Фостер Провост і Том Фоусетт

DATA SCIENCE ДЛЯ БІЗНЕСУ

Як збирати, аналізувати і використовувати дані

2-ге видання

Переклала з англійської
Анастасія Дудченко

«Наш формат» · Київ · 2020

УДК 004.6:33](0.062)
П78

Провост Фостер, Фоусетт Том

П78 Data Science для бізнесу. Як збирати, аналізувати і використовувати дані / пер. з англ. Анастасія Дудченко. — 2-ге вид. — К. : Наш формат, 2020. — 400 с.
ISBN 978-617-7730-03-2 (паперове видання)
ISBN 978-617-7730-04-9 (електронне видання)

Протягом останніх років не лише технологічні гіганти, а й інші компанії навчилися збирати дані про операційну роботу, результати маркетингових кампаній і поведінку своїх клієнтів. Проте не всі вміють застосовувати їх на користь власній справі. У цій книжці експерти Фостер Провост і Том Фоусетт пояснюють, як оцінити роль даних у вашому бізнесі, як їх трактувати й узагальнювати та якими принципами керуватися, щоб використати зібрану інформацію для розвитку вашого бізнесу.

УДК 004.6:33](0.062)

Перекладено за виданням: Foster Provost, Tom Fawcett. *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking* (Sebastopol, O'Reilly, 2013, ISBN 978-1-449-36132-7).

Головна редакторка *Ольга Дубчак.* Літературна редакторка *Анна Весній.* Коректорка *Алла Кравченко*. Верстальниця *Наталія Коваль*. Технічна редакторка *Ірина Щепіна*. Обкладинку адаптувала *Оксана Гаджій*. Випускова редакторка *Вікторія Шелест*. Художня редакторка *Оксана Гаджій*. Відповідальна за випуск *Ілона Замоцна*.

Дякуємо за допомогу в підготовці видання *Владиславові Шовковому*.

Надруковано в Україні видавництвом «Наш формат» у ТОВ «Конві Прінт», вул. Антона Цедіка, 12, м. Київ, 03680. Свідоцтво ДК № 6115 від 29.03.2018. Замовлення № 001066. Підписано до друку 28.12.2019. Тираж 1000 прим. Термін придатності необмежений. ТОВ «НФ», пров. Алли Горської, 5, м. Київ, Україна, 01032, тел. (044) 222-53-49, pub@nashformat.ua. Свідоцтво ДК № 4722 від 19.05.2014. Висновок Держ. сан.-епідем. експертизи № 05.03.02.-04/51017 від 16.11.2015.

Науково-популярне видання

ISBN 978-617-7730-03-2 (паперове видання)
ISBN 978-617-7730-04-9 (електронне видання)

ЗМІСТ

Присвячується нашим татам

ПЕРЕДМОВА

«Data Science для бізнесу» призначена для декількох типів читачів:

- Людей бізнесу, які працюватимуть зі спеціалістами з аналізу даних, управлятимуть проектами, які орієнтовані на дані, або робитимуть венчурні інвестиції в data science компанії;
- Розробників, які будуть впроваджувати пов'язані з data science рішення в життя, і
- Людей, які хочуть у майбутньому професійно працювати з даними.

Це не книжка про алгоритми, і вона не може замінити книжку про алгоритми. Ми свідомо вирішили не зосереджуватись на алгоритмах. Ми вважаємо, що в основі того, як діставати корисну інформацію з даних, лежить відносно невелика кількість фундаментальних концепцій або принципів. Ці концепції і стали *основою* для багатьох відомих алгоритмів добування даних. Навіть більше: ці концепції лежать в основі аналізу датоцентричних управлінських проблем, створення та оцінки рішень із використанням data science і оцінки загальних стратегій та пропозицій data science. Відповідно й ми організували цю книжку довкола цих загальних принципів, а не конкретних алгоритмів. Там, де обов'язково потрібно пояснювати деталі з процедури, ми поєднували текст і діаграми — вважаємо, що так зрозуміліше, ніж просто крок за кроком прописати алгоритм.

