DOI:
УДК 004.04
О.М. Міхайлуца, к.т.н., доцент, elenamikhaylutsa7@gmail.com
А.В. Пожуєв, к.ф.-м.н., доцент, scorpio6828@gmail.com,
В.В. Тищенко, магістр
Запорізький національний університет, Запоріжжя

МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАННИХ ТА ЇХ ЗАСТОСУВАННЯ У СФЕРІ ЕЛЕКТРОННОЇ КОМЕРЦІЇ

Розглянуто питання, пов'язані з областю інтелектуального аналізу даних. Досліджено основні методи та інструменти Data Mining, які застосовуються у високопродуктивній інтелектуальній аналітичній обробці даних. Досліджено проблему автоматизованого аналізу покупок. Представлені результати застосування методів інтелектуального аналізу даних - Data Mining, зокрема, розробленого парсеру та системи попередньої обробки даних. Розроблено клієнт-серверний додаток з веб-інтерфейсом для аналізу та класифікації товарів на основі їх характеристик.
***Ключові слова**: статистичний аналіз; машинне навчання; Data Science; Data Mining.*

Here, the issues related to the data mining sphere are considered. The basic methods and tools of Data Mining that are used in high-performance intelligent analytical processing are investigated. The problem of automated purchases analysis is also researched. The results of Data Mining methods application, in particular, the developed parser and the system of preliminary data processing are presented. Client-server application with web interface for analysis and classification of goods based on their characteristics is developed.
***Keywords**: statistical analysis; machine learning; Data Science; Data Mining.*

Постановка проблеми

За допомогою методів галузі Data Science можна перетворити дані в цінні знання на основі яких можна приймати рішення і робити передбачення. Це особливо корисно в комерційних цілях, де правильні рішення і прогнози приносять прибуток бізнесу, а неправильні — призводять до збитків та втрати клієнтів. Переважна більшість існуючих систем для аналізу продукції орієнтується на дані з рівню продажів або кількості переглядів товарів, а це вимагає тривалого часу функціонування бізнесу для можливості отримання якісних результатів. Очевидно, що аналіз товарів на основі їх характеристик краще підходить для підтримки бізнесу на початкових стадіях, а також може залучити нових клієнтів та покращити стан вже існуючої бази. Отримати з даних корисні знання для прогнозування і прийняття рішень можна за допомогою статистичних методів, методів Data Mining'у та машинного навчання, які й застосовуються для вирішення поставлених задач у даній роботі.

Для розробки системи аналізу вкрай важливим питанням є підбір методів, що будуть узгоджуватись з наявними даними, оскільки деякі алгоритми висувають вимоги до розподілу даних та наявності викидів.

Серед методів статистичного аналізу доцільно використати кореляційний та дисперсійний аналіз. За допомогою отриманих результатів можна дослідити взаємозв'язки між параметрами товарів і позбутись неважливих або сильно корельованих предикторів у навчальних вибірках для машинного навчання. Для задач класифікації та регресії слід звернути увагу на ансамблеві методи [1], особливо ті, в основі яких лежать дерева прийняття рішень, оскільки вони не висувають жодних вимог щодо розподілу даних і є стійкими до викидів.

Важливим етапом аналізу є збір та підготовка даних, що може займати до 80% часу, коли для побудови моделей та інтерпретації результатів залишається лише 20% часу. Отже, застосування методів Data Science, крім іншого, включає також збір і попередню обробку даних. Такий підхід дозволить провести аналіз товарів маючи лише їх характеристики. Якщо зібрати з

відкритих джерел дані про вартість товарів за певний період, то додатково можна виконати аналіз часових рядів для дослідження та прогнозування майбутньої поведінки цін.

Аналіз останніх досліджень та публікацій

Проблема обробки великих обсягів даних та інформації виникла разом з розвитком обчислювальної техніки. Дослідження методів інтелектуального аналізу даних, а також прийняття рішень на підставі цього аналізу розглянуто у багатьох наукових роботах. Так, у роботі [2] розглядається поняття «знання», значення знань для САПР, концепція та історія розвитку баз і сховищ даних, використання знань в системах підтримки прийняття рішень, актуальні напрямки інтелектуального аналізу даних, а також значення генетичних алгоритмів і еволюційного моделювання для інтелектуального аналізу даних.

Наукове дослідження [3] присвячено Технології Інтелектуального Аналізу Даних (Data Mining) - однієї з областей інформаційних технологій, що активно розвивається, яка призначена виявленню корисних знань з баз даних різної природи. Аналіз великих наборів даних - так званих Big data - стане ключовою основою конкуренції, яка буде лежати в основі нових хвиль росту продуктивності, інновацій та споживчого надлишку, якщо будуть діяти правильні політики та інструменти підтримки. Дослідження, проведене Глобальним інститутом McKinsey і офісом бізнес-технологій McKinsey [4], аналізує стан цифрових даних і документує значну цінність, яку потенційно можна розблокувати. Великі дані допоможуть створити нові можливості зростання і зовсім нові категорії компаній. Багато з них будуть компаніями, які знаходяться в центрі великих інформаційних потоків, де можна збирати і аналізувати дані про продукти і послуги, покупців і постачальників, перевагах і наміри споживачів.

