

ML: аналізування великих даних

Виконав: Колесніков Антон Олеговича

Науковий керівник: к.т.н., доц., зав. каф. Жежерун Олександр Петрович

Актуальність теми

Машинне навчання: аналізування великих даних в електронній комерції

Е-commerce перевищив \$6 трильйонів

У 2023 обсяг е-commerce перевищив \$6 трлн (Statista, 2024)

Швидкість рішень визначає успіх

Успіх залежить від швидкості рішень щодо асортименту, цін, закупівель.

Дані зростають експоненційно

Обсяг даних та інтернет-магазинів зростає експоненційно (McKinsey, 2024).

Зростає інтерес до автоматизації аналітики

Машинне навчання стає ключовим інструментом для аналізу даних в е-commerce.

Мета і завдання дослідження

Мета

Розробка масштабованої ML-системи для автоматизованого аналізу даних у сфері e-commerce з інтеграцією в Telegram-бот

Завдання

- 1 Збір та очищення даних
- 2 Реалізація моделей для класифікації, регресії, кластеризації, виявлення аномалій
- 3 Telegram-інтеграція
- 4 Експерименти
- 5 Рекомендації для бізнесу

Об'єкт, предмет і методи дослідження

Об'єкт

Процес автоматизованого аналізу великих e-commerce-даних

Предмет

ML-методи та алгоритми аналізу e-commerce-даних

Методи

- 1 Аналіз літератури та наукових джерел
- 2 Розробка ПЗ мовою Python
- 3 Машинне навчання (Random Forest, XGBoost, HDBSCAN, Isolation Forest, CLIP)
- 4 Експерименти та валідація моделей
- 5 Графічна візуалізація результатів

Наукова новизна

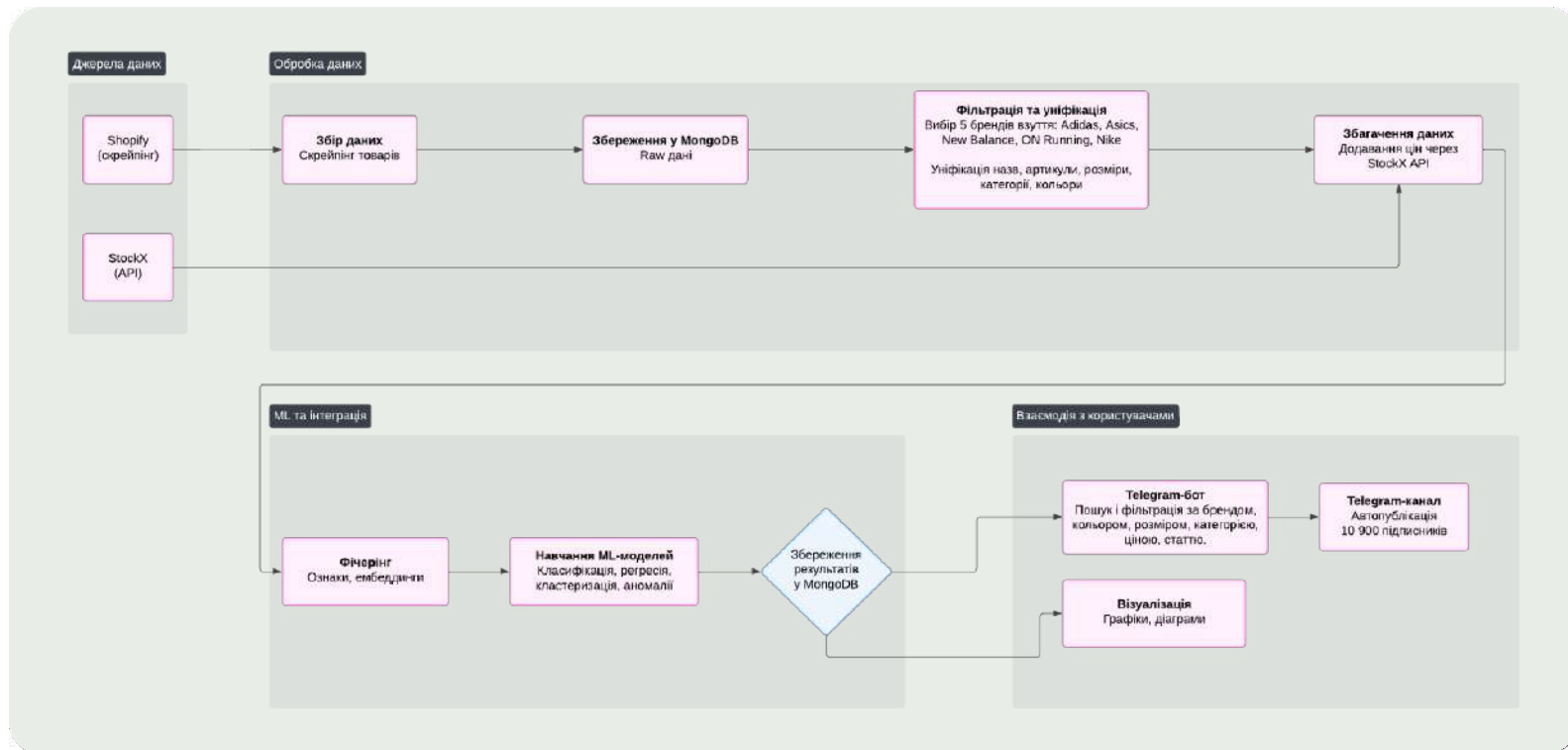
Побудовано ML-конвеєр для
великомасштабної обробки
e-commerce-даних

