**Міністерство освіти і науки України**

**Національний технічний університет України «КПІ» імені Ігоря Сікорського**

**Кафедра обчислювальної техніки ФІОТ**

**ЗВІТ**

**з лабораторної роботи №2**

**з навчальної дисципліни «Вступ до технології Data Science»**

**Тема:**

**СТАТИСТИЧНЕ НАВЧАННЯ З ПОЛІНОМІАЛЬНОЮ РЕГРЕСІЄЮ**

**Виконав:**

Студент 3 курсу кафедри ІПІ ФІОТ,

Навчальної групи ІП-11

Лошак В.І.

**Перевірив:**

Професор кафедри ОТ ФІОТ

Писарчук О.О.

**Київ 2023**

**І. Мета:**

Виявити дослідити та узагальнити особливості реалізації процесів статистичного навчання із застосуванням методів обробки Big Data масивів та калмановської рекурентної фільтрації з використанням можливостей мови програмування Python.

**ІІ. Завдання:**

***Реалізація проекту триває та спрямовано на збільшення функціональності програмної компоненти***

*Лабораторія провідної ІТ-компанії реалізує масштабний проект розробки універсальної платформи з обробки Big Data масиву статистичних даних поточного спостереження для виявлення закономірностей і прогнозування розвитку контрольованого процесу. Платформа передбачає розташування back-end компоненти на власному хмарному сервері з наданням повноважень користувачам заздалегідь адаптованого front-end функціоналу універсальної платформи.*

**ІІІ. Завдання IV рівня(максимум 15 балів):**

Реалізувати групу вимог 1 та (або) 2 з імплементацією однієї з групи вимог 3. Докладно опитати отримані R&D рішення

Група вимог\_1:

1. Отримання вхідних даних із властивостями, заданими в Лр\_1;
2. Модель вхідних даних із аномальними вимірами;
3. Очищення вхідних даних від аномальних вимірів. Спосіб виявлення аномалій та очищення обрати самостійно;
4. Визначення показників якості та оптимізація моделі (вибір моделі залежно від значення показника якості). Показник якості та спосіб оптимізації обрати самостійно.
5. Статистичне навчання поліноміальної моделі за методом найменших квадратів (МНК – LSM) – поліноміальна регресія для вхідних даних, отриманих в п.1,2. Спосіб реалізації МНК обрати самостійно;
6. Прогнозування (екстраполяцію) параметрів досліджуваного процесу за «навченою» у п.5 моделлю на 0,5 інтервалу спостереження (об’єму вибірки);
7. Провести аналіз отриманих результатів та верифікацію розробленого скрипта.

Група вимог 3:

3.1. Здійснити розробку власного алгоритму виявлення аномальних вимірів та / або «навчання» параметрів відомих алгоритмів «бачити» властивості статистичної вибірки.

**IV. Результати виконання лабораторної роботи.**

1. **Отримання вхідних даних із властивостями, заданими в Лр\_1**

Для отримання даних попередньої лабораторної використано бібліотеку pandas. Загружені дані про тренд та залишки скомбіновано в один data frame. В лабораторній 1 ці дані вже було очищено від аномалій.

A screen shot of a computer

Description automatically generated

Рис. 1— Візуалізація тренду та комбінованої моделі даних.

1. **Модель вхідних даних із аномальними вимірами;**

Для того щоб додати аномалії до даних використано метод додавання кратної певному мультиплікатору кількості стандартних відхилень. Згенеровані аномалії розподілені рівноміро.

A screen shot of a computer code

Description automatically generated

A graph showing a line

Description automatically generated with medium confidence

Рис. 2— Візуалізація часового ряду з доданими аномаліями що відхиляються на 5 стандартних відхилень.

1. **Очищення вхідних даних від аномальних вимірів. Спосіб виявлення аномалій та очищення обрати самостійно;**

3.1. Здійснити розробку власного алгоритму виявлення аномальних вимірів та / або «навчання» параметрів відомих алгоритмів «бачити» властивості статистичної вибірки.