Велика кількість наукових робіт розглядає переваги застосування інтелектуального аналізу у бізнесі та електронній комерції. Зосередивши увагу в першу чергу на прикладах з галузі охорони здоров'я, в статті [5] коротко пояснюється, чому «великі дані» дійсно відрізняються через їх впливу на усталені підходи до створення знань. У роботі [6] виявлено переваги використання великих даних в електронній комерції в порівнянні з традиційною комерцією. На основі проведеного дослідження автором пропонується використовувати систему обслуговування електронної комерції на основі великих даних за допомогою майданчики електронної комерції. У статті [7] розглядаються проблеми статистичного аналізу споживчих переваг в електронній комерції (на прикладі готельної галузі). Автором проведено множинний регресійний аналіз для оцінки впливу різних чинників на споживчий вибір і інтернет-продажу в даній галузі. З використанням ціннісно-орієнтованого маркетингового підходу в статті запропоновано метод розрахунку індексу сприймається споживчої цінності (карти цінності) в електронній комерції.

Формулювання мети дослідження

Предметом дослідження даної роботи є методи статистичного аналізу, дата-майнінгу та машинного навчання у застосуванні до класифікації товарів, а також особливості побудови моделей інтелектуального аналізу та налаштування їх параметрів.

Виклад основного матеріалу

У випадку відсутності достатньої кількості даних для аналізу їх можна зібрати з різноманітних веб-ресурсів в мережі Інтернет. Добування даних з мережі Інтернет можна виділити в окремий напрямок, який називається веб-майнінгом. Цей процес, як правило, носить більш практичну складову, ніж теоретичну, і зводиться до розробки парсерів web-сторінок з HTML-кодом. Існує п’ять основних підходів для добування даних таким чином:

- Аналіз DOM – дерева сторінки.
- Парсинг рядкових даних.
- Використання регулярних виразів.
- Парсинг XML.
- Візуальний парсинг.

Перед тим як подати дані на вхід аналітичної моделі чи алгоритму машинного навчання, необхідно виконати їх очистку та попередню обробку [8], що є важливими етапи, які забезпечують ефективне використання набору даних та точність отриманих результатів. Отримані в результаті збору дані повинні відповідати певним критеріям якості. Даними високої якості вва-

жаються повні, точні, своєчасні дані, які піддаються інтерпретації. Серед типових проблем з якістю даних виділяють такі, як: неповнота, дані не містять атрибутів або є пропущені значення; шум, дані містять помилкові записи або викиди; неузгодженість, наявні конфліктуючі записи або розходження; відмінності у форматах запису. При наявності проблем для перевірки якості даних оцінюють такі показники як: кількість записів, кількість атрибутів, типи даних атрибутів, кількість пропущених значень, правильність формату даних, узгодженість даних, наявність викидів.

Використання методів статистичного аналізу дозволяє дослідити зв'язок між параметрами, оцінити їх значущість та позбавитись неважливих або сильно корельованих предикторів перед формуванням тестових та навчальних вибірок для алгоритмів машинного навчання.

Зв'язок між параметрами виявляють за допомогою кореляційного аналізу [9], який дозволяє визначити силу і напрям стохастичної взаємодії між змінними (випадковими величинами). Якщо змінні виміряні, як мінімум, в інтервальній шкалі і мають нормальний розподіл, то кореляційний аналіз здійснюється шляхом обчислення коефіцієнта кореляції Пірсона, в іншому випадку використовується кореляція Спірмена або Кендала. Для дослідження впливу однієї незалежної якісної змінної або групи таких змінних (факторів) на залежну кількісну змінну (відгук) застосовується дисперсійний аналіз (AnalysisOfVariance) [10]. Для поставлених задач особливо цікавим буде дослідження впливу окремих факторів на ціну.

Аналіз часових рядів і прогнозування цін проводиться за ARIMA-моделлю. Для налаштування моделі використовувався інформаційний критерій Акаіке [11], який реалізовано в пакеті forecast для мови R. Критерій засновано на теорії інформації, він пропонує відносні оцінки втраченої інформації при застосуванні даної моделі для представлення процесу, що породжує дані, досягаючи при цьому компромісу між пристосованістю моделі та її складністю. Для вирішення задач регресії та класифікації розглядаються методи, в основі яких лежать дерева прийняття рішень, зокрема, ансамблеві алгоритми машинного навчання.

Серед методів класифікаційного аналізу виділяють дерева класифікації, що дозволяє визначати належність об'єктів до того чи іншого класу в залежності від відповідних значень ознак, що характеризують об'єкти. На відміну від класичного дискримінантного аналізу, дерева класифікації здатні виконувати одновимірне розгалуження за змінними різних типів: категоріальним, порядковим, інтервальним. По аналогії з дискримінантим аналізом метод дає можливість аналізувати вклад окремих змінних в процедуру класифікації. Можливість графічного представлення результатів і простота інтерпретації багато в чому пояснюють популярність дерев класифікації в прикладних областях, проте, найважливіші відмінності дерев класифікації — це їх ієрархічність та широка застосовність. Даний підхід дозволяє за керованими параметрами будувати дерева довільної складності, досягаючи мінімальної помилки класифікації. Оскільки, за складним деревом досить важко класифікувати новий об'єкт через велику кількість правил, то при побудові дерев класифікації необхідно шукати розумний компроміс між складністю дерева та трудомісткістю процедури класифікації.

