tasatanattakool and C. Techapanupreeda, "Blockchain: Challenges and applications," 2018 International Conference on Information Networking (ICOIN), 2018, pp. 473-475, doi: 10.1109/ICOIN.2018.8343163.

[9] A Guide to Cryptocurrency Fundamental Analysis. URL: https://academy.binance.com/en/articles/a-guide-to-cryptocurrency-fundamental-analysis

[10] The 7 Key Factors Influencing Cryptocurrency Value. URL: https://www.makeuseof.com/factors-influencing-the-cryptocurrency-value/

[11] S. Boshuis, T. Braam, A. Pedroza Marchena and S. Jansen, "The Effect of Generic Strategies on Software Ecosystem Health: The Case of Cryptocurrency Ecosystems," 2018 IEEE/ACM 1st International Workshop on Software Health (SoHeal), 2018, pp. 10-17.

[12] Jiangtao Ma, Yaqiong Qiao, Guangwu Hu, Yongzhong Huang, Arun Kumar Sangaiah, Chaoqin Zhang, et al., "De-anonymizing social networks with random forest classifier", IEEE Access, vol. 6, pp. 10139-10150, 2017.

[13] Vynokurova O., Peleshko D., Zhernova P., Perova I., Kovalenko A. (2021) Solving Fraud Detection Tasks Based on Wavelet-Neuro Autoencoder. In: Babichev S., Lytvynenko V., Wójcik W., Vyshemyrskaya S. (eds) Lecture Notes in Computational Intelligence and Decision Making. ISDMCI 2020. Advances in Intelligent Systems and Computing, vol 1246. Springer, Cham. https://doi.org/10.1007/978-3-030-54215-3_34

[14] T. Radivilova, L. Kirichenko, D. Ageiev and V. Bulakh, "Classification Methods of Machine Learning to Detect DDoS Attacks," 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2019, pp. 207-210, doi: 10.1109/IDAACS.2019.8924406.

[15] F. A. Cahyadi, A. I. Owen, F. Ricardo and A. A. S. Gunawan, "Blockchain Technology behind Cryptocurrency and Bitcoin for Commercial Transactions," 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI), 2021, pp. 115-119, doi: 10.1109/ICCSAI53272.2021.9609790.

[16] How Bitcoin Works. URL: https://www.investopedia.com/news/how-bitcoin-works.

[17] S. Pillai, D. Biyani, R. Motghare and D. Karia, "Price Prediction and Notification System for cryptocurrency Share Market Trading," 2021 International Conference on Communication information and Computing Technology (ICCICT), 2021, pp. 1-7, doi: 10.1109/ICCICT50803.2021.9510122.

[18] X. Li and C. A. Wang, "The technology and economic determinants of cryptocurrency exchange rates: The case of bitcoin", Decision Support Systems, vol. 95, pp. 49-60, 2017.

[19] Caia.org, 2022. [Online]. Available: https://caia.org/sites/default/files/metcalfeslaw_websiteupload_7-5-18.pdf. [Accessed: 21- Jan- 2022].

[20] Y. Hilpisch, Python for finance. Beijing: O'Reilly Media, 2020.

[21] Mdpi.com, 2022. [Online]. Available: https://www.mdpi.com/2673-2688/2/4/30/pdf. [Accessed: 21- Jan- 2022].

[22] "Random Walk Theory", Investopedia, 2022. [Online]. Available: https://www.investopedia.com/terms/r/randomwalktheory.asp. [Accessed: 06- Feb- 2022].
[23] L. Cocco, R. Tonelli and M. Marchesi, "Predictions of bitcoin prices through machine learning based frameworks", PeerJ Computer Science, vol. 7, p. e413, 2021. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8022579.
[24] K. Smelyakov, A. Chupryna, M. Hvozdiev and D. Sandrkin, "Gradational Correction Models Efficiency Analysis of Low-Light Digital Image," 2019 Open Conference of Electrical, Electronic and Information Sciences (eStream), 2019, pp. 1-6, doi: 10.1109/eStream.2019.8732174.
[25] "What is bitcoin and how does it work?", New Scientist, 2022. [Online]. Available: https://www.newscientist.com/definition/bitcoin/. [Accessed: 18- Feb- 2022].
[26] "What is DeFi", Coinbase, 2022. [Online]. Available: https://www.coinbase.com/learn/crypto-basics/what-is-defi. [Accessed: 24- Feb- 2022].
[27] "'I Lost Everything': How Squid Game Token Collapsed | CoinMarketCap", *CoinMarketCap Alexandria*, 2022. [Online]. Available: https://coinmarketcap.com/alexandria/article/i-lost-everything-how-squid-game-token-collapsed. [Accessed: 18- Mar- 2022].
[28] K. Smelyakov, M. Shupyliuk, V. Martovytskyi, D. Tovchyrechko and O. Ponomarenko, "Efficiency of image convolution," 2019 IEEE 8th International Conference on Advanced Optoelectronics and Lasers (CAOL), 2019, pp. 578-583, doi: 10.1109/CAOL46282.2019.9019450.
[29] The flowchart of random forest (RF) for regression. URL: https://www.researchgate.net/figure/The-flowchart-of-random-forest-RF-for-regression-adapted-from-Rodriguez-Galiano-et_fig3_303835073.
[30] Seasonality, Holiday Effects, And Regressors", *Prophet*, 2022. [Online]. Available: https://facebook.github.io/prophet/docs/seasonality,_holiday_effects,_and_regressors.html. [Accessed: 08- Apr- 2022].
[31] Trend Changepoints", *Prophet*, 2022. [Online]. Available: https://facebook.github.io/prophet/docs/trend_changepoints.html. [Accessed: 18- Apr- 2022].
[32] O. Triebe, N. Laptev and R. Rajagopal, "AR-Net: A simple Auto-Regressive Neural Network for time-series", *arXiv.org*, 2022. [Online]. Available: https://arxiv.org/abs/1911.12436. [Accessed: 18- Apr- 2022].
[33] "PyTorch", *Pytorch.org*, 2022. [Online]. Available: https://pytorch.org/. [Accessed: 19- Apr- 2022].
[34] "Prophet vs. NeuralProphet", *Medium*, 2022. [Online]. Available: https://towardsdatascience.com/prophet-vs-neuralprophet-fc717ab7a9d8. [Accessed: 20- Apr- 2022].

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

[2] K. Smelyakov, A. Chupryna, D. Sandrkin and M. Kolisnyk, "Search by Image Engine for Big Data Warehouse," 2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), 2020, pp. 1-4, doi: 10.1109/eStream50540.2020.9108782.

[3] K. Smelyakov, D. Karachevtsev, D. Kulemza, Y. Samoilenko, O. Patlan and A. Chupryna, "Effectiveness of Preprocessing Algorithms for Natural Language Processing Applications," 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T), 2020, pp. 187-191, doi: 10.1109/PICST51311.2020.9467919.

[24] K. Smelyakov, A. Chupryna, M. Hvozdiev and D. Sandrkin, "Gradational Correction Models Efficiency Analysis of Low-Light Digital Image," 2019 Open Conference of Electrical, Electronic and Information Sciences (eStream), 2019, pp. 1-6, doi: 10.1109/eStream.2019.8732174.

[28] K. Smelyakov, M. Shupyliuk, V. Martovytskyi, D. Tovchyrechko and O. Ponomarenko, "Efficiency of image convolution," 2019 IEEE 8th International Conference on Advanced Optoelectronics and Lasers (CAOL), 2019, pp. 578-583, doi: 10.1109/CAOL46282.2019.9019450.