УДК 339.138: 004.89

Алгоритми Data Science у моделюванні бізнес-процесів

Гнот Т.В.
аспірант
Національного університету біоресурсів і природокористування України
Негрей М.В.
кандидат економічних наук, доцент
Національного університету біоресурсів і природокористування України

Успішне функціонування бізнесу передбачає прийняття рішень в умовах невизначеності та великої кількості інформації. Необхідною передумовою ефективного управління бізнес-процесами є використання сучасних підходів до моделювання, які базуються на алгоритмах Data Science. У статті розглянуто базові підходи Data Science – моделі керованого навчання та моделі некерованого навчання, їх основні алгоритми та показано їх практичну реалізацію у моделюванні бізнес-процесів.
Ключові слова: Data Science, алгоритм, модель, бізнес-процес, рішення.

Гнот Т.В., Негрей М.В. АЛГОРИТМЫ DATA SCIENCE В МОДЕЛИРОВАНИИ БИЗНЕС-ПРОЦЕССОВ
Успешное функционирование бизнеса предполагает принятие решений в условиях неопределенности и большого количества информации. Необходимым условием эффективного управления бизнес-процессами является использование современных подходов моделирования, основанных на алгоритмах Data Science. В статье рассмотрены базовые подходы Data Science – модели управляемого обучения и модели неуправляемого обучения, их основные алгоритмы, показана их практическая реализация в моделировании бизнес-процессов.
Ключевые слова: Data Science, алгоритм, модель, бизнес-процесс, решение.

Hnot T.V., Nehrey M.V. DATA SCIENCE ALGORITHMS IN BUSINESS PROCESSES MODELING
Successful business involves making decisions under uncertainty using a lot of information. Modern modeling approaches based on Data Science algorithms are a necessity for the effective management of business processes. Data science involves principles, processes, and techniques for understanding business processes through the analysis of data. The main goal of this article is to improve decision making using data science algorithms. In research, there are described basic approaches of Data Science – Supervised learning models and Unsupervised learning models, their main algorithms and their practical implementation in modeling business processes.
Keywords: Data Science, algorithm, model, business process, decision.

Постановка проблеми у загальному вигляді. Сучасне бізнес-середовище характеризується значною невизначеністю, зростанням конкуренції, глобалізацією. Для успішного функціонування бізнесу необхідно приймати рішення, враховуючи велику кількість факторів та значний обсяг інформації. Ефективність прийняття бізнес-рішень значною мірою залежить від уміння аналізувати наявну інформацію, прогнозувати розвиток бізнес-процесів та системного бачення усього бізнесу. Моделювання бізнес-процесів – це найбільш складна частина в їх аналізі. Вдосконалення процесу прийняття бізнес-рішень можливе за умови коректного застосування сучасних методів і моделей бізнес-аналізу, зокрема Data Science.

Аналіз останніх досліджень і публікацій. Питання застосування різних алгоритмів для моделювання бізнес-процесів вивчала значна кількість науковців. Серед вітчизняних дослідників, які використовували алгоритми Data Science для моделювання бізнес-процесів, варто зазначити праці Б. Павлишенка, А. Матвійчука, К. Ковальчука, В. Кравченка, В. Соловйова, П. Григорука та ін. Проте питанню вибору алгоритму для моделювання бізнес-процесів приділено недостатньо уваги.

Формулювання цілей статті (постановка завдання). Основним завданням статті є визначення сутності та особливостей застосування алгоритмів Data Science для моделювання бізнес-процесів з метою підвищення ефективності прийняття бізнес-рішень.

Виклад основного матеріалу дослідження. Загалом Data Science – це наука про отримання знань із даних. Data Science є продовженням Data Mining та Predictive Analytics. Даний підхід є міждисциплінарним, оскільки поєднує в собі методи та моделі таких дисциплін, як математика, статистика, теорія ймовірності, інформаційні технології, вклю-

чаючи обробку сигналів, імовірнісні моделі, машинне навчання, статистичне навчання, інтелектуальний аналіз даних, бази даних, розпізнавання об'єктів, візуалізацію, моделювання невизначеності, сховищ даних, стиснення даних, комп'ютерне програмування і високопродуктивні обчислення.

Суть Data Science полягає у видобутку інформації на основі знань і навичок із різних сфер діяльності, необхідних для отримання знань. Склад подібного набору значною мірою залежить від сфери дослідження. Для фахівців у цьому напрямі досліджень – Data Scientist – розроблено узагальнені кваліфікаційні вимоги.

Data Science розвивається швидкими темпами. Велика кількість інформації, яка зростає з кожним роком, дає можливість будувати високоточні моделі, які спрощують і частково автоматизують процес прийняття рішення. Створюються моделі, у яких реалізуються ключові алгоритми Data Science для прийняття рішень у бізнесі [8].

Основними підходами у Data Science є моделі керованого навчання (Supervised learning models) та моделі некерованого навчання (Unsupervised learning models).