Серйозний математичний бекґраунд для цієї книжки не потрібен. Але матеріал сам по собі дещо технічний — у нас було на меті розповісти найважливіше з того, що потрібно знати про data science, а не просто наговорити великих незрозумілих загальних слів. Загалом, ми постаралися мінімізувати математику й пояснити все якомога більш «схематично».

Колеги з галузі кажуть, що книжка просто безцінна для тих, кому треба в бізнесі постійно працювати і зі стандартними процесами, і з технічно-

виробничою стороною, і з людьми, що займаються data science. Але група людей, про яку вони говорять, насправді дуже маленька — і нам було цікаво, скільком людям це насправді може бути важливо (див. розділ 5!). Ідеальну картину ми бачимо так: коли дата-науковцю потрібно почати працювати з новою командою розробників чи менеджерів, він дає їм цю книжку і каже: якщо ви справді хочете придумати/застосувати в бізнесі data science рішення екстра-класу, нам усім потрібно знати з цієї теми одне й те саме.

Колеги також кажуть, що книжка згодилася їм так, як ми того не очікували: щоб готуватися до співбесід із кандидатами на посади з data science. У бізнесі дуже сильний попит на спеціалістів з data science, і він щоразу зростає. Як результат, дедалі більше людей, які шукають роботу, починають називати себе дата-спеціалістами. Кожен кандидат на посаду такого спеціаліста має розуміти фундаментальні засади професії, які описані в цій книжці. (Колеги з галузі розповідають, що вони дуже здивовані тим, скільки людей насправді цих засад не розуміють. Ми напівсерйозно пишемо про це в «Записках Кліффа до співбесіди з кандидатом по data science»).

Наш концептуальний підхід до data science

У цій книжці ми зібрали колекцію найфундаментальніших принципів data science. Деякі з них винесено в заголовки розділів, про інші просто згадано у тексті (і тому вони не обов'язково виокремлені як фундаментальні принципи). Цими принципами описаний весь процес, від передбачення того, яка виникне проблема, до застосування технік data science і використання результатів на те, щоб обрати кращий із варіантів рішення. Ті самі принципи лежать в основі великої кількості аналітичних методів і технік, які застосовуються в бізнесі.

Ці принципи можна розбити на три великі групи:

1. Принципи, які стосуються того, яке взагалі у data science місце в компаніях і конкуренції, включно з тим, яким способом можна зацікавлювати й структурувати команди по data science і робити з них успішну частину компанії, яким способом перетворити data science на конкурентну перевагу і які можна використовувати тактики, щоб добре давати собі раду із проектами, пов'язаними з data science.
2. Загальні принципи аналітичного мислення з використанням даних. Вони допоможуть визначити прийнятні дані й застосувати відповідні методи. Серед цих принципів — *процес майнингу даних*, а також підбірка різних *завдань на майнинг даних високого рівня*.
3. Загальні принципи того, яким чином отримувати із даних потрібну інформацію — те, для чого власне й існує велика кількість завдань з data science та алгоритми їхнього виконання.

Наприклад, один із фундаментальних принципів — визначити, чи схожі два описаних даними суб'єкти. Із такою властивістю одразу з'являється чимало різноманітних завдань. Це можна використовувати напряму для *пошуку* клієнтів, схожих на заданого клієнта. Можна утворити основу для кількох алгоритмів *передбачення*, які будуть приблизно вираховувати потрібне значення — наприклад, який очікується коефіцієнт використання ресурсу для конкретного клієнта, або яка вірогідність того, що клієнт зреагує на пропозицію.

Можна також зробити це основою для техніки *групування*, за якої суб'єкти групуються за спільними ознаками, але без конкретної мети. На подібності засновані алгоритми *пошуку й видачі інформації*, за якими формується список релевантних сторінок у відповідь на пошуковий запит. І нарешті, вона лежить в основі кількох основних алгоритмів *рекомендацій*. У звичайній книжці, де багато говорили би про алгоритми, кожному з цих завдань приділили б окремий розділ, і описували би загальні принципи із купою деталей щодо самого алгоритму або математичних формул. Але в цій книжці ми більше зосереджені на тих принципах, які все це об'єднують, а конкретні завдання й алгоритми тут — просто природна демонстрація роботи цих принципів.

