Дослідження ефективності алгоритмів машинного навчання для задачі управління активами на ринку криптовалют з урахуванням оцінки довіри користувачів до криптовалюти

Виконав:
студент 4 курсу, групи КА-93
Кротенко Нікіта Сергійович

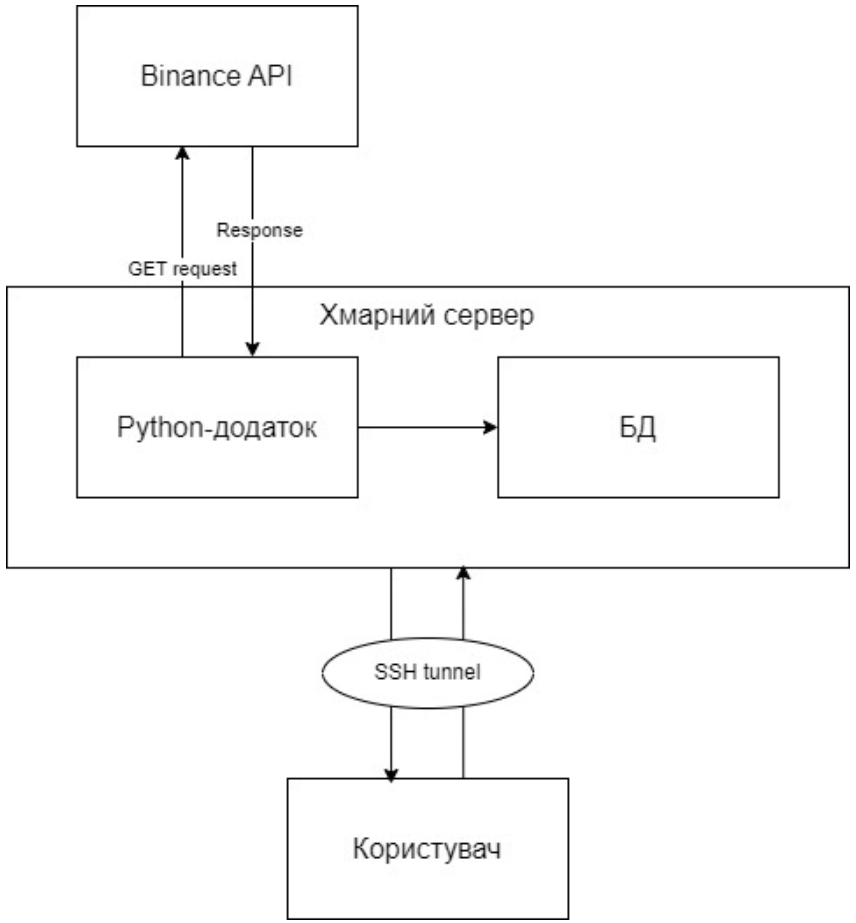
Керівник:
Ассистент кафедри ММСА
Канцедал Георгій Олегович

Об’єкт дослідження – дані про хвилинні свічки і глибину ринку (bid/ask)

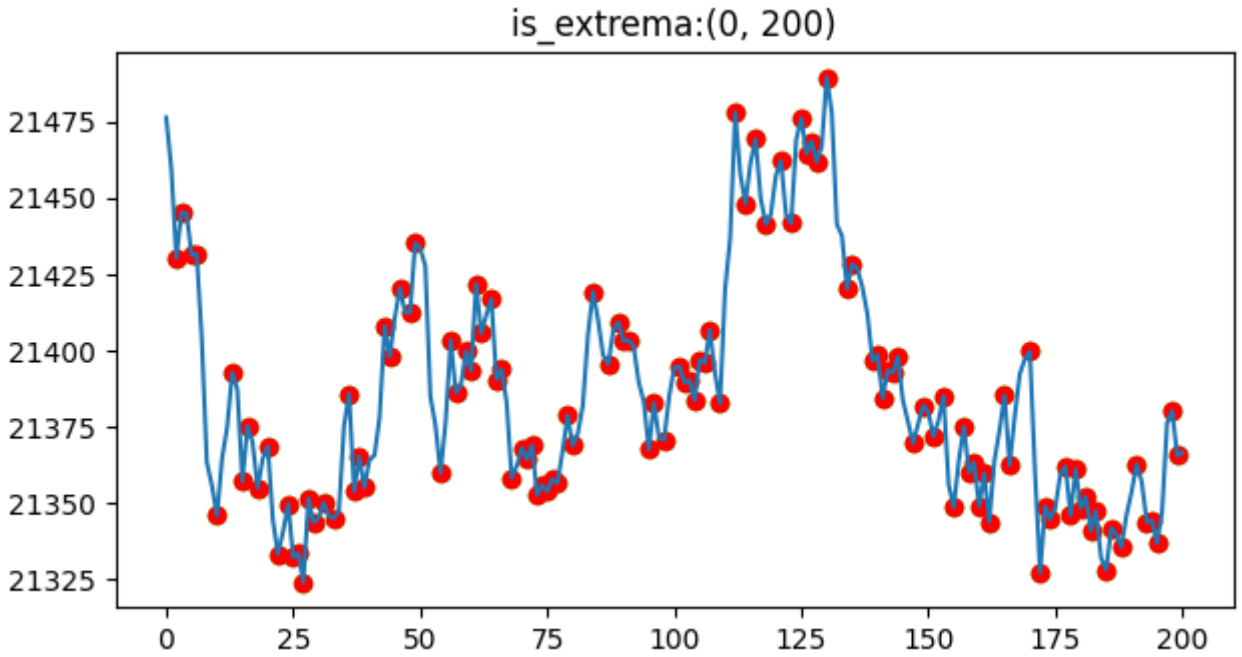
Предмет дослідження – моделі для передбачення часових рядів.

Мета роботи – відтворенні усього циклу реального дослідження, включаючи збір і підготовку даних, визначення цільових метрик та безпосередню розробку і налаштування моделей машинного навчання, для їх подальшого порівняння.