Кероване навчання. Кероване навчання є одним із методів машинного навчання, у ході якого модель навчається на основі розмічених даних. За допомогою керованого навчання можна розв'язувати два типи задач: регресію та класифікацію. Основна відмінність між ними полягає у типі змінної, яку прогнозують за допомогою відповідного алгоритму. За регресійного навчання – це неперервна змінна, за класифікації – категоріальна змінна. Для розв'язку цих задач розроблено велику кількість алгоритмів. Одні з найбільш поширених – це лінійна та логістична регресії, дерева рішень.

Лінійна регресія. Регресійний аналіз можна вважати основою статистичного дослідження. Цей підхід включає широкий спектр алгоритмів, що використовуються для прогнозування залежної змінної, використовуючи один або декілька факторів (незалежних змінних). Залежність між змінними виражається лінійною функцією:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n,$$

x_i – *i-ий* фактор (*i-та* характеристика, на основі якої будується прогноз);

b_i – *i-ий* параметр моделі, який виражає вплив фактору;

y – залежна змінна, для якої будується прогноз.

Перевагою застосування такого підходу до моделювання є простота та зрозумілість результатів, швидкість навчання та видача прогнозу. Недоліком – не завжди достатньо велика точність (оскільки у бізнес-процесах лінійна залежність між змінними спостерігається рідко).

Один із прикладів застосування лінійної регресії у моделюванні бізнес-процесів – побудова трендів для часових рядів, коли за незалежну змінну приймають часові значення або індекси значень (наприклад від 1 до n, де n – кількість елементів у часовому ряді). Тренд дає змогу спрогнозувати значення на

Рис. 1. Середньорічна температура повітря та ґрунту в Україні

наступний період. Так, наприклад, проведено дослідження реальних змін клімату на основі аналізу середньорічної температури ґрунту та повітря в різних областях за 1990–2011 рр. (рис. 1) [9]. Для моделювання бізнес-процесів аграрного сектору важливо аналізувати цю тенденцію, оскільки це дає змогу оцінити умови посіву культур, розробити заходи захисту рослин, спрогнозувати врожайність сільськогосподарських культур тощо.

Іншим прикладом застосування лінійної регресії є задача оптимізації цін. Розглянемо приклад оптимізації цін на кренделі на тиждень в одному з магазинів. Маючи історичні дані ціни і попиту, цю задачу можна розв'язати в декілька етапів:

1. Спрогнозувати попит на продукт за допомогою аналізу часового ряду на наступний період, ураховуючи сезонні особливості і приріст споживачів.

2. Оцінити лінійну функцію залежності попиту від ціни, попередньо обчисливши рівень попиту на наступний тиждень, на основі прогнозу. Крім того, можна додати такі залежні змінні, як акційний товар, наявність реклами в буклетах/на дисплеях, у магазині тощо (рис. 2). У складніших випадках ця залежність може виражатися й нелінійною функцією (наприклад, сигмоїдальною).

3. Урахувавши оцінену функцією попиту, оптимізувати квадратичну функцію доходу (рис. 3):

$$D = pu = p(a + bp) = bp^2 + ap,$$

де D – дохід, p – ціна, u – попит.

Логістична регресія використовується, коли потрібно прогнозувати вихід бінарної змінної, використовуючи датасет неперервних або категоріальних змінних. Ситуації, коли вихідна змінна має більше ніж два можливих значення, можна змоделювати підходом one-vs-all, коли будуються логістичні класифікатори для кожного можливого виходу, або ж one-vs-one, коли будуються логістичні класифікатори для кожної можливої комбінації категорій вихідної змінної.

Залежність між незалежними і залежною змінними в логістичній регресії залишається лінійною, єдиною відмінністю від лінійної регресії є сигмоідальна функція, яка перетворює лінійний результат в ймовірність приналежності до класу в межах [0; 1]:

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n)}},$$

де x_i – *i-ий* фактор (*i-та* характеристика, на основі якої будується прогноз);

b_i – *i-ий* параметр моделі, який виражає вплив фактору;

p – ймовірність приналежності до класу 1.

Переваги і недоліки логістичної регресії перекликаються з перевагами і недоліками лінійної регресі. Це – швидкість роботи алгоритму та ймовірнісна інтерпретація результатів, з одного боку, і невелика точність – з іншого.