Якщо вже оцінювати те, наскільки широко можна застосувати окремий принцип, то як інший приклад можна навести поняття *підйому* — принцип, поширений куди більше, ніж очікувалось — який постійно дає про себе знати в data science. Його використовують, щоб оцінювати різноманітні види патернів у різних контекстах. Алгоритми таргетингу реклами оцінюються так, що рахується підйом, який реклама отримує при таргетингу на конкретну аудиторію. Підйом використовують, щоб оцінити, наскільки аргумент за якесь рішення чи проти нього сильний. Підйом допомагає визначити, коли спільна поява (асоціація) даних цікава, а коли це просто звичайний збіг через популярність в аудиторії.

Ми вважаємо, що коли такі фундаментальні принципи data science пояснювати — це не тільки допоможе читачеві, а й налагодить контакт між акціонерами бізнесу і дата-спеціалістами. Так обидві сторони говоритимуть однією мовою і краще розумітимуть одна одну. Якщо всі будуть розуміти ці принципи — то стане можливою тісніша співпраця, і можуть виплисти критично важливі питання, які інакше ніколи б не були порушені.

Для інструктора

Цю книжку успішно використовували як підручник на дуже багатьох курсах з data science. Взагалі, книжка з'явилася ще восени 2005 року, коли почали розвиватись мультидисциплінарні лекції Фостера з data science

у Школі бізнесу Леонарда Штерна*. На початку цей курс номінально був для студентів MBA і магістрів наук у галузі інформаційних систем, але туди тяглися й студенти з інших факультетів. Найцікавіше про цей курс було не те, що він подобався студентам MBA та інфо-системникам, для яких був створений, а те, що студенти, які добре зналися на комп'ютерному навчанні та інших технічних дисциплінах, теж вважали його дуже цінним. Схоже було, що причина тут, зокрема, в тому, що у їхньому розкладі не було предметів, де викладалися би фундаментальні принципи та інші теми, що не стосуються алгоритмів.

Тепер у Школі Штерна ми використовуємо цю книжку в багатьох пов'язаних із даними навчальних програмах: для безпосередньо студентів MBA та інфо-системників, для бакалаврів бізнес-аналітики, для нового набору магістрів бізнес-аналітики і на курсі «Вступ до data science» для нового набору магістрів data science. Окрім усього, (ще до публікації) книжку адаптували у більш ніж двадцяти інших університетах для навчання в дев'яти країнах (і це ще не кінець), у бізнес-школах, на курсах з програмування, і курсах, де дають більш загальну інформацію з data science.

Щоб отримати корисний інструкторський матеріал — слайди для лекцій, приклади питань і задач для домашніх завдань, інструкції з тестового проекту на основі фреймворків із книжки, питань до іспитів та іншого — стежте за оновленнями на сайті книжки.

Ми ведемо список відомих нам адаптацій на сайті книжки (www.data-science-for-biz.com). Клікайте на Who's Using It угорі.

Інші вміння та принципи

Окрім базових принципів data science, дата-спеціалісту знадобиться багато інших принципів і вмінь. Ці принципи та вміння ми обговоримо в розділах 1 і 2. Зацікавленого читача будемо раді бачити на сайті книжки, де він знайде список матеріалів, за якими ці додаткові вміння можна отримати (наприклад, виконання скриптів на мові Python, обробка командного рядка на Unix, датафайли, поширені формати даних, бази даних і постановка запитів, архітектури й системи біґ-дати, наприклад, MapReduce і Hadoop, візуалізація даних та інші теми, які зачіпають і нашу).

* Ясна річ, кожен автор беззаперечно вважає, що доклав до роботи над книжкою куди більше зусиль, аніж інший.

Секції та умовні позначення

Окрім виносок, у книжці будуть «бокові панелі». По суті, це будуть розширені примітки. Туди ми винесли матеріал, який вважаємо цікавим і вартим уваги, але він завеликий для виноски, і надто сильно відходить від основної теми.

Попереду технічні деталі — Відмітка там, де позначено «зірочкою»
Інколи траплятимуться математичні деталі — вони будуть винесені під «зірочку». Зірочка стоятиме у назвах параграфів, і перед ними завжди буде йти параграф, оформлений, як ось цей. В абзацах під «зірочкою» буде більше математичних і/або технічних деталей, ніж в інших частинах книжки, а в цих вступних параграфах буде пояснюватись, для чого вони потрібні. Книжку написано так, що ці розділи можна пропускати, й ви не перестанете далі розуміти, про що йдеться, але ми декілька разів ще нагадаємо читачеві, що в таких розділах можна порозбиратися з деталями.