Для задач класифікації та регресії використовують алгоритм машинного навчання – градієнтний бустинг [12], який виробляє прогностичну модель у формі ансамблю слабких моделей. Даний алгоритм будує моделі у послідовній манері як і інші алгоритми бустингу і узагальнює їх, дозволяючи оптимізацію довільної диференційованої функції втрат. До переваг градієнтного бустингу відносять такі: висока точність отриманих результатів; велика кількість модифікацій алгоритму; можливість підставляти різні моделі. При оптимізації параметрів часто вибирають невелике число дерев і підлаштовують під нього значення інших параметрів. При побудові фінального алгоритму збільшують число дерев і підлаштовують під нього темп навчання не змінюючи інших параметрів. Такий підхід дозволяє досягати заданого рівня точності.

З практичної точки зору алгоритм машинного навчання Random Forest (випадковий ліс) [13] має значну перевагу, оскільки саме він майже не потребує конфігурування. Даний алгоритм для отриманих даних випадковим чином створює множину дерев прийняття рішень, а потім усереднює результати їх передбачень. На відміну інших алгоритмів машинного навчання, будь то регресія чи нейронна мережа, Random Forest має лише один важливий параметр, що вимагає налаштування (розмір випадкової підмножини, що обирається на кожному кроці побу-

дови дерева). Проте, навіть використовуючи значення за замовчуванням, можна отримати досить прийнятні результати. Класифікація об'єктів проводиться шляхом голосування: кожне дерево множини відносить об'єкт, який класифікується, до одного з класів, і перемагає клас, за який проголосувало найбільше число дерев. Оптимальне число дерев підбирається таким чином, щоб мінімізувати помилку класифікатора на тестовій вибірці. У разі її відсутності, мінімізується оцінка помилки out-of-bag: частка прикладів навчальної вибірки, неправильно класифікованих множиною дерев, якщо не враховувати голоси дерев на прикладах, що входять в їх власну навчальну підвибірку. Випадкові ліси можуть бути природним чином використані для оцінки важливості змінних в задачах регресії та класифікації. Першим кроком в оцінці важливості змінної в тренувальному наборі є тренування випадкового лісу на цьому наборі. Під час процесу побудови моделі для кожного елементу тренувального набору вважається так звана out-of-bag помилка. Потім для кожної сутності така помилка опосередковується по всьому випадковому лісі.

Для оцінки реальних можливостей моделі і налаштування її параметрів в задачах машинного навчання досить часто використовують крос-валідацію. При цьому виділяється деяка множина розбиттів вихідної вибірки на навчальну та контрольну підвибірки, в подальшому для кожного розбиття алгоритм спочатку налаштовується по навчальній підвибірці, а потім оцінюється його середня помилка на контрольній підвибірці. Оцінкою такої перевірки називають середню по всім розбиттям величину помилки на контрольних підвибірках. Для незміщеної оцінки ймовірності помилки, отриманої шляхом крос-валідації, необхідно, щоб навчальна та контрольна вибірки утворювали підмножини, які не перетинаються, що дозволяє уникнути перенавчання.

Зазвичай система аналітики спирається більше на технічні характеристики продукції, ніж на аналіз рівню продаж або статистики переглядів товарів. Крім того, існуючи аналітичні системи не мають необхідних інструментів для класифікації товарів на групи за вигідністю для покупки. Отже, застосування чи модифікація існуючих рішень, зважаючи на вартість їх придбання та надлишкову, або місцями недостатню функціональність, не є доцільним рішенням для вирішення поставлених задач. Тому доцільним є розроблення програмного комплексу, що складається з трьох взаємопов'язаних системи (парсер, система попередньої обробки даних, система аналізу), результати роботи кожної з яких подаються на вхід наступної. Діаграма варіантів використання для зазначених систем наведена на рис. 1.