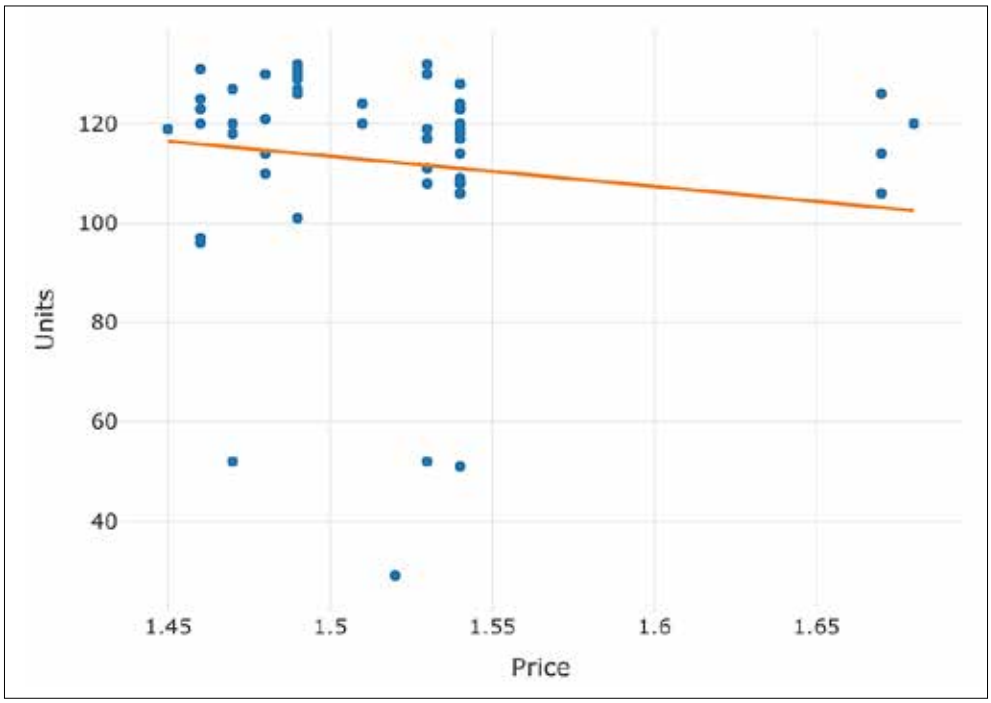

Рис. 2. Лінійна залежність між ціною і попитом на кренделі

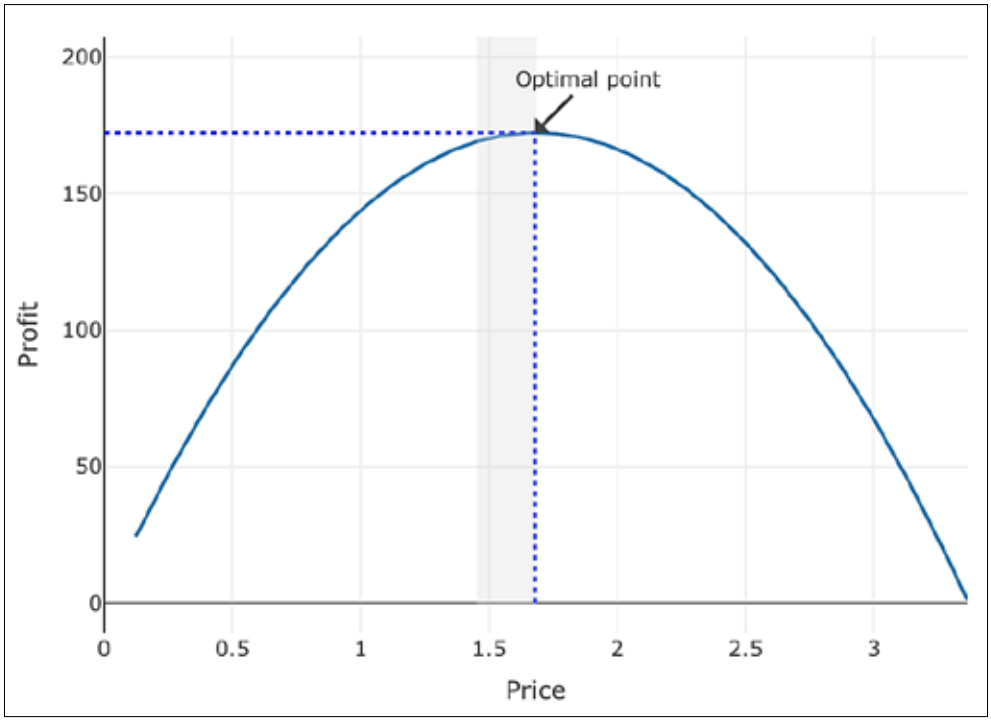

Рис. 3. Функція доходу на кренделі

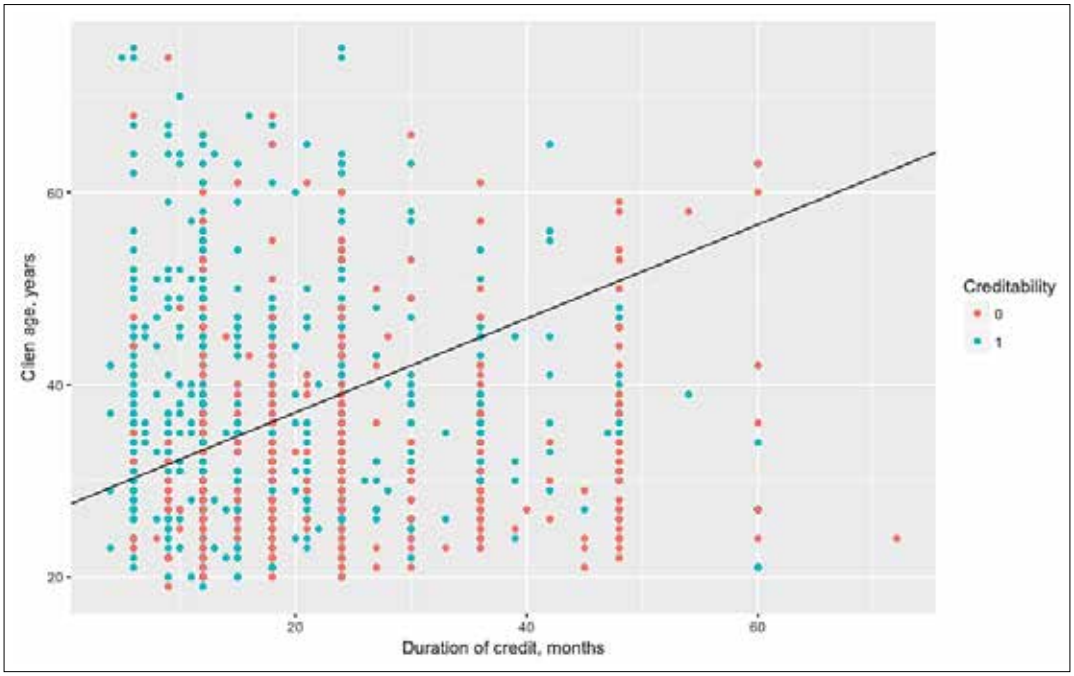

Рис. 4. Скорингова модель кредитоспроможності клієнтів

Логістична регресія часто використовується для побудови скорингових моделей. Важливим фактором при цьому є інтерпретабельність її результатів. Вплив кожного фактору чітко виражається величиною коефіцієнта *b*, що дає можливість чітко визначити, які з них позитивно і якою мірою впливають на прийняття рішення. На рис. 4 зображено просту скорингову модель, яка прогнозує кредитоспроможність клієнтів на основі двох факторів: вік клієнтів і термін кредиту. Дана модель побудована на основі 1000 кейсів із German Credit Risk датасету. Як видно з рисунку, модель прогнозує вищу кредитоспроможність клієнтам із терміном кредиту до двох років і віком 30–40 років. Точність такої моделі ~ 60%, за побудови логістичної регресії на всіх 20-ти атрибутах можна досягнути

точності 80%. Чорна лінія на графіку відображає границю рішень моделі: вище неї ймовірність позитивної відповіді > 50%.

Дерева рішень – ще один підхід як до регресії, так і до класифікації. Вони широко використовуються в інтелектуальному аналізі даних. Дерево рішень складається з «листів» і «гілок». На листях дерев знаходяться атрибути, які використовуються для прийняття рішень. Для того, щоб прийняти рішення, потрібно спуститися на низ дерева рішень. Послідовність атрибутів у дереві, а також значення, які розбивають листя на гілки, залежить від таких параметрів, як кількість інформації або ентропії, яку вносить атрибут у прогнозну змінну.