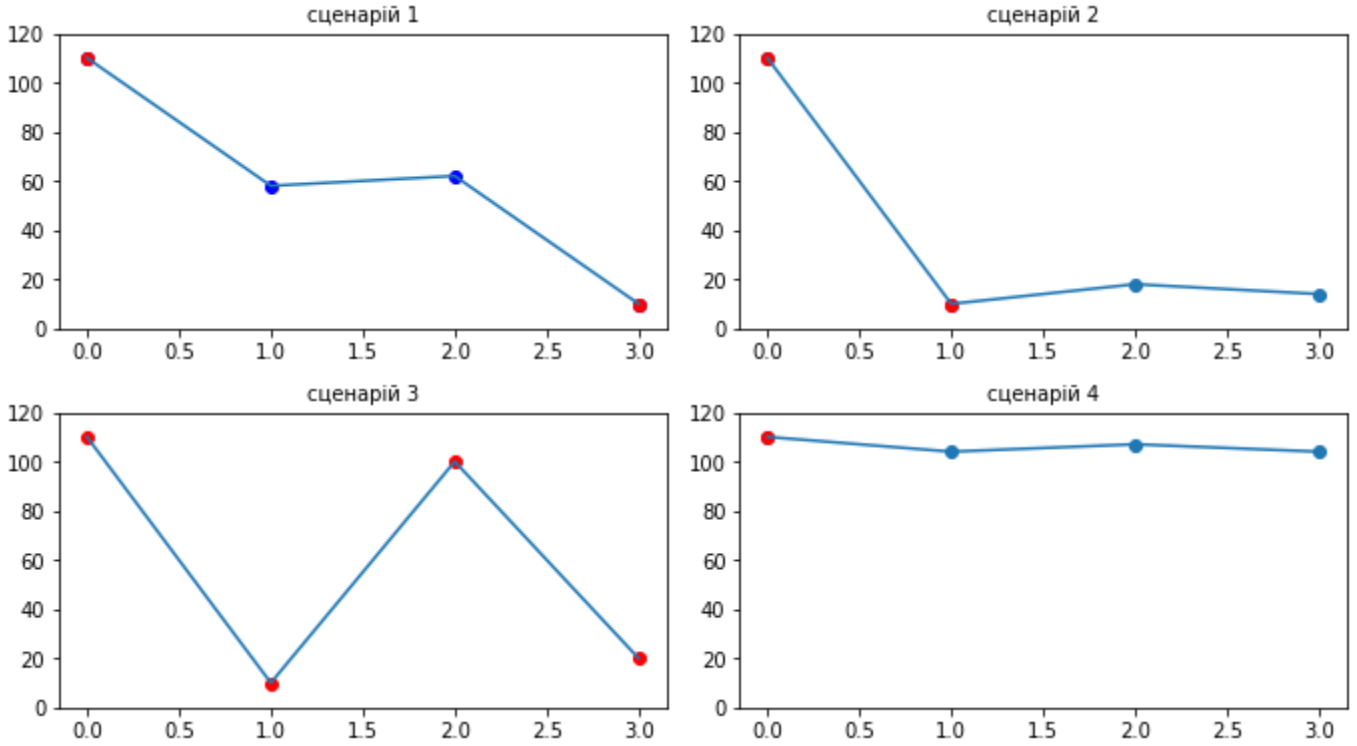
Блок-схема організації хмарного сервера для збору даних



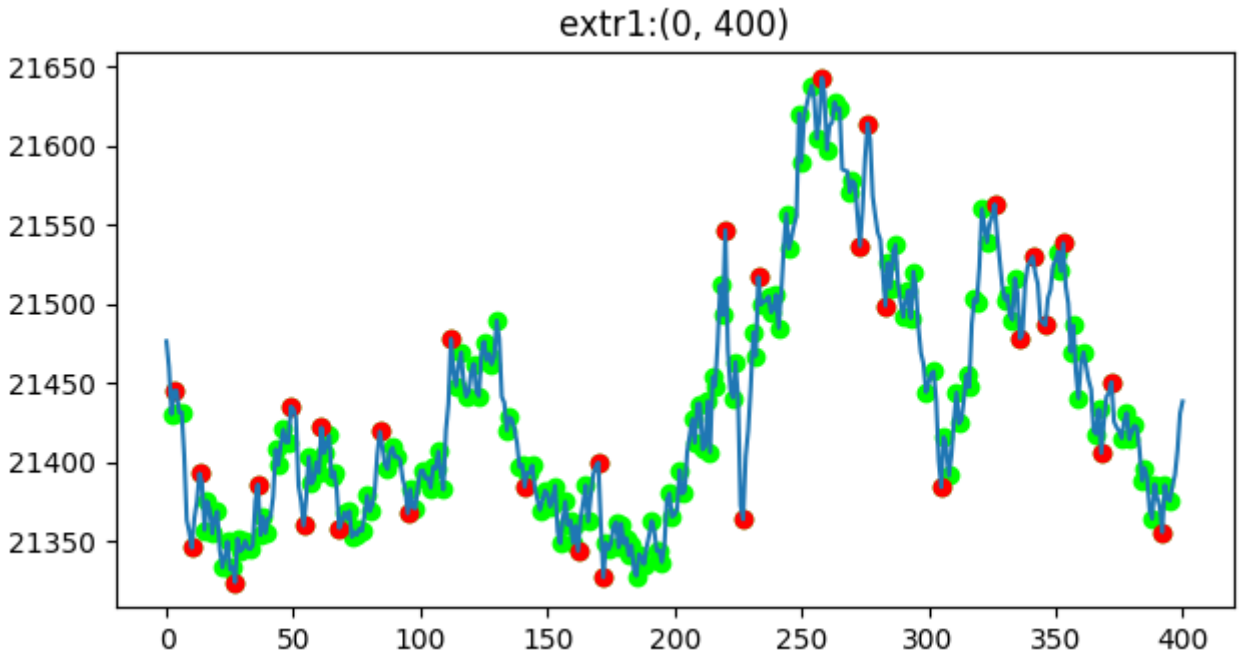
Створення розмітки даних
1. Виділення локальних екстремумів



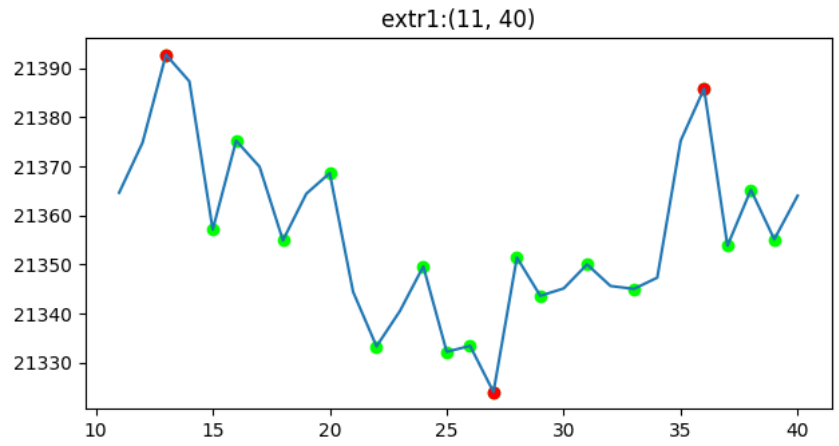
Створення розмітки даних
2. Пошук початкових екстремумів