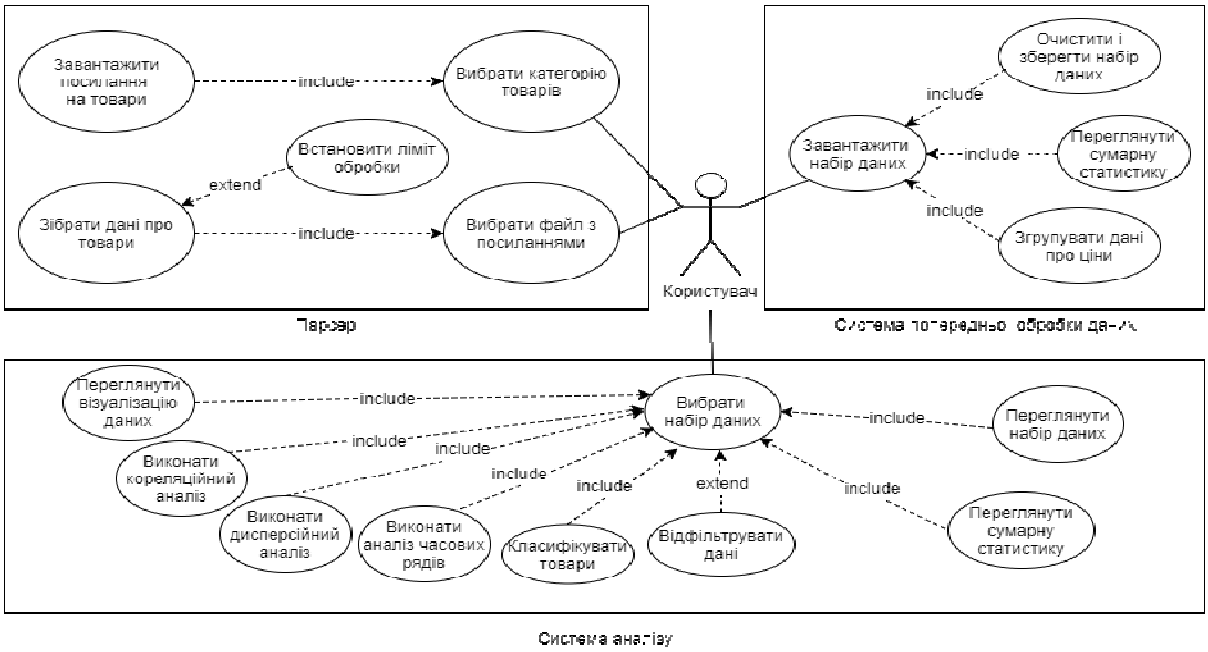

Рис. 1. Діаграма варіантів використання парсеру, системи попередньої обробки даних та системи аналізу

Система парсингу є десктопним додатком, що використовується для збору даних про товари з мережі Інтернет, і не надається для використання кінцевим користувачам. Для розробки парсеру веб ресурсів було вибрано мову програмування C# [14], технологію WPF для розробки користувацького інтерфейсу і середовище розробки Microsoft Visual Studio. Для досягнення продуктивності, розглянуто класичний інтерпретатор мови R та інтерпретатор Microsoft R Open, що пропонує розпаралелювання обчислень по різних потоках.

Система попередньої обробки інформації призначена для очистки зібраних даних за допомогою парсеру даних. До її функцій входить обробка текстових даних з метою приведення їх до одного формату запису, пошук і заповнення пропущених даних, видалення дублікатів, агрегація даних для аналізу часових рядів, збереження сформованих наборів даних. Архітектура системи попередньої обробки даних складається з набору скриптів наступного призначення: завантаження наборів даних у форматі .txt до системи у формат data.table; скриптів, що виконують очищення завантажених даних і їх подальше збереження у відповідному форматі; скрипти для агрегації часових рядів; скрипти для перегляду сумарної статистики по окремих змінних з набору даних. Використання системи здійснюється через виконання скриптів або окремих їх команд з середовища розробки і не надається для використання кінцевим користувачам.

Система аналізу призначена для проведення кореляційного та дисперсійного аналізу, аналізу часових рядів та класифікації товарів на вигідні та не рекомендовані для покупки. Функціональність системи включає зручний інтерфейс для перегляду, фільтрування та візуалізації даних. Робота з системою аналізу може виконуватись в режимі offline за допомогою середовища розробки та браузера, але такий варіант не надається для використання кінцевим користувачам.

Парсер веб-ресурсів володіє мінімалістичним десктопним Windows-інтерфейсом, котрий складається з одного вікна. Серед елементів керування доступний список категорій для парсингу, а також поля, що відображають кількість доступних для обробки елементів (сторінок або товарів), кількість вже оброблених елементів та витрачений на виконання час.

Система попередньої обробки даних представлена набором скриптів, що виконуються в режимі командного рядку в середовищі розробки. Інтерфейс середовища може бути налаштований індивідуально, але зазвичай складається з чотирьох елементів:

- Вікно коду скрипту, який можна виконувати повністю або покомандно, виділивши необхідні рядки.
- Вікно консолі, що містить форматований вивід результатів і дозволяє виконувати окремі команди не вносячи змін до скрипту.
- Вікно зі списком змінних, що містить коротку інформацію про кожний об’єкт у пам’яті, а також дозволяє переглянути повну структуру і зміст таких об’єктів як датафрейм, список і т.п.
- Вікно попереднього перегляду графічного виводу, що містить графічні зображення або html-сторінки.

Система аналізу володіє адаптивним веб-інтерфейсом побудованим за допомогою UI-фреймворку Bootstrap, вся функціональність якої логічно розділена на змістовні вкладки: Explore, Summary, Visualisation, Correlation, ANOVA, Time-series, Classification (рис.2).

В рамках аналізу можливостей створеної системи за допомогою розробленого парсеру зібрано дані для аналізу про більше ніж 8000 видів цифрової техніки. На наступному кроці будується матриця коефіцієнтів кореляції розрахованих за методом Спірмена, дослідження якої дозволяє проаналізувати зв’язки між параметрами техніки, а також, виявити висококорельовані предиктори і незначні параметри. Для подальшого полегшення моделі рекомендовано видалення виявлених параметрів з набору даних перед використанням його в алгоритмах класифікації чи регресії. В межах проведення дисперсійного аналізу для поставлених задач досліджено вплив різних параметрів техніки на ціну. Окрім того, система дозволяє детальніше розглянути вплив змінних одна на одну дослідивши їх розподіл, наприклад розподіл ціни по виробниках (рис. 3).