Конструкції такого типу: (Сміт і Джонс, 2003) позначають, що ми посилаємось на бібліографію (у цьому випадку, на книжку або статтю, яку в 2003 році написали Сміт і Джонс). «Сміт і Джонс (2003)» означатиме те саме. Повна бібліографія є наприкінці книжки.

У цій книжці ми постаралися звести математику до мінімуму, а ту, що є, — максимально спростили, щоб ніхто не плутався. Для тих читачів, які мають технічний бекґраунд, можуть стати в пригоді деякі коментарі щодо цих спрощень:

1. Ми не використовували позначення сіґми (Σ) та пі ($\prod$), які зазвичай використовують у підручниках на позначення сум і добутків. Натомість ми просто писали рівняння із трикрапками, як ось тут:

   $$f(x) = w_1x_1 + w_2x_2 + \cdots + w_nx_n$$

   У технічних виносках під «зірочкою» ми інколи використовуємо знаки сіґми й пі, коли використовувати трикрапки надто незручно. Ми припускаємо, що люди, які читають ці частини, більш підковані технічно і математичні позначки їх не лякають.

2. Книжки зі статистики зазвичай дуже чітко позначають різницю між собівартістю та її приблизним значенням за допомогою «шапочки» на тих змінних, що позначають приблизні значення. У таких книжках справжня вірогідність зазвичай позначається p, а її приблизне значення — $\hat{p}$. У цій книжці мова майже завжди йтиме про приблизні

значення даних, а якщо всюди розставляти шапочки, то рівняння стають надто багатослівні й незугарні. Усе потрібно вважати за приблизне значення, якщо не буде вказано інше.

3. Ми спрощуємо позначення і прибираємо сторонні змінні, якщо вважаємо, що сенс від цього не зміниться. Наприклад, коли ми обговорюватимемо класифікатори з математичного погляду, технічно ми будемо мати справу із предикатами рішень за векторами ознак. Формально це треба буде виразити ось таким рівнянням:

$$\hat{f}_R(\mathbf{x}) = x_{\text{Age}} \times -1 + 0{,}7 \times x_{\text{Balance}} + 60$$

Натомість ми вибираємо більш читабельний спосіб:

$$f(\text{x}) = \mathit{Age} \times -1 + 0{,}7 \times \text{Balance} + 60,$$

розуміючи при цьому, що x — це вектор, а *Age* і *Balance* — його складові.

Ми постаралися всюди оформлювали все однаково, а шрифти зі сталою шириною (наприклад, `sepal_width`) лишити для того, щоб позначати атрибути або ключові слова в даних. Наприклад, у розділі про майнинг тексту такі слова, як «*обговорення*», позначатимуть слово в документі, а обговорити може бути результуючим пакетом даних.
У цій книжці використано такі типографічні норми:

Курсив
: Позначає нові терміни, посилання, електронні адреси, назви файлів і розширення файлів.

`Стала ширина`
: Використовується для описів програм, а також у тексті на позначення програмних елементів: назв змінних або функцій, баз даних, типів даних, змінних оточень, виражень та ключових слів.

`Стала ширина з курсивом`
: Показує, що цей текст можна замінити значеннями користувача або значеннями, яких вимагає контекст.

Ми також розставили у книжці спеціальні підказки та попередження щодо матеріалу. Вони будуть по-різному оформлені в паперових виданнях, PDF і електронних книгах, але матимуть такий вигляд:

Позначені так речення і параграфи — це підказки або пропозиції.

Цей текст і картинка означають, що тут стоїть загальна примітка.

Позначений ось так текст — це попередження. Інформація в цих примітках важливіша, ніж прості підказки, які використовуються лише інколи.

Використання прикладів

Окрім того, що ця книжка — вступ до data science, її ще можна використовувати в обговореннях цієї теми і в щоденній роботі з нею. Вам не потрібен особливий дозвіл, щоб відповісти на питання, процитувавши цю книжку і приклади з неї.