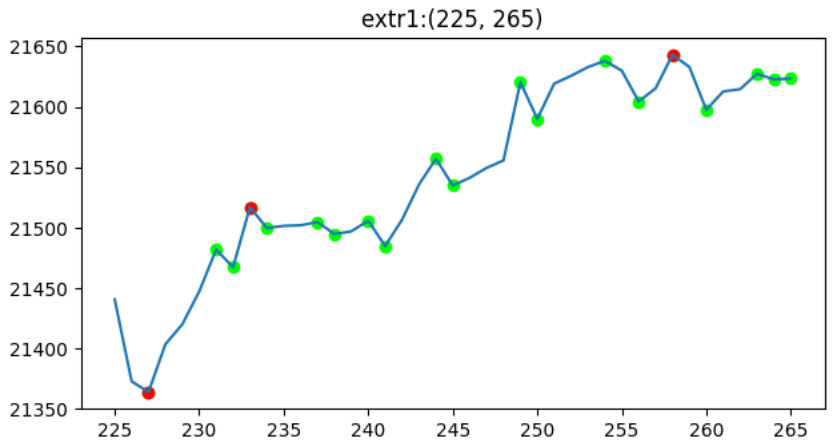


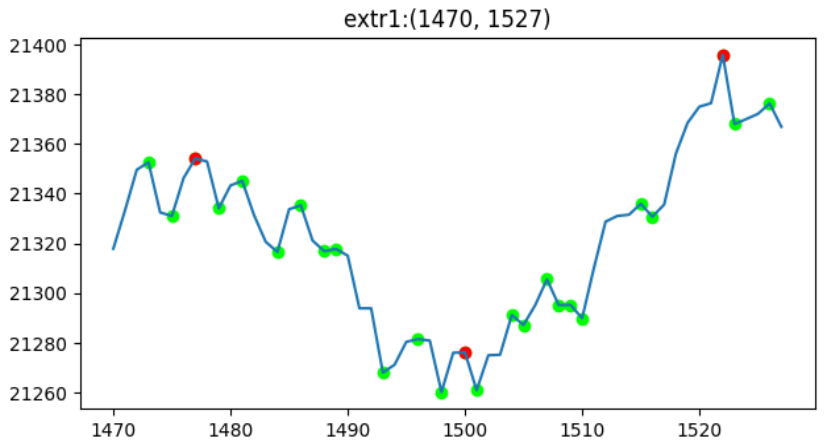
Створення розмітки даних
2. Пошук початкових екстремумів



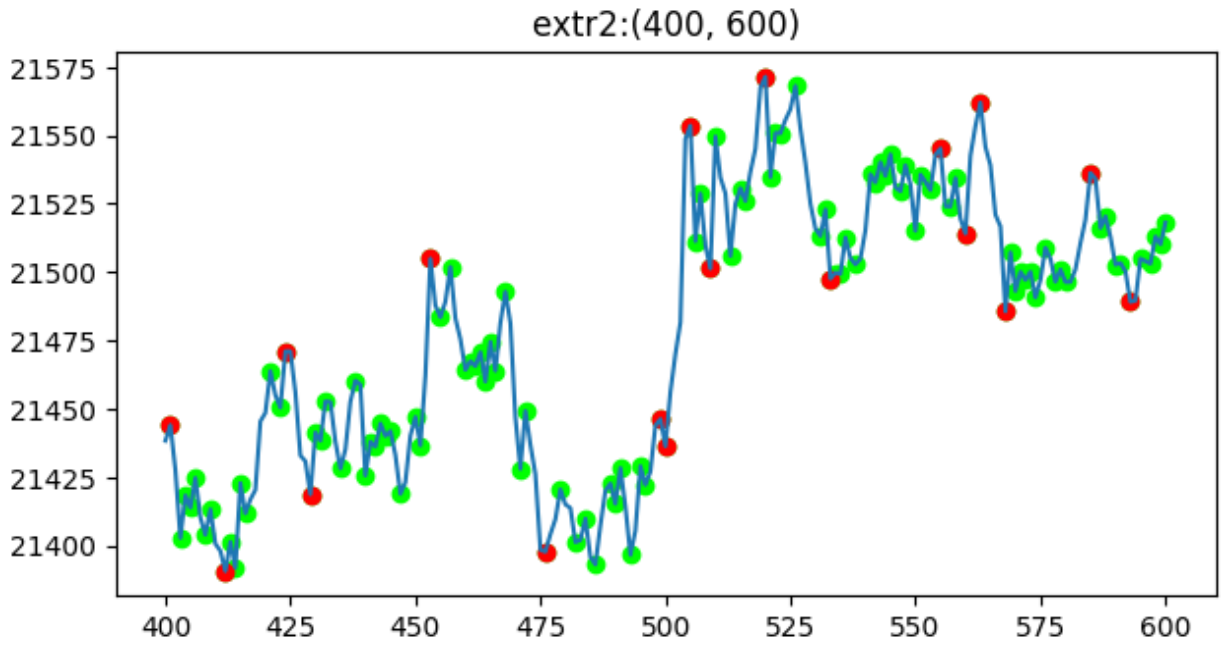
Створення розмітки даних
3. Пошук глобальних екстремумів на відрізках