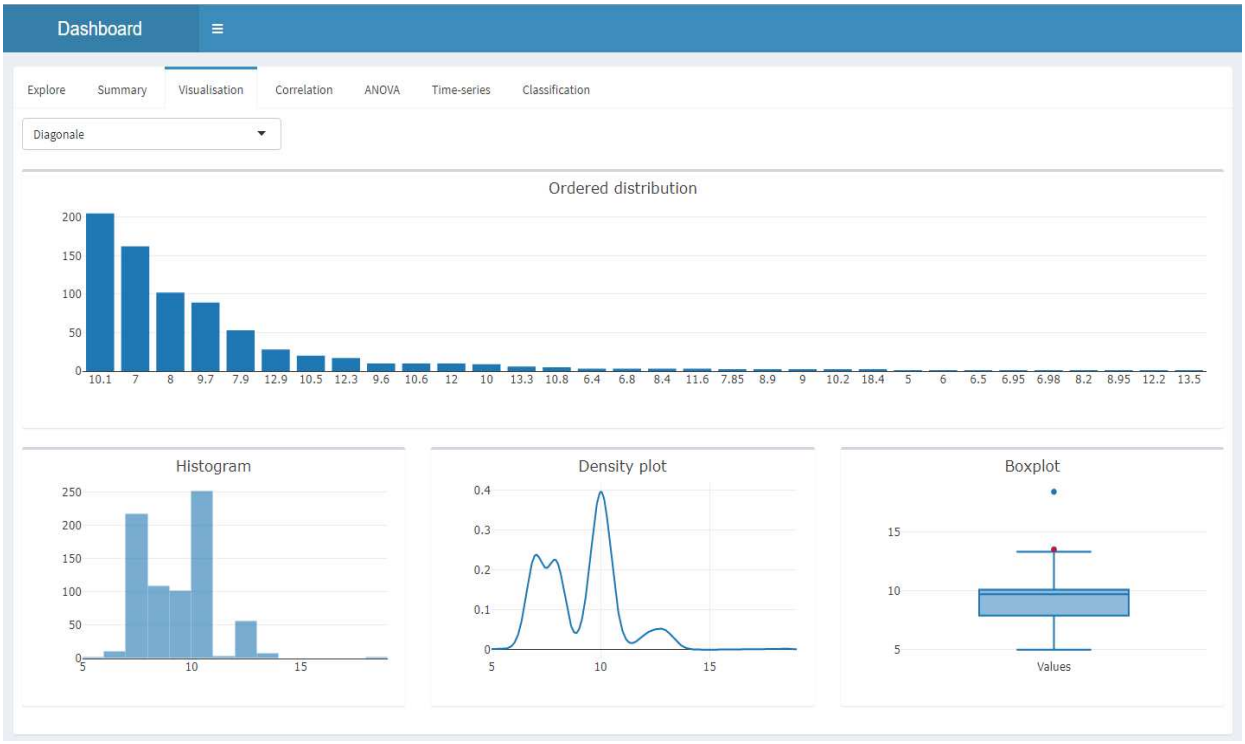

Рис. 2. Інтерфейс вкладки Visualisation

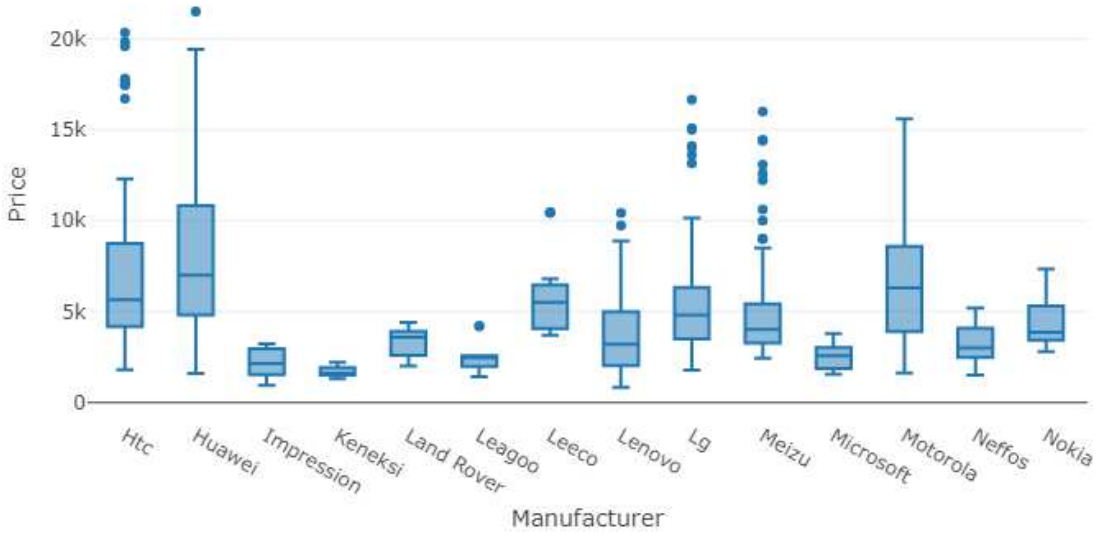

Рис. 3. Розподіл цін на мобільні телефони по окремих виробниках

Для застосування ARIMA-моделі для аналізу дані часового ряду спочатку групуються по місяцях, а потім виконується декомпозиція ряду для виділення окремих компонентів: тренду, сезонної та випадкової складової. Тренд є довгостроковою складовою і відображає поведінку ряду з плином часу, сезонна складова відображає мінливість ціни по місяцях. Також, за допомогою розробленої моделі можна спрогнозувати поведінку цін у майбутньому (рис. 4).