Перевагами дерев рішень є: простота інтерпретації, більша точність у моделюванні прийняття рішень порівняно з регресійними моделями, простота візуалізації, природне моделювання категоріальних змінних (у регресійних моделях їх необхідно кодувати штучними змінними). Проте дерева рішень мають один суттєвий недолік – досить низьку прогнозну точність [3].

Прикладом застосування дерева рішень є визначення алгоритму класифікації клієнтів компанії Walker – побудови Loyalty Matrix [7]. Усі клієнти при цьому поділяються на чотири групи (Truly Loyal, Accessible, Trapped, High risk) на основі відповіді на питання анкети від одного до п'яти. На рис. 5 зображено дерево, яке на основі трьох питань дає змогу з 98%-ю точністю спрогнозувати клас клієнта.

Некероване навчання. На противагу керованому навчанню некероване описує більш складну ситуацію, у якій для кожного спостереження $i = 1,\ldots,n$, є спостереження вектора вимірів x_i, але без ніяких змінних на виході y_i. На таких даних побудова моделей лінійної чи логістичної регресії є неможливою, оскільки відсутня змінна, яку прогнозують. У такій ситуації проводять так званий «сліпий» аналіз. Така задача належить до класу задач некерованого навчання через відсутність вихідної змінної, яка б керувала аналізом. Алгоритми некерованого навчання можна поділити на алгоритми зменшення простору й алгоритми кластеризації. Основна задача кластеризації полягає у пошуку патернів у даних, які дають змогу розбити дані на групи і потім певним чином проаналізувати їх і дати їм інтерпретацію.

K-середніх (k-means) – один із найпопулярніших алгоритмів кластеризації, основна задача якого полягає у розподіленні *n*-спостережень на *k*-кластерів. При цьому мінімізується сума квадратів відстаней кожного спостереження до центру відповідного кластеру. Цей алгоритм є ітераційним, на кожному кроці перераховуються центри кластерів і перерозподіляються спостереження між ними до того моменту, поки не досягнуто стабільного результату.