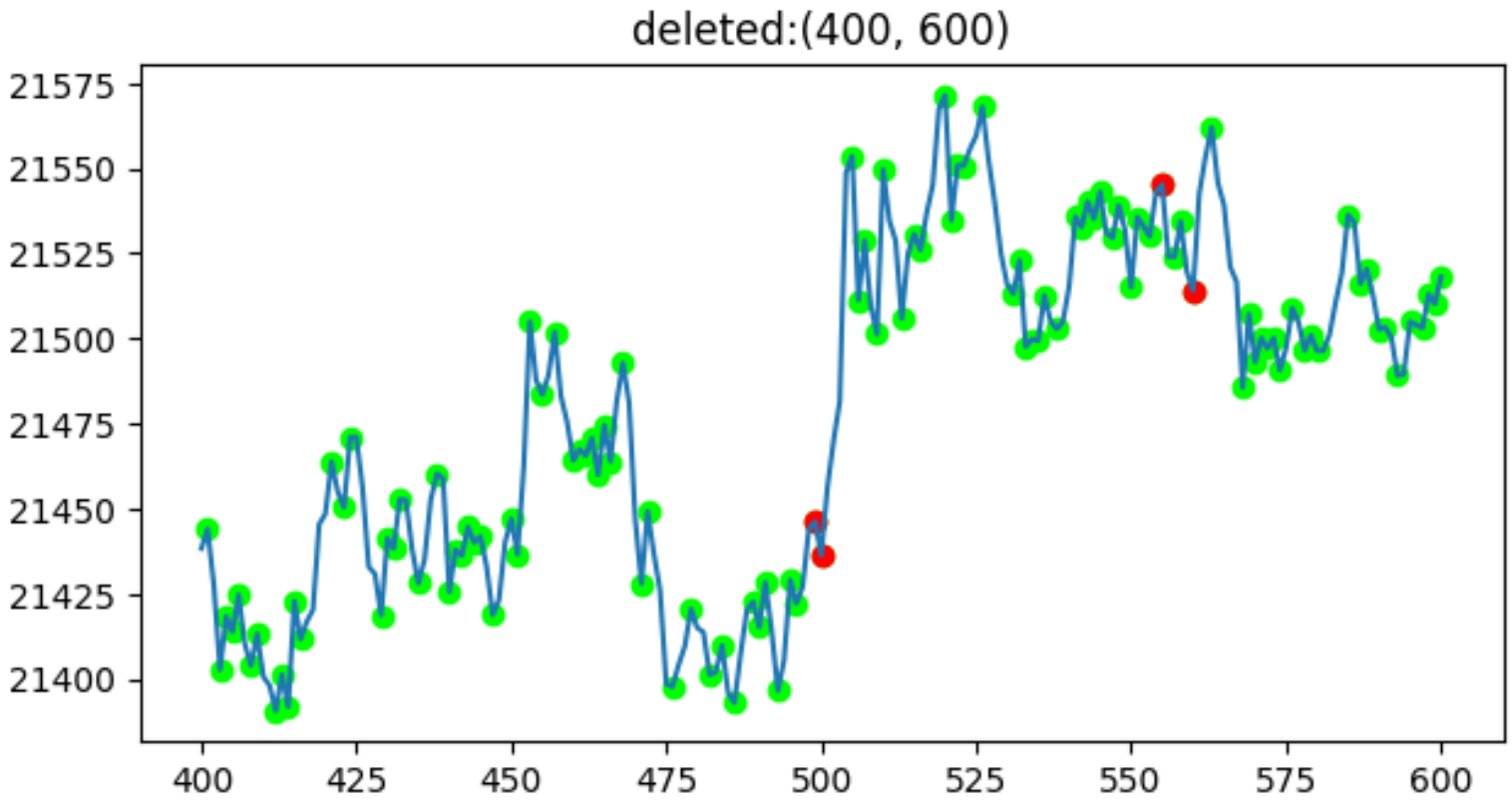




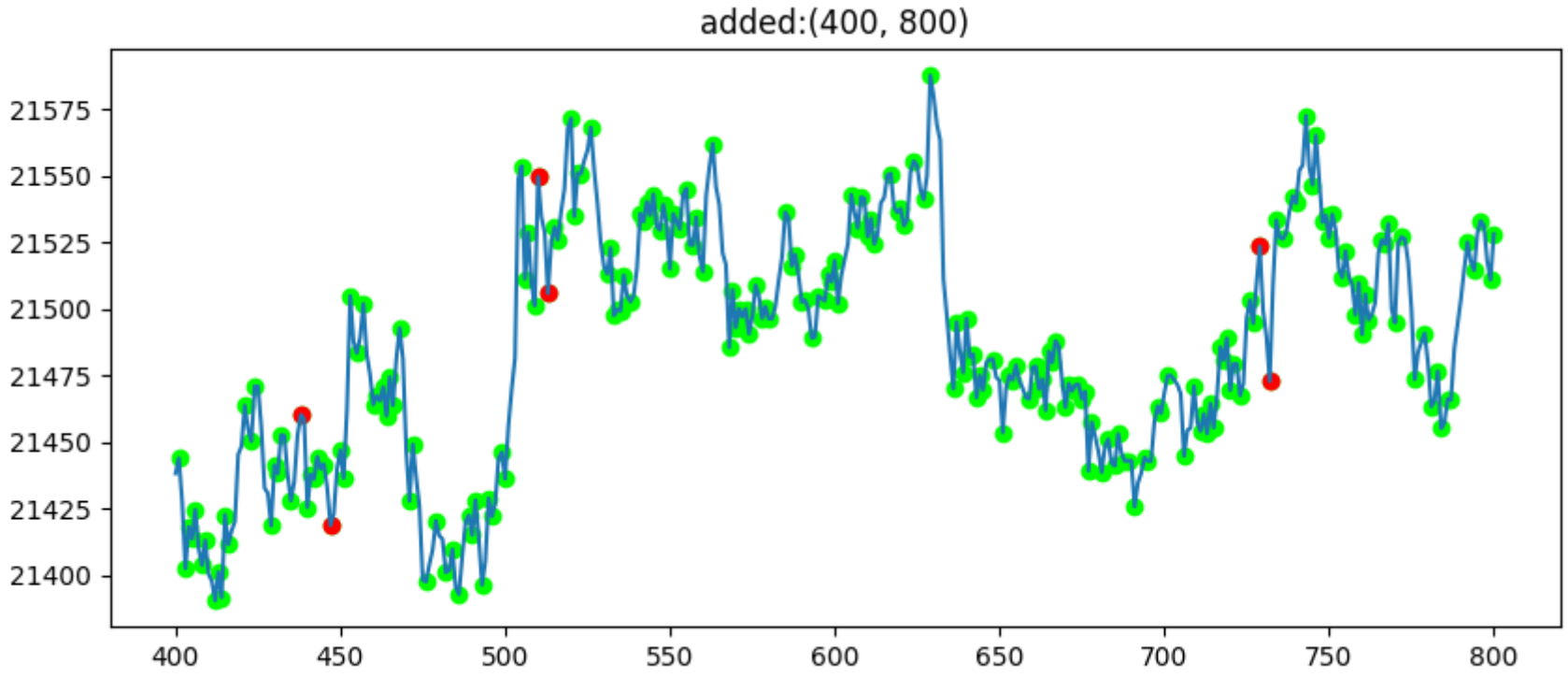
Створення розмітки даних
3. Пошук глобальних екстремумів на відрізках