Рис. 4. Прогноз поведінки цін на мобільні телефони

Оскільки вимога стаціонарності є необхідною для отримання адекватної моделі і достовірних прогнозів, у випадку нестаціонарних часових рядів їх можна зробити стаціонарними шляхом взяття різниць різних порядків або інших перетворень, наприклад, шляхом логарифмування. Для класифікації товарів на вигідні і не рекомендовані для покупки порівнюються алгоритми машинного навчання випадковий ліс та градієнтний бустинг. Наведені алгоритми використовуються на задачі регресії з метою оцінки вартості товарів і порівняння отриманого значення з реальною ціною. Для налаштування параметрів моделей використовується процедура крос-валідації, вибірка розбивається по принципу 80 на 20, налаштування параметрів моделі виконується за допомогою пакету caret. Налаштування параметрів моделі для градієнтного бустингу для вказаного набору даних зайняло близько 5 хвилин, що не є хорошим результатом для роботи в режимі реального часу з мінливими даними. Це зумовлено великою кількістю параметрів алгоритму. Також, градієнтний бустинг не працює з категоріальними даними і вимагає приведення їх до числових значень. Хоча точність даного алгоритму є однією з кращих на теперішній час, було вирішено відмовитись від його використання. Іншим алгоритмом для вирішення поставленої задачі є випадковий ліс. Даний алгоритм не потребує тривалого налаштування, досить добре працює "з коробки" і може оброблювати категоріальні дані, яких досить багато в наявних наборах даних для аналізу. Налаштування параметрів моделі зайняло близько 10 секунд, що є набагато кращим результатом, ніж у градієнтного бустинга. Підібраний параметр mtry є розміром випадкової підмножини параметрів розбиття, що вибирається на кожному кроці. По замовчуванню це значення приймає значення кількості предикторів розділеної на три і округленої донизу. На поставлених задачах значення по замовчуванню дає приблизно ту саму похибку. Якщо додатково побудувати графік залежності середньоквадратичного відхилення прогнозних і реальних цін в залежності від кількості дерев, то можна помітити що приблизно після 100 дерев збільшення їх кількості не призводить до суттєвого покращення точності, вони виходить на асимптоту (рис. 5).

Отже дослідним шляхом підібрано і кількість дерев для побудови моделі. Побудована модель має середньоквадратичне відхилення прогнозних цін від реальних, для діапазону від 6000 до 8000 гривень це складає менше 5%, що говорить про досить високу точність моделі.

Проведені дослідження довели важливість параметрів товарів як критеріїв розбиття. Приклад класифікації товарів наведено на рис. 6.