Перевагами такого алгоритму кластеризації є простота, швидкість і можливість опрацьовувати великі масиви даних. Недоліком – користувач повинен зазначити кількість кластерів, які він хоче використати для кластеризації, перед проведенням обчислень; нестабільність результату (він залежить від початкового розділення точок між кластерами).

На рис. 6 зображено приклад застосування k-середніх для кластеризації користувачів Інтернет-сервісу за координатами. Це дає змогу поділити їх на групи і сформувати зони доставки.

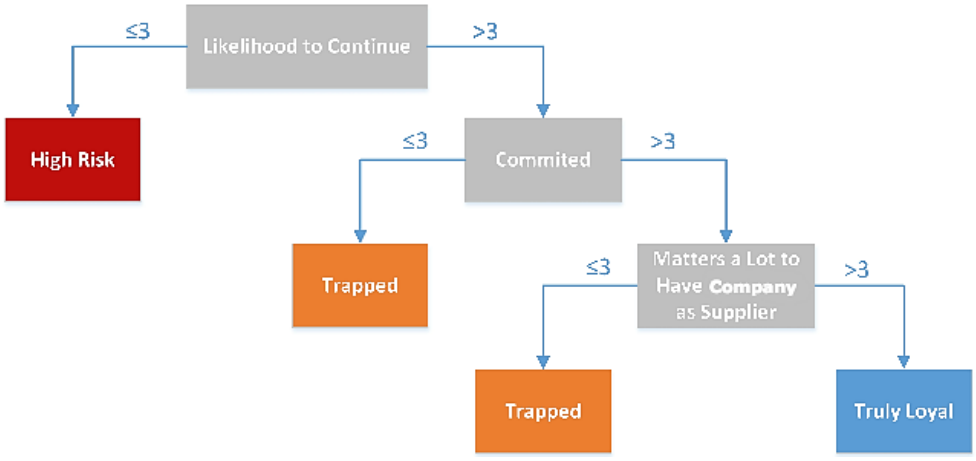

Рис. 5. Дерево рішень класифікації клієнтів компанії Walker

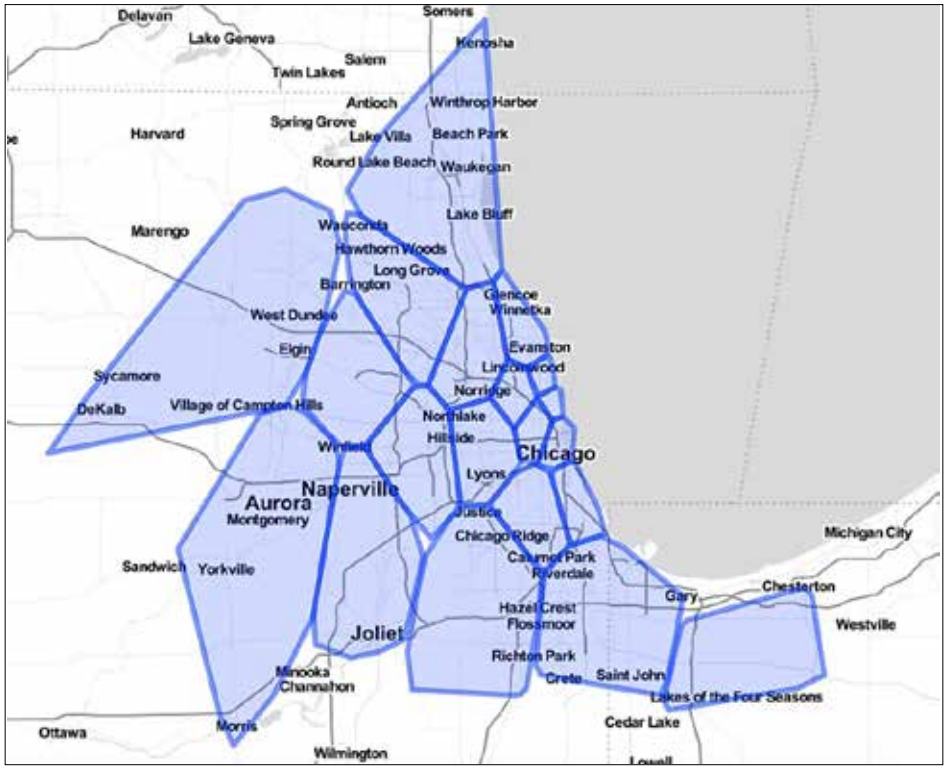

Рис. 6. Кластеризація Інтернет-клієнтів на основі їх координат

Ієрархічна кластеризація є альтернативним підходом до кластеризації, який не вимагає попереднього визначення кількості кластерів. Більше того, ієрархічна кластеризація гарантує стабільність результату і дає на виході привабливу візуалізацію, яка базується на деревовидній структурі спостережень/кластерів – дендрограма. Даний алгоритм кластеризації використовує різні метрики відстані і критерії агломеративного об'єднання кластерів, що робить його дуже гнучким до даних, на яких проводиться кластеризація. Проте недоліком ієрархічної кластеризації є необхідність обчислення матриці відстаней між спостереженнями перед проведенням агломерації, що ускладнює застосовування даного алгоритму для великих даних і даних із багатьма вимірами.