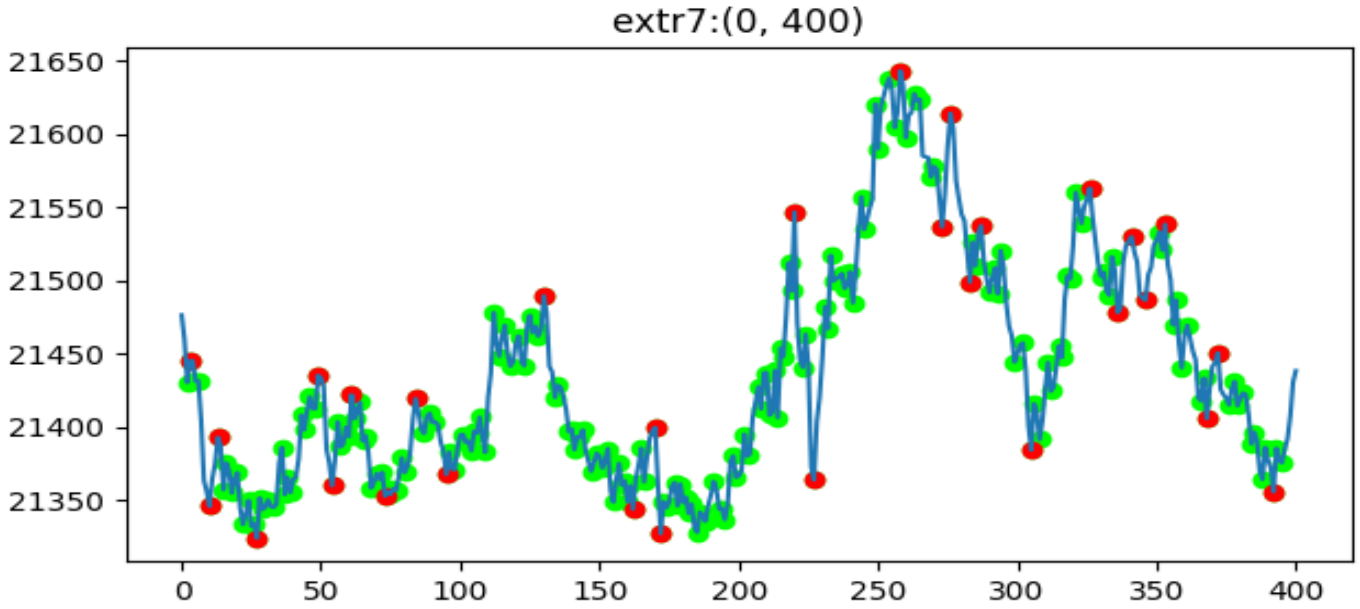
Створення розмітки даних
4. Пошук неправильно визначених екстремумів



Створення розмітки даних
5. Пошук пропущених екстремумів



Створення розмітки даних
Результати

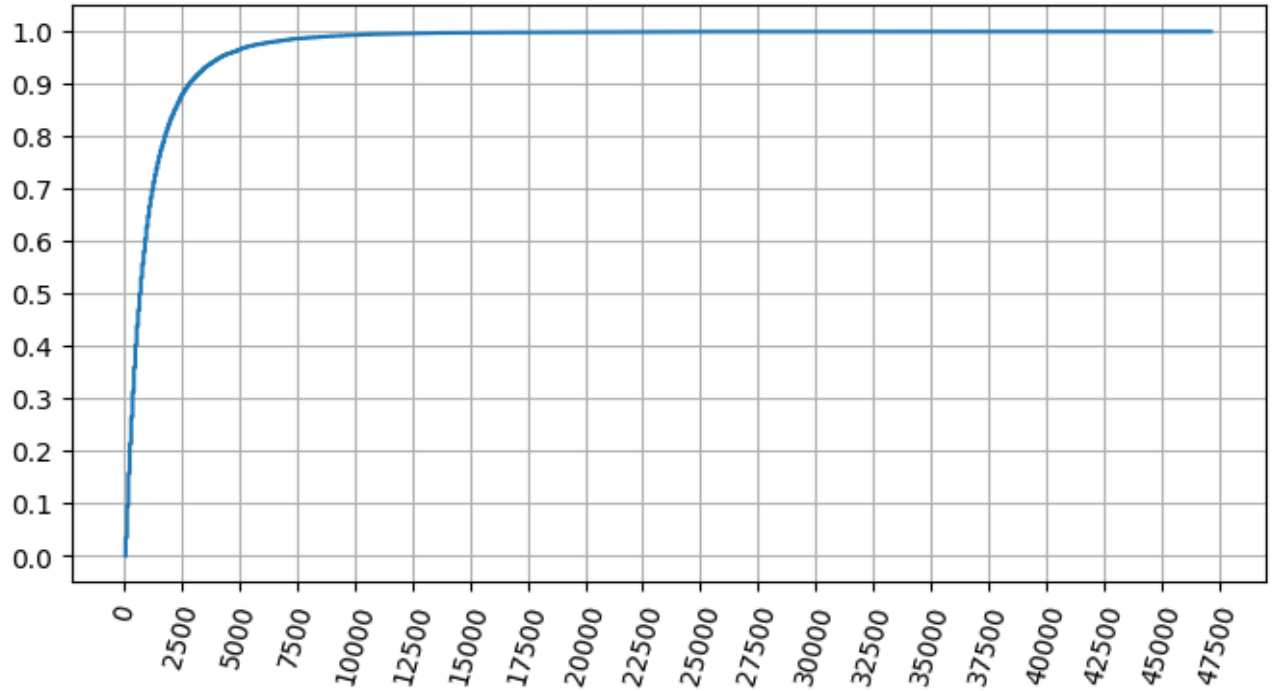


1. Екстремумів всього	188275
2. Екстремумів після виділення початкових	18620
3. Екстремумів після знайдення глобальних по відрізкам	16024
4. Екстремумів після видалення зайвих	15078
5. Екстремумів після додавання попередніх	18406
6. Екстремумів повтору кроків 4 і 5	18324

Огляд створеної розмітки

Кумулятивний розподіл часу тривалості контрактів:

- 80.77% контрактів тривають до 30 хвилин;
- 93.48% контрактів тривають до 60 хвилин.



Генерація нових ознак
1. По-хвилинні

$$high\_low\_diff_{perp} = high_{perp} - low_{perp}$$

$$open\_close\_diff_{perp} = open_{perp} - close_{perp}$$

$$bid\_ask\_spread_{perp} = \frac{askPrice_{perp} - bidPrice_{perp}}{askPrice_{perp}}$$

$$bid\_ask\_spread\_abs_{perp} = askPrice_{perp} - bidPrice_{perp}$$

Та інші

Генерація нових ознак
1. Ознаки групи «Короткострокові очікування»

$$short\_term\_expectations_1 = \frac{open_{perp} - open_{cq}}{time\_left + 86400}$$

$$short\_term\_expectations_5 = \frac{(open_{perp} - open_{cq})^2}{time\_left + 86400 * 9}$$

$$short\_term\_expectations_9 = \frac{(open_{perp} - open_{cq})^2}{\sqrt{time\_left + 86400 * 90}}$$

$$short\_term\_expectations_{10} = \frac{\sqrt{|open_{perp} - open_{cq}|}}{time\_left + 86400}$$

$$short\_term\_expectations_{14} = \frac{\sqrt{|open_{perp} - open_{cq}|}}{\sqrt{time\_left + 86400 * 9}}$$

Та інші

Генерація нових ознак
1. По-контрактні

Згрупурувавши дані по контрактам, з колонок:

['volume_perp', 'trades_perp', 'volume_cq', 'trades_cq’,
'bidQty_perp', 'askQty_perp', 'bidQty_cq', 'askQty_cq’]

беремо показники:

- sum – сума;
- mean – середнє;
- total_delta – різниця в ціні відкриття і ціні закриття контракту;
- minmax_rel_diff – відносна різниця в максимальній і мінімальній ціні що були зафіксовані на проміжку доки було відкрито контракт.

