

Тищенко В.В., магістрант, Михайлуца О.М., к.т.н., доцент, науковий керівник

ДОСЛІДЖЕННЯ ПРОБЛЕМИ АВТОМАТИЗОВАНОГО АНАЛІЗУ ПОКУПОК З ВИКОРИСТАННЯМ МЕТОДІВ DATA SCIENCE

Запорізька державна інженерна академія, кафедра ПЗАС

Пошук, аналіз та вибір вигідних покупок є досить складним завданням для людського мозку, оскільки вимагає маніпулювання багатьма характеристиками товарів, що розглядаються. Порівняння десятка товарів може викликати труднощі, порівняння сотень і тисяч – фізично неможливе. Зі сторони бізнесу аналіз популярності товарів, дослідження коливань цін та об'єму продаж може стати основним елементом у побудові ефективної і прибуткової стратегії розвитку.

Якість подання товару покупцеві і роз'яснення що і скільки коштує у вибраному товарі є важливим фактором для покупки, а рекомендації на основі його вподобань підштовхнуть його до інших покупок або придбання додаткових аксесуарів чи комплектуючих. В результаті і покупець, і продавець матимуть вигоду, оскільки перший буде бачити за що віддає гроші і зможе детально визначити які параметри обраного товару для нього найважливіші, а другий — отримає додатковий прибуток і матиме можливість відслідковувати тренди в різних категоріях товарів.

Основним засобом дослідження і вирішення поставлених задач є використання методів галузі Data Science, що включають в себе методи статистичного аналізу, Dataminingта машинного навчання [1]. Для визначення патернів у коливанні цін доцільно застосувати аналіз часових рядів. Адекватно оцінити взаємозв'язок і важливість предикторів моделі лише за допомогою кореляційного аналізу, простої чи множинної лінійної регресії — не вийде. Це зумовлено тим, що наврядчб можна говорити про лінійну залежність ціни від регресорів, кількість яких може досягати 20 одиниць. Так само не можна використовувати і поліноміальні чи нелінійні моделі не знаючи наперед вид залежності того чи іншого регресора від ціни окремо взятої одиниці товару. Залишається розглянути алгоритми машинного навчання, в основі яких лежать дерева прийняття рішень — випадковий ліс і градієнтний бустинг. Оптимальним вибором буде той алгоритм, котрий краще покаже себе при перехресній перевірці (крос-валідації).

Ідея градієнтного бустингу полягає в побудові ансамблю послідовно уточнюючих одна одну елементарних моделей, кожна з яких навчається на помилках попередніх.

До переваг даного алгоритму відносять:

- високу точність отриманих результатів;
- велику кількість модифікацій алгоритму;
- можливість підставляти різні моделі — лінійні, узагальнені лінійні, дерева рішень.

До недоліків наведеного алгоритму слід віднести:

- велику кількість параметрів (більше двадцяти);
- необхідність переналаштування параметрів в залежності від даних, що розглядають.

Інший алгоритм — випадковий ліс (random forest). Він будує множину дерев прийняття рішень і усереднює результати їх передбачень. Класифікація об'єктів проводиться шляхом голосування: кожне дерево множини відносить об'єкт який класифікується до одного з класів, після чого перемагає той клас, за який проголосувало найбільше число дерев.

Основні переваги алгоритму «Випадковий ліс»:

- ефективна обробка даних з великою кількістю параметрів;
- однаково добре обробляються як безперервні, так і дискретні ознаки;

- невелика кількість параметрів для налаштування;
- стійкість до викидів.

Основні недоліки наведеного алгоритму:

- схильність до перенавчання;
- великі витрати пам'яті і розмір отримуваних.

Поєднавши розглянуті підходи, можна виконати комплексний аналіз продукції, на основі якого можна оптимізувати стратегію розвитку бізнесу, розробити рейтингову чи рекомендаційну систему продукції і максимально якісно подати товар покупцеві.

Базуючись на актуальності наведених вище проблем, було розроблено систему аналізу товарів, що складається з трьох компонентів:

- система збору даних (завантаження характеристик товарів та даних про їх попит і коливання цін в неочищеному вигляді);
- система попередньої обробки та нормалізації даних (приведення даних до одного формату, зменшення розмірності даних, відсіювання викидів та незначних параметрів, заповнення або фільтрація пропущених значень);
- система аналізу товарів з візуалізацією отриманих результатів та генерацією звітів.

Отже, розроблена система дозволяє детально і всесторонньо аналізувати продукцію як покупцям, так і продавцям, і має наступні можливості:

- зменшення області пошуку;
- визначення вигідності покупок;
- визначення важливостей характеристик товарів;
- визначення впливу кожної характеристики товару на його ціну;
- пошук патернів у коливанні цін та їх прогнозування;
- планування строків та об'ємів закупок продукції;
- є основою для розробки рейтингових та рекомендаційних систем.

Основним засобом реалізації вибрано мову програмування R [1,2]. До її переваг варто віднести відповідну область застосування, гнучкість та наявність багатьох пакетів з реалізованими методами статистичного аналізу та машинного навчання. Також, мова R має широкий набір інструментів для обробки та візуалізації даних, генерації звітів, а також — можливість будувати інтерактивні клієнт-серверні веб-додатки.

Висновки:

- досліджено питання автоматизованого пошуку та аналізу товарів;
- визначено і проаналізовано основні методи та алгоритми вирішення поставлених завдань;
- порівняно зручність та продуктивність технічних засобів розробки;
- розроблено систему збору і попередньої обробки даних;
- розроблено клієнт-серверну систему аналізу і веб-інтерфейс для візуалізації отриманих даних та перегляду звітів.

Література

1. Hadley Wickham. R for Data Science. / Hadley Wickham, Garret Grolemond. – Canada.: O'Reilly Media, Inc., 2016. – 522 с.
2. Наглядная статистика. Используем R! / [Шипунов А.Б., Балдин Е.М., Волкова П.А. и др.]; М.: ДМК Пресс, 2014. – 298 с.