Для розробки власного алгоритму для виявлення аномалій було вирішено модифікувати алгоритм Sliding\_Window\_AV\_Detect\_medium наведений Писарчук О.О. в файлі L\_1\_3\_Statistical\_learnin. В ході модифікації було створено алгоритм для підбору оптимальних параметрів для заданого часового ряду, а також модифіковано спосіб збільшення вікна. Оновлений алгоритм medium має вигляд:

Крок 1: Створити вікно даних з заданою параметром розмірністю

Крок 2: Визначити статистичні характеристики поточного вікна

Крок 3: Зсунути вікно вздовж інтервалу на одну поділку

Крок 4: Визначити статистичні характеристики нового вікна

Крок 5: Порівняти стат. Характеристики нового вікна з даними попереднього вікна і замінити включене до вікна значення на медіанне значення по поточному вікну.

Крок 6:встановити характеристики поточного вікна замість минулих характеристик

Крок 7: Якщо не досягнуто кінця даних — повернутися на крок 3.

A computer screen shot of a program code

Description automatically generated

Рис. 3— Оновлений алгоритм детекції аномалій.

Для підбору параметрів моделі було розроблено алгоритм що проходиться через усі значення розміру вікна та трешхолду в певному діапазоні значень, та для кожної комбінації оцінює якість очистки використовуючи MSE.

A screenshot of a computer program

Description automatically generated

Рис. 4— Алгоритм підбору параметрів для очищення вибірки від аномалій способом sliding\_window\_medium.

A screenshot of a computer program

Description automatically generated

Рис. 5— Значення підібраних параметрів алгоритму sliding\_window\_medium

A screen shot of a computer

Description automatically generated

Рис. 6—Ряд очищений від аномалій з використанням оптимальних параметрів очищення.

1. **Визначення показників якості та оптимізація моделі (вибір моделі залежно від значення показника якості). Показник якості та спосіб оптимізації обрати самостійно.**

Для створення прогнозу було обрано побудувати поліноміальну модель. Для оцінки якості моделі вибрано показник R2. Для підбору оптимального значення параметру поліному використано алгоритм пошуку що використовує R2. Пошук параметра здійснено в діапазоні 1—10.

A screen shot of a computer code

Description automatically generated

Рис. 7— Метрика R2 для оптимальної поліноміальної моделі.

1. **Статистичне навчання поліноміальної моделі за методом найменших квадратів (МНК – LSM) – поліноміальна регресія для вхідних даних, отриманих в п.1,2. Спосіб реалізації МНК обрати самостійно;**

A screen shot of a computer program

Description automatically generated

Рис. 8— Алгоритм підбору параметрів поліному.

A graph with blue lines and orange lines

Description automatically generated

Рис. 9— Поліноміальна модель натренована за МНК після проведення підбору параметрів

1. **Прогнозування (екстраполяцію) параметрів досліджуваного процесу за «навченою» у п.5 моделлю на 0,5 інтервалу спостереження (об’єму вибірки);**

З вище наведених метрик зроблено висновок що модель достатньо добре описує дані одже ми можемо використати її для прогнозування.

A screenshot of a computer screen

Description automatically generated

Рис. 10— Прогноз з використанням поліноміальної моделі на 1.5 та +10 відповідно до розміру вибірки.

1. **Провести аналіз отриманих результатів та верифікацію розробленого скрипта.**

З графіків видно що в короткостроковій перспективі модель прогнозує різкий стрибок вгору ціни Bitcoin з подальшим його стрімким падінням, що приводить нас до висновку що така волатильність може бути використана для спекуляцій на ринку. В той же час різке падіння відображає факт того що поліноміальна модель містить в собі певні надоліки в плані прогнозування оскільки не враховує історичні дані про діапазон в якому ціни біткоїн зазвичай перебувають що і приводить до появи від’ємних значень ціни і безпрецедентного прогнозу падіння. Незважаючи на вище сказане, в короткостроковій перспективі модель показує себе задовільно.

**IV. Висновки.**

У ході цієї роботи, я практикував побудову поліноміальної моделі для прогнозування часових рядів. Мною було створено/модифіковано алгоритм sliding\_window\_medium очищення даних від аномальних вимірів що адаптується до наданих даних за своїми параметрами і якісно відтворює оригінальні дані.

Виконав: студент ФІОТ Лошак В.І. ІП-11