Та інші.

Вибір роду задачі: Регресія чи Класифікація

Недоліки задачі класифікації:

1. Модель не може отримати переваги від хвилинних коливань ціни і повинна закривати або відкривати контракти лише у фіксовану секунду хвилини і лише за ринковими ордерами.
2. Цільовий показник незбалансований, що ускладнює вибір метрики і потребує додаткових маніпуляцій над даними.

Переваги задачі регресії:

1. З можливістю відкривати лімітні ордери ми можемо отримати додатковий прибуток за рахунок коливання ціни контракту протягом хвилини.
2. За допомогою введення другої змінної ми можемо не відкривати контракти на інтервалах, на яких для отримання бажаного прибутку тривалість контракту перевищуватиме комфортні для нас 30 хвилин.

Враховуючи вищезазначені властивості було вирішено розглядати задачу як задачу регресії.

Перший цільовий показник

Додатково визначимо 2 колонки:

$$up\_border = \frac{high + max(open, close)}{2} \qquad low\_border = \frac{low + min(open, close)}{2}$$

Значення цільового показника 1:

- Для мінімумів - low_border
- Для максимумів - up_border

Другий цільовий показник

Додатково визначимо 2 колонки:

- min_over_upcoming_30min – мінімальний показник low_border, який буде зафіксовано протягом наступних 30 хвилин;
- max_over_upcoming_30min – максимальний показник up_border, який буде зафіксовано протягом наступних 30 хвилин.

Значення цільового показника 2:

Для максимумів - min_over_upcoming_30min

Для мінімумів - max_over_upcoming_30min

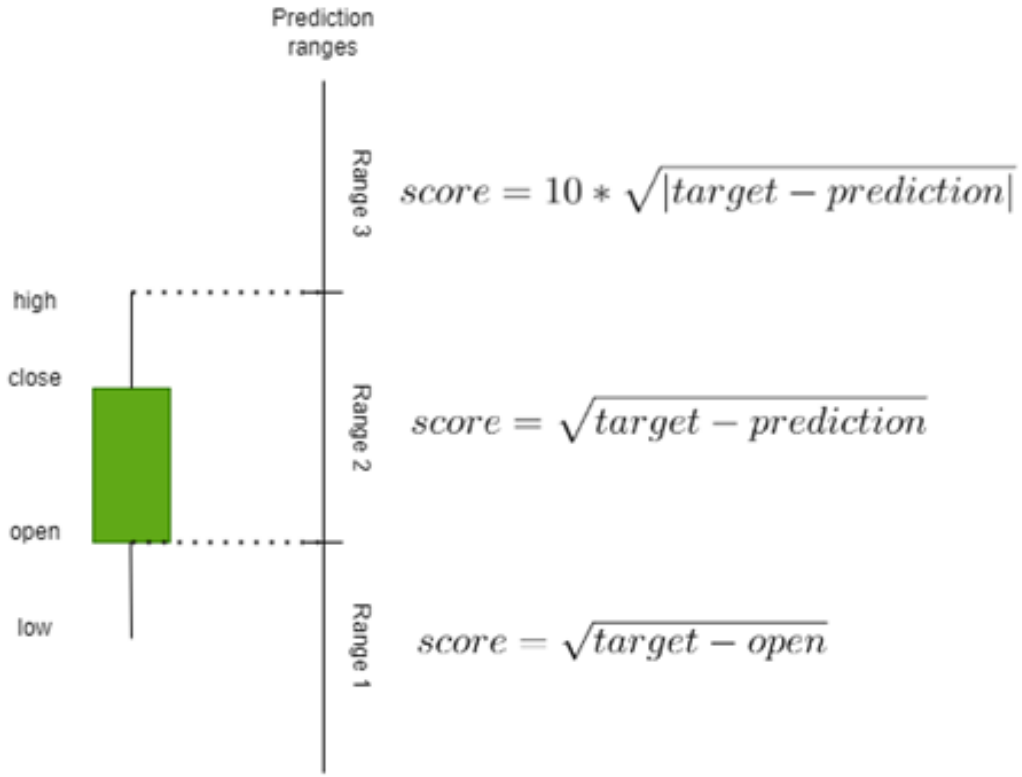
Фінальна обробка цільових показників

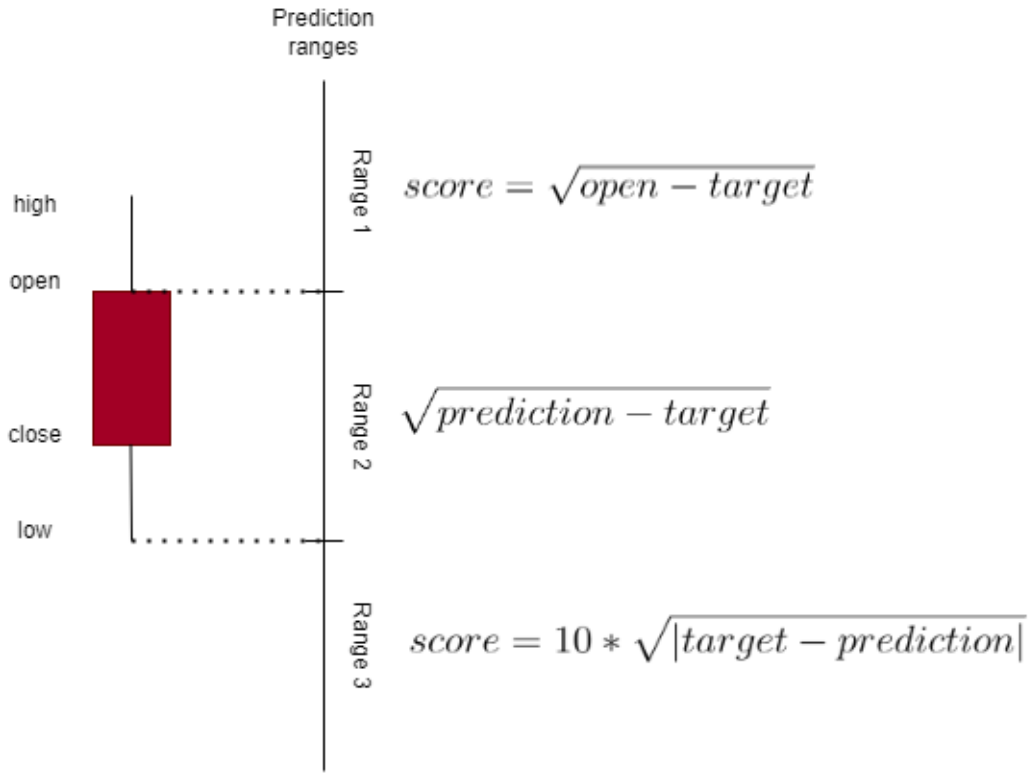
Тепер, маючи цільові показники для всіх рядків, що є екстремумами ми пересуваємо усі значення в цих колонках на клітинку назад і заповнюємо пропуски методом “backfill” – кожен пропуск заповнюється значенням наступної непорожньої клітинки.