На рис. 7 наведено дендрограму сегментації клієнтів за такими ознаками, як кількість транзакцій у вихідні/будні, середня кількість покупок за тиждень тощо. Сегментація дає змогу виділити групи «подібних» клієнтів, наприклад ті, що роблять покупки тільки у вихідні; ті, які купують переважно товари зі знижкою, тощо. Таке розбиття дає змогу поліпшити цільовий маркетинг.

Аналіз часових рядів. Часовий ряд формується спостереженнями, які були зібрані з фіксованим інтервалом. Це може бути щоденний попит або ж щомісячні показники приросту прибутку, рівень інфляції тощо. Аналіз часових рядів посідає важливе місце в аналізі даних, який покриває області, починаючи з аналізу валютних курсів і закінчуючи прогнозуванням продажів [4]. Одна із задач аналізу часових рядів полягає у виділенні трендової і сезонних компонент і побудові прогнозу. Для цього було розроблено велику кількість алгоритмів, розглянемо такі моделі, як ARIMA і Prophet.

ARIMA. Алгоритм ARIMA є одним із найпоширеніших алгоритмів прогнозування часових рядів. Основна ідея полягає у використанні попередніх значень часового ряду для прогнозування майбутніх. При цьому може використовуватися будь-яка кількість лагів, що робить такий підхід складним у налаштуванні, оскільки потрібно так підібрати параметр, щоб водночас мінімізувати помилку і не перенавчити модель. ARIMA часто використовується для короткострокового прогнозування. Недоліком є складність навчання моделі в умовах багатьох сезонностей.

На рис. 8 зображено приклад прогнозування на один тиждень кількості замовлень у ресторані [5]. Можна чітко побачити сезонність в один день, яка притаманна рядам такого роду.

Prophet – це алгоритм, який розроблений компанією Facebook на початку 2017 р. для