Використано Telegram-інтерфейс для
інтерактивного доступу до результатів

Інтегровано модель CLIP для аналізу
зображень товарів

Розроблено гнучку систему нормалізації
назв, мультивалютних цін і категорій

Архітектура системи



Моделі машинного навчання

Класифікація маржинальності

XGBoost — визначення рівня маржинальності для взуття топ-брендів. F1-score: 0.9987, ROC-AUC: 1.00, урахуванням ринкових даних. R^2 : 0.80, MAE: \$917, Accuracy: 99.85%

Прогнозування цін

XGBoost Regressor — оцінка оптимальної ціни з урахуванням ринкових даних. R^2 : 0.80, MAE: \$917, RMSE: \$2265

Виявлення аномалій

Isolation Forest — автоматичне виявлення підозрілих цін. Silhouette Score: 0.71, Частка аномалій: 5%

Кластеризація товарів

HDBSCAN — групування схожих товарів за CLIP-ембеддингами. Silhouette Score: 0.62, Davies-Bouldin Index: 0.41

Векторні представлення (ембеддинги)

OpenCLIP — мультимодальні ембеддинги, використані для знаходження схожих товарів за зображенням і назвою.

Експериментальні результати

Обсяг оброблених даних

1,1 МЛН

товарів з 90 магазинів

26 000

очищених, нормалізованих
позицій для навчання моделей

Практичне значення та впровадження

Впроваджено систему для щоденного аналізу товарів і цін на основі ринкових даних

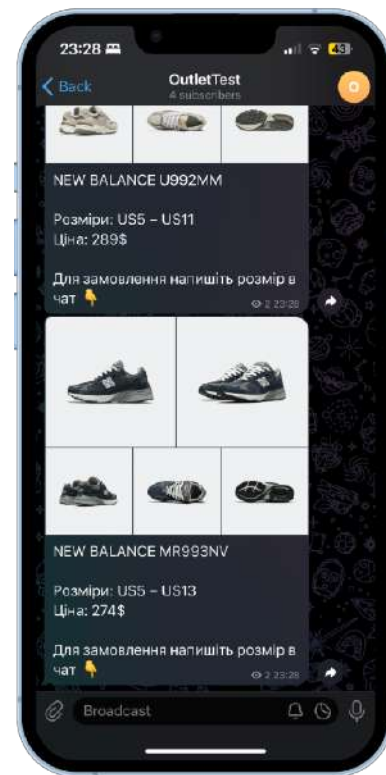
Здійснюється щоденний моніторинг даних з 90 сайтів

Telegram-бот використовується для взаємодії з результатами в реальному бізнесі (10 900 підписників)

Архітектура підтримує масштабування: додавання нових сайтів, брендів, валют і категорій



Інтерфейс адміністратора



Інтерфейс клієнта

Експериментальні результати

Пошук схожих товарів: результат моделі CLIP



Оригінальний товар



Схожість: 0.6297



Схожість: 0.6135

Висновки

- 1 Система проаналізувала 1.1 млн товарів із 90 сайтів
- 2 Побудовано 4 ML-моделей (класифікація, регресія, аномалії, кластери)
- 3 Інтеграція з Telegram-ботом для реального бізнесу
- 4 Досягнута висока точність у задачах ціноутворення та аналізу товарів
- 5 Архітектура готова до масштабування: додавання сайтів, брендів, форматів