Таким чином, після закриття кожної хвилинної свічки відкриватиметься новий лімітний ордер по ціні передбаченій в цільовому показнику 1.

У випадках, коли прибуток від контракту відкритого по ЦП1 і закритого по ЦП2 не перевищує 1.075% – актуальний контракт буде закрито по ціні ЦП1, але нового контракту відкрито не буде.

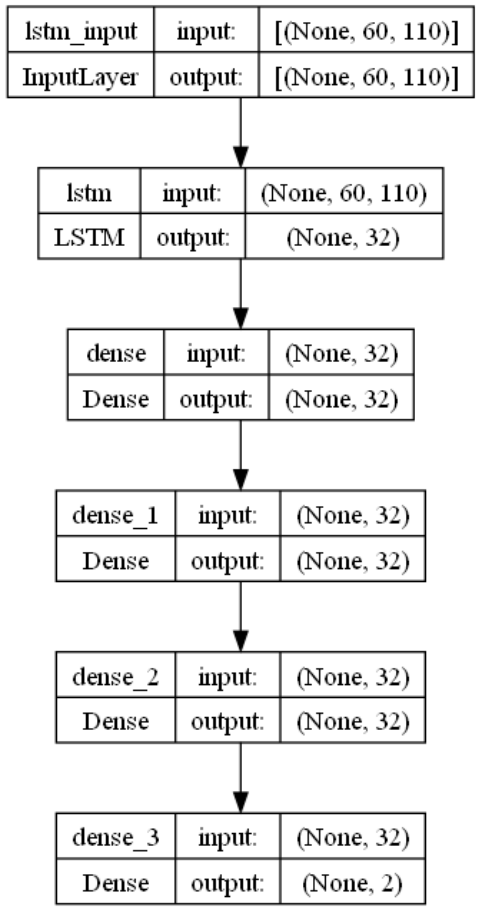
Метрика


Prediction ranges
Range 3
Range 2
Range 1
high
close
open
low
$score = 10 * \sqrt{|target - prediction|}$
$score = \sqrt{target - prediction}$
$score = \sqrt{target - open}$

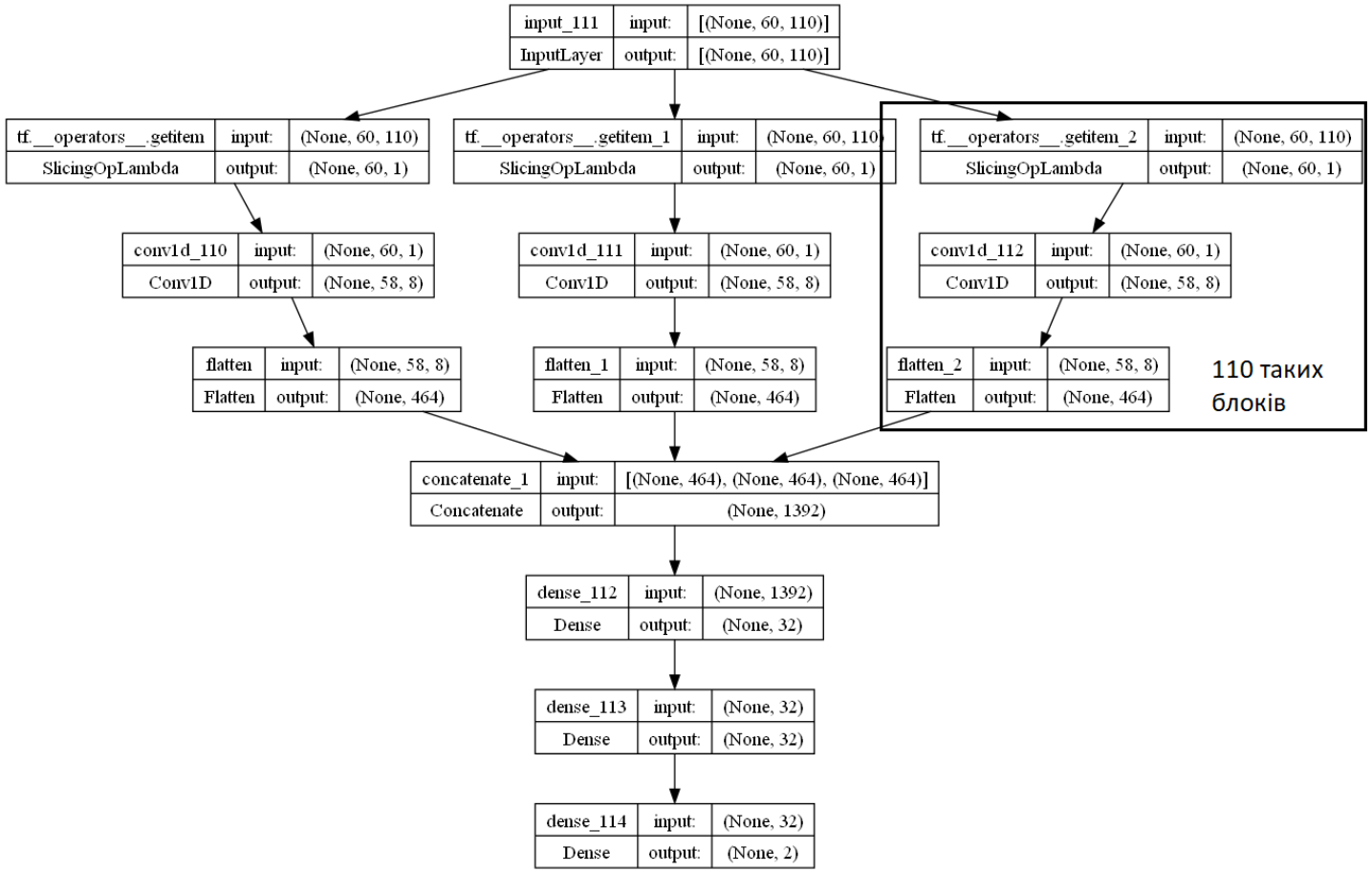

Prediction ranges
Range 1
Range 2
Range 3
high
open
close
low
$score = \sqrt{open - target}$
$\sqrt{prediction - target}$
$score = 10 * \sqrt{|target - prediction|}$