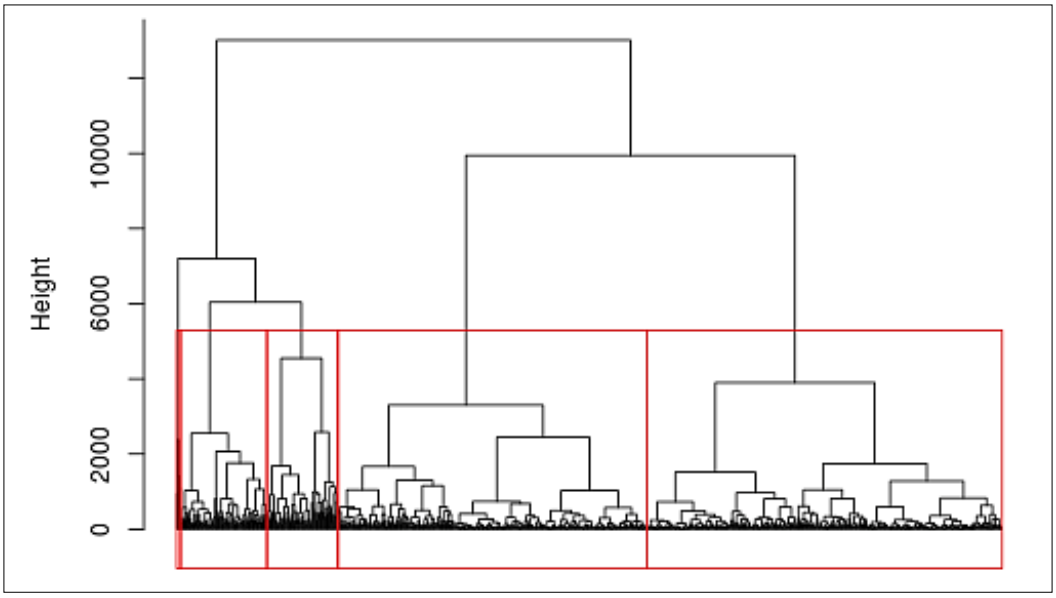

Рис. 7. Дендрограма сегментації клієнтів

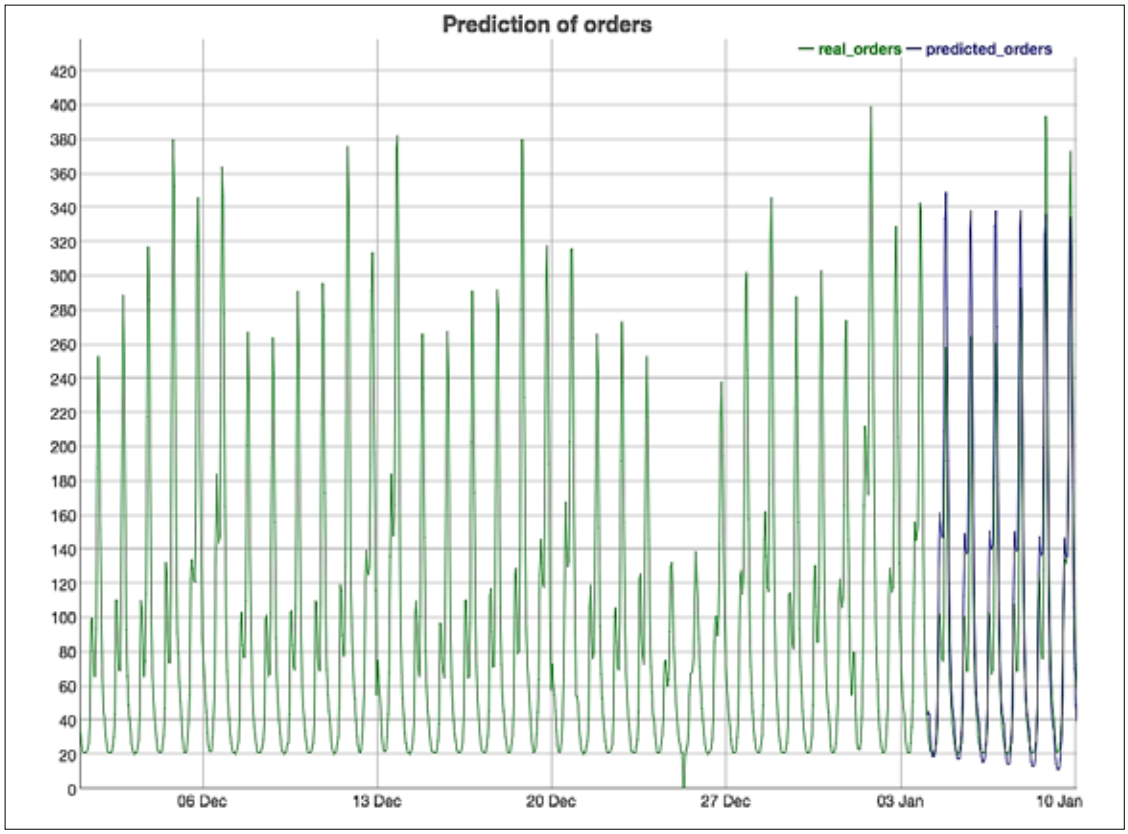

Рис. 8. Прогнозування кількості замовлень на основі ARIMA

прогнозування на основі часових рядів [6]. Він базується на адитивній моделі, в якій нелінійні тренди знаходяться з річною і тижневою сезонністю. Даний підхід також дає змогу змоделювати святкові дні і вихідні, тим самим даючи можливість прогнозувати викиди в ряді. Також Prophet є нечутливим до пропущених значень, зміщень у тренді і значних викидів, що є важливою його перевагою над ARIMA. Ще однією перевагою є досить велика швидкість навчання, а також можливість використання великих за обсягом часових рядів.

На рис. 9 наведено приклад прогнозування з допомогою Prophet. На першому з графіків –