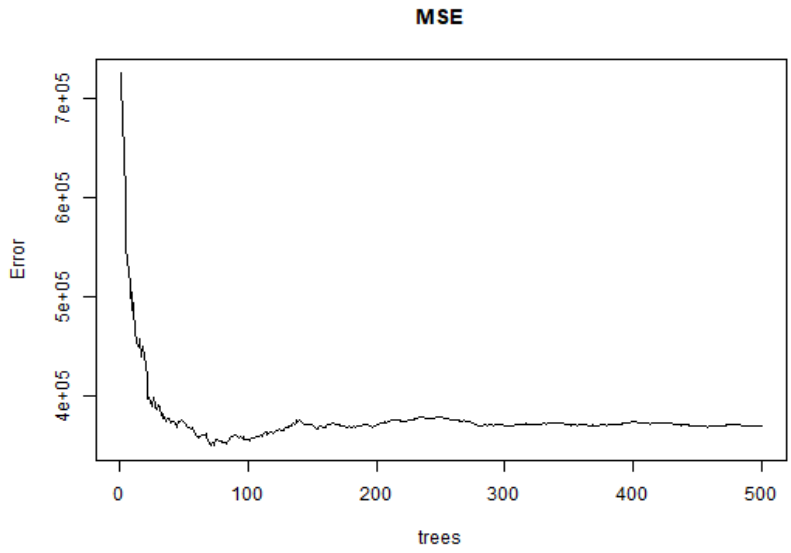

Рис. 5. Залежність СКВ похибки цін від кількості дерев

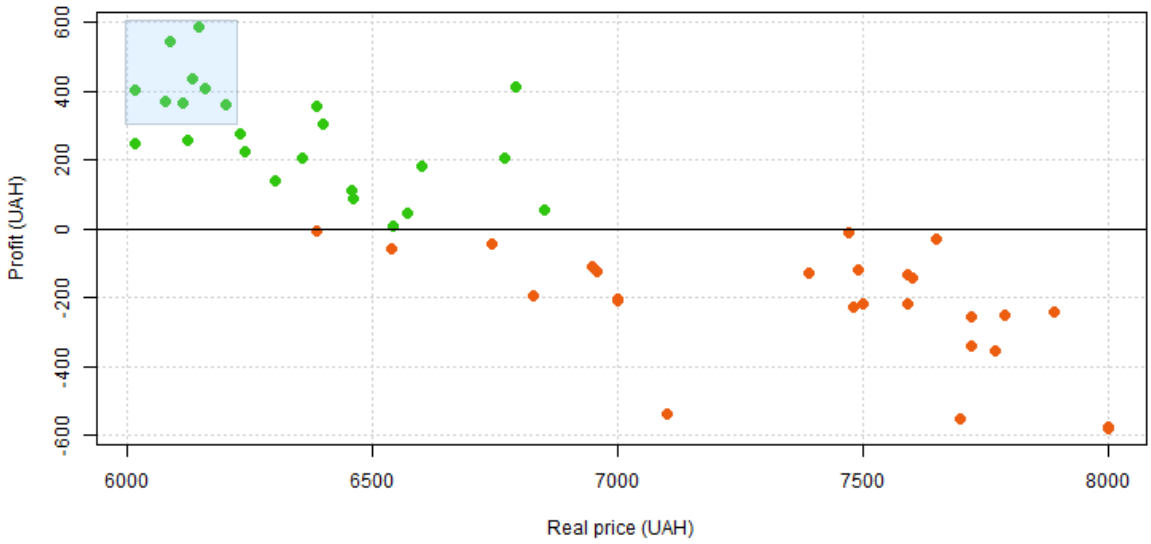

Рис. 6. Приклад класифікованих товарів

Для перегляду інформації про товар необхідно виділити потрібні спостереження на графіку і вони з'являться в таблиці під ним (табл. 1).

Таблиця 1. Інформація про виділені класифіковані товари

Model	RealPrice	Predicted	Profit	%
Lenovo Tab 3 Plus TB-8703X 16GB LTE Deep Blue	6144	6729.45	585.45	9.53
Lenovo Tab 3 Plus X70F 3G 16GB Slate Black	6085	6632.41	547.41	9.00
Lenovo Yoga Tablet 3 X50M LTE 16Gb Black	6131	6565.92	434.92	7.09
ONDA V919 3G Air 64GB (Gold)	6015	6419.39	404.39	6.72
Apache A83	6157	6567.19	410.19	6.66
Dell Venue 11 Pro 64GB (11-2500, FTCWD02H)	6076	6445.41	369.41	6.08
GoClever TAB R104	6113	6480.12	367.12	6.01
Samsung Galaxy Tab A 9.7 16GB LTE (Smoky Titanium)	6199	6561.96	362.96	5.86

Висновки та перспективи подальших досліджень

В ході виконання роботи було досліджено ряд методів галузі Data Science, що включають кореляційний та дисперсійний аналіз, аналіз часових рядів, градієнтний бустинг та випадковий ліс для задачі регресії. На основі проведеного аналізу методів статистичного аналізу, дата-майнінгу та машинного навчання використано набір алгоритмів для аналізу покупок на основі тільки їх характеристик, в той час як практично всі існуючі системи спираються на дані про рівень продажів або кількість переглядів товару, що вимагає тривалого часу функціонування бізнесу. За допомогою розробленої системи досліджено взаємозв'язок між характеристиками товарів, їх вплив на ціноутворення, досліджено і прогнозовано поведінку цін, визначено товари з завищеною або заниженою ціною.

Результат роботи можна використовувати у сфері електронної комерції типу «Business to business», а саме, в електронній торгівлі.

Список використаної літератури

1. История развития ансамблевых методов классификации в машинном обучении [Електронний ресурс] /Ю.С. Кашницкий, 2015. Режим доступу: https://www.researchgate.net/publication/278019662_Istoria_razvitia_ansamblevyh_metodov_klassifikacii_v_masinnom_obucenii (дата звернення: 04.03.2020). Назва з екрану.
2. Коваленко О. С. Обзор проблем и перспектив анализа даннях. *Информатика, вычислительная техника и инженерное образование*. 2010. № 2. С. 15–31.
3. Степанов Р. Г. Технология Data Mining: интеллектуальный анализ даннях. Казань: Казанский государственный университет им. В.И.Ульянова-Ленина (КГУ), 2008. 58с.
4. Manyika J., Chui M. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011. 156 p.
5. Dhar V. Data Science and Prediction. *Communications of the ACM*. 2013. Vol. 56, № 12. P. 64–73.
6. Цинбяо С. Особенности применения больших данных в электронной коммерции. *Международная торговля и торговая политика*. 2017. № 4. С. 114–119.
7. Никитина О. В. Статистический анализ потребительских предпочтений в электронной коммерции. *Вопросы статистики*. 2015. № 6. С. 46–52.
8. Хань Ц., Кэмберб М., Пей Ц.. Интеллектуальный анализ данных: концепции и методы. Morgan Kaufmann Publishers, 2011. 132 с.
9. Шорохова И. С., Кисляк Н. В., Мариев О. С. Статистические методы анализа: [учеб. пособие]. Екатеринбург: Изд-во Урал. ун-та, 2015. 300 с.
10. Уикем Х., Гроулмунд Г.. Язык R в задачах науки о данных: импорт, подготовка, обработка, визуализация и моделирование даннях. Вильямс, 2017. 592 с.
11. Taddy M. Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions. New York: McGraw-Hill, 2019. 90 p.
12. Abe N., Freund Y., Schapire R. A Short Introduction to Boosting [Електронний ресурс] / N. Abe, Y. Freund, R. Schapire, 1999. – Режим доступу: https://www.semanticscholar.org/paper/A-Short-Introduction-to-Boosting-Abe-Freund/147f1c125a26d3115ee78a033df0c33eeb8c430b/pc.
13. Breiman L. Random Forests. *Machine Learning.* 2001. № 45(1). P. 5–32.
14. Троелсен Э. Язык программирования C# 5.0 и платформа .NET 4.5. М.: ООО "И.Д. Вильямс", 2013. 1312 с.