Розбиття даних на тренувальну, тестову і валідаційну вибірки

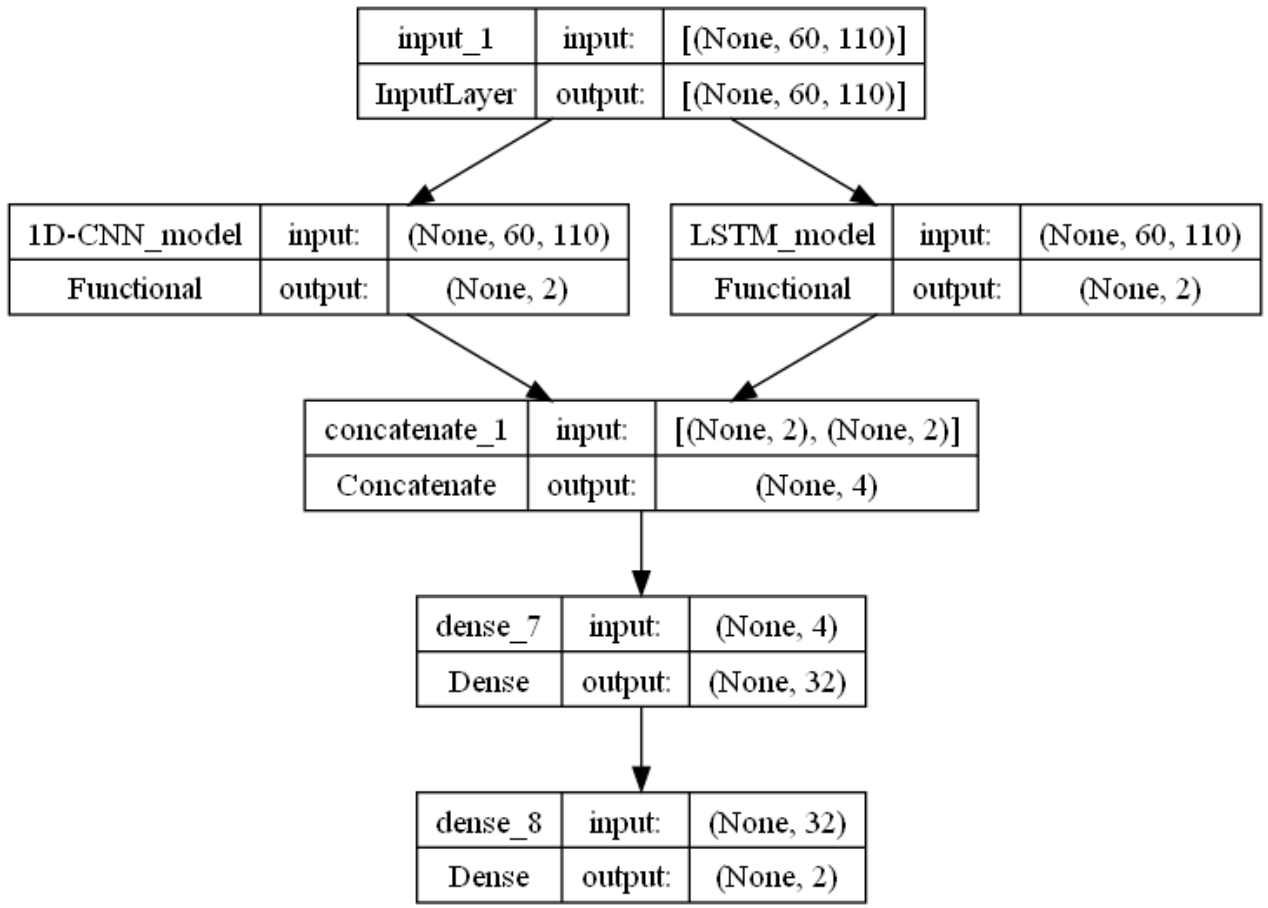
Структура LSTM моделі



Структура 1D-CNN моделі



Структура ансамблевої моделі



Порівняння моделей

Модель	Час навчання	Кількість епох	Похибка валідаційної вибірки	Похибка тренувальної вибірки
LSTM	5:33	140	6.184e-4	5.903e-5
1D-CNN	43:08	100	2.885e-4	1.358e-4
Ensemble	22:45	53	9.904e-4	1.841e-4

Чому не можна напевно стверджувати, що знайдені архітектури мереж є оптимальними.

1. Множина можливих комбінацій гіпер-параметрів неповна (навіть в адекватних межах);
2. Початкові ваги нейронної мережі випадково ініціалізуються перед тренуванням. Ці ваги впливають на те, які важливі функції та особливості модель може навчитися розпізнавати. Так як ми тренували кожну мережу лише по одному разу - результати можуть значно змінитись при перетренуванні навіть вже існуючих мереж.

Як можна покращити результати моделей

1) хоча й було налаштовано систему сповіщень, в даних все одно наявні пропуски в моментах, коли виникали помилки з боку Binance API або Python додатку. Так як пропуски даних виникали не часто і зазвичай не перевищували 20 хвилин - дані було заповнено методом інтерполяції;
2) реалізувати фільтрацію ознак;
3) пошук моментів відкриття/закриття був реалізований спираючись на ціну відкриття свічки. Оптимальніше було б це зробити спираючись на ознаки “low_border”, “high_border”, створені пізніше;
4) реалізувати автоматичну генерацію ознак за допомогою бібліотек tsfresh/tsfel;
5) реалізувати моделі на основі дерев рішень (RandomForest) і моделі на основі градієнтного бустингу (XGBoost, LightGBM, CatBoost) та інших моделей, спробувати інші ансамблі моделей

Висновки

Незважаючи на те, що не всі заплановані підходи вдалося втілити, і не було повністю реалізовано всі потенційні можливості для зросту, в даній роботі було детально описано весь цикл такого виду досліджень. Він включав збір і підготовку даних, визначення цілей і ключових метрик, а також підготовку та налаштування моделей для подальшого порівняння.

Розглянута тема дуже широка і актуальна, а тому вона потребує подальшого дослідження. Це дипломна стане чудовою основою для майбутньої магістрської роботи.

Дякую за увагу!