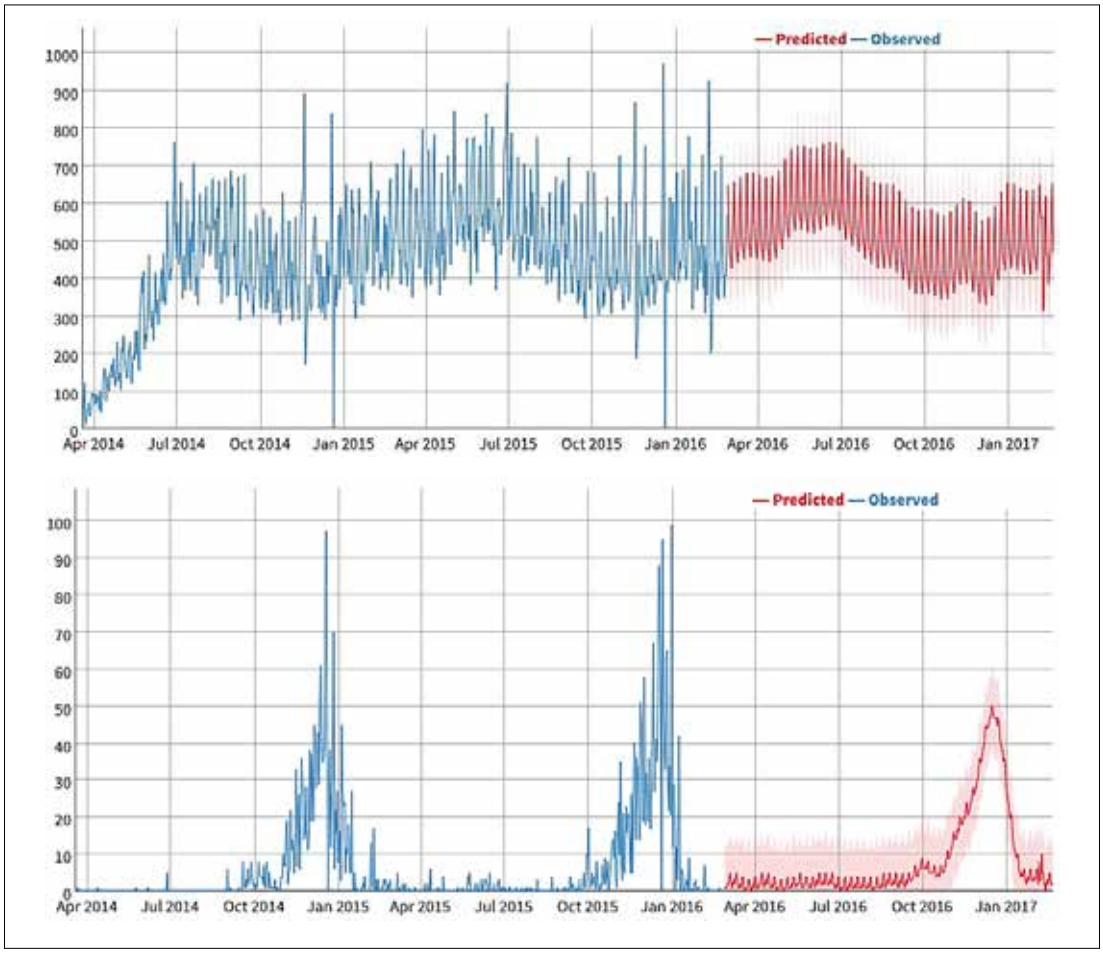

Рис. 9. Моделювання часових рядів за допомогою Prophet

прогнозування цілої категорії товарів, на другому – тих товарів, які купують на Різдво. При цьому в другому випадку враховуються тільки сезонні компоненти і не моделюється holiday-складник.

Хоча всі вищезазначені алгоритми не становлять повного переліку алгоритмів Data Science, проте, на нашу думку, вони становлять базу, яка необхідна для моделювання бізнес-процесів. Перелік інструментів Data Science, які доцільно використовувати в моделюванні бізнес-процесів, можна продовжити такими алгоритмами: ANOVA, нейронні мережі, метод головних компонент, факторний аналіз та ін.

Висновки з цього дослідження. Для забезпечення успішного розвитку бізнесу необхідно приймати рішення, використовуючи сучасні підходи до бізнес-аналітики – методи Data Science. Застосування алгоритмів Data Science дає можливість глибоко проаналізувати та зрозуміти бізнес-процеси, сприяє структуризації проблем, забезпечує систематизацію бізнес-процесів. Моделювання бізнес-процесів, в основі якого покладено алгоритми Data Science, дає можливість обґрунтовувати рішення і навіть автоматизувати процеси прийняття бізнес-рішень.

ЛІТЕРАТУРА:

1. Chen H., Chiang R.H., Storey V. C. Business intelligence and analytics: From big data to big impact. MIS quarterly.– 2012. – № 36(4).

2. Davenport T.H. Data scientist: the sexiest job of the 21st century / Davenport T.H., and Patil D.J. – Harv Bus Rev, – Oct. 2012.

3. James G. An Introduction to Statistical Learning /Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. – New York. : Springer, 2014. – 440 p.

4. Pavlyshenko B.M. Linear, machine learning and probabilistic approaches for time series analysis / Pavlyshenko Bohdan M. – In: Data Stream Mining & Processing (DSMP), IEEE First International Conference on. IEEE, 2016. – P. 377–381.

5. RPubs research code: http://rpubs.com/tarashnot/orders_full.

6. Taylor S.J. Forecasting at Scale / Sean J. Taylor, Benjamin Letham. – Menlo Park. : PeerJ Preprints., 2017 – 25 p.

7. Walker Loyalty Matrix: https://www.walkerinfo.com/docs/WP-The-Walker-Loyalty-Matrix.pdf.

8. Гнот Т.В. DATA SCIENCE в аналізі проблем природокористування / Т.В. Гнот, М.В. Негрей // Збірник матеріалів III Міжнародної науково-практичної конференції «Глобальні та регіональні проблеми інформатизації в суспільстві та природокористуванні 2015» (Київ, 25–26 червня 2015 р.). – К. : Інтерсервіс, 2015. – С. 35–36.

9. Негрей М.В. Зміна клімату: ризики і можливості для сільського господарства України / М.В. Негрей // Збірник матеріалів IV Міжнародної науково-практичної конференції «Глобальні та регіональні проблеми інформатизації в суспільстві та природокористуванні 2016» (Київ, 25–26 червня 2016 р.). – К. : Інтерсервіс, 2016. – С. 16–18.