METHODS OF DATA MINING AND THEIR APPLICATIONS IN THE ELECTRONIC COMMERCE

Mikhailutsa O.M., Pozhuyev A.V., Tishchenko V.V.

Abstract

The present paper deals with the problem of intellectual data analysis using statistical analysis, data mining and machine learning. An important step in the analysis is the collection and preparation of data, which can take up to 80% of the time, while only 20% remains to build models and interpret the results. A number of mathematical and statistical methods have been analyzed to select the optimal algorithms on which the software package is based. The study of the influence of individual factors on price and the relation between parameters is evaluated using correlation and analysis of variance. Time series analysis and pricing are based on the ARIMA model. For classification and regression problems, machine learning algorithms are used - gradient boosting and random forest. The real-world capabilities of the model and the tuning of its parameters in machine learning tasks are performed by cross-validation. Based on these algorithms, a software has been developed consisting of three interconnected systems (parser, pre-processing system and analysis system), the working results of which are served to the input of the next.

Parsing is a desktop application used to collect data about products from the Internet and is not provided for end-users. The functions of the pre-processing module include the processing of text data in order to bring them into one format of recording, finding and filling in the missing data, removing duplicates, aggregating data for analysis of time series and saving the formed data sets. The analysis system is designed to perform correlation and variance analysis, time series analysis and product classification for advantageous and not recommended for purchase. The system's functionality includes a user-friendly interface for viewing, filtering and visualizing data.

As a part of the capabilities analysis of the created system, the developed parser collected data for the analysis of more than 8000 types of digital technology. Using the analysis of variance, the impact of variables on one another is considered by examining their distribution. The use of the ARIMA model for analysis allowed us to build a trend change that reflects the behavior of a series over time and a seasonal component. In addition, the developed model can predict the behavior of prices in the future. To classify goods as advantageous and not recommended for purchase, such machine learning algorithms as random forest and gradient boosting are compared. The study found that a random forest does not require long tuning, works well out of the box, and can handle categorical data, which is abundant in the available datasets for analysis. The result can be used in "Business to business" e-commerce.

References

[1] Kashnitsky Yu.S. (2015). Istoriya razvitiya ansamblevykh metodov klassifikatsii v mashinnom obuchenii [The history of the development of ensemble classification methods in machine learning]. Retrieved from https://www.researchgate.net/publication/278019662_Istoria_razvitia_ansamblevyh_metodov_klassifikacii_v_masinnom_obucenii

[2] Kovalenko O.S. (2010). Obzor problem i perspektiv analiza dannykh [Overview of problems and prospects of data analysis]. *Informatika, vychislitel'naya tekhnika i inzhenernoye obrazovaniye – Computer science, computer engineering and engineering education, 2,* 15–31 [in Ukrainian].

[3] Stepanov R.G. (2008). *Tekhnologiya Data Mining: intellektual'nyy analiz dannykh [Data Mining Technology: Data Mining]*. Kazan': Kazanskiy gosudarstvennyy universitet im. V.I.Ul'yanova-Lenina [in Russian].

[4] Manyika J. & Chui M. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

[5] Dhar V. (2013). Data Science and Prediction. *Communications of the ACM. (Vols. 56), 12,* 64–73.

[6] Tsinbyao S. (2017). Osobennosti primeneniya bol'shikh dannykh v elektronnoy kommertsii [Features of the use of big data in e-commerce]. *Mezhdunarodnaya torgovlya i torgovaya politika – International trade and trade policy, 4,* 114–119 [in Russian].
[7] Nikitina O. V. (2015). Statisticheskiy analiz potrebitel'skikh predpochteniy v elektronnoy kommertsii [Statistical analysis of consumer preferences in electronic commerce]. *Voprosy statistiki – Questions of statistics, 6,* 46–52 [in Russian].
[8] Han C., Camberb M. & Pei C. (2011). *Data Mining: Concepts and Methods.* Morgan Kaufmann Publishers.
[9] Shorokhova I.S., Kislyak N.V. & Mariev O.S. (2015). *Statisticheskiye metody analiza [Statistical analysis methods].* Yekaterinburg: Izd-vo Ural. un-ta [in Russian].
[10] Wickem H. & Groulmund G. (2017). *Language R in the problems of data science: import, preparation, processing, visualization and modeling of data.* Williams.
[11] Taddy M. (2019). *Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions.* New York: McGraw-Hill.
[12] Abe N., Freund Y. & Schapire R. (1999). A Short Introduction to Boosting. Retrieved from https://www.semanticscholar.org/paper/A-Short-Introduction-to-Boosting-Abe-Freund/147f1c125a26d3115ee78a033df0c33eeb8c430b/pc.
[13] Breiman L. (2001). Random Forests. *Machine Learning, 45(1),* 5–32.
[14] Troelsen E. (2013). *Programming language C # 5.0 and platform .NET 4.5.* Williams.