Якщо ви поставите посилання на нас, ми будемо раді, але це не обов'язково. Формальне посилання містить назву, автора, видавництво та ISBN. Наприклад: «"Data Science для бізнесу", Фостер Провост, Том Фоусетт (О'Рейлі). 2013, Фостер Провост і Том Фоусетт, 978–1–449–36132–7».

Якщо, за вашими відчуттями, ви використовуєте приклади так, що це вже стає до нас несправедливо і не підпадає під описане вище, звертайтеся до нас за адресою permissions@oreilly.com.

Як із нами зв'язатися

Будь ласка, надсилайте коментарі й питання стосовно книжки видавцеві:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800–998–9938 (in the United States or Canada)
707–829–0515 (international or local)
707–829–0104 (fax)

У нас є дві інтернет-сторінки цієї книжки, де ми публікуємо список друкарських помилок, приклади та іншу додаткову інформацію. Адреса сторінки — *www.oreil.ly/data-science і сторінки автора www.data-science-for-biz.com.*

Технічні питання й коментарі щодо книжки надсилайте на *bookquestions@oreilly.com*.

Більше інформації про книжки, курси, конференції і новини «О'Рейлі Медіа» шукайте на сайті *www.oreilly.com*.

Шукайте нас на Facebook: *www.facebook.com/oreilly*

Стежте за нами в Twitter: *www.twitter.com/oreillymedia*

Дивіться нас на YouTube: *www.youtube.com/oreillymedia*

Подяки

Дякуємо всім-всім колегам та іншим, хто пропонував безцінні ідеї, фідбек, критику, пропозиції, хто підтримував нас, обговорював з нами книжку, читав чернетки. Ризикуємо когось забути, але особливо хочется подякувати: Паносу Адамополосу, Мануелю Арріазі, Джошу Аттенберґу, Солону Барокасу, Рону Бекерману, Джошу Блюменстоку, Огаду Бразілею, Аарону Бріку, Джесіці Кларк, Нітешу Чаулі, Пітеру Девіто, Васанту Дхару, Яну Еймке, Теосу Євгеніоу, Джастіну Ґапперу, Томеру Ґеві, Деніелу Ґіліку, Шаундрі Гілу, Нідхі Катурії, Ронні Когаві, Маріосу Кокодісу, Тому Лі, Філіппу Мареку, Девіду Мартенсу, Софі Могін, Лорену Муресу, Алану Мюрею, Ніку Нішимурі, Баладжі Падманабгану, Джейсону Пену, Клодії Перліч, Ґреґорі П'ятецкі-Шапіро, Тому Філіпсу, Кевіну Рейлі, Майтал Саар-Цечанській, Евану Садлеру, Ґаліту Шмуелі, Роджеру Стейну, Ніку Стріту, Кірілу Цемехману, Крейгу Вону, Крісу Волинські, Валлі Вону, Джофу Веббу, Деббі Ястер і Рону Женґу. Хотілося би також подякувати всьому студентству з курсів Фостера — Дата-майнинг для бізнес-аналітики, Практична data science, Вступ у data science, і Рісерч-семінару по data science. Питання і складнощі, які виникали під час використання попередньої версії книжки, прекрасно допомогли нам її покращити.

Дякуємо всім колегам, які роками вчили нас data science і того, як її викладати. Особлива подяка Майтал Саар-Цечанській і Клодії Перліч. Майтал милостиво поділилася з Фостером своїми записами для курсу по дата-майнингу, який вела багато років тому.

Приклад із деревом класифікацій із розділу 3 (особлива подяка за візуалізацію «тіл») — це переважно її ідея та її приклад; її ідеї й приклади лежали в основі візуалізації порівняння структуризації екземпляра простору за допомогою дерев та лінійних дискримінантних функцій у розділі 4, приклад під назвою «Чи відповість Девід» з розділу 6 оснований на її прикладі, і є, мабуть, ще щось, про що ми забули. Останні кілька років Клодія разом із Фостером викладала узагальнювальні секції на курсах Дата-майнинг для бізнес-аналітики/Вступ до data science, і в процесі навчила його багато чого про data science (і не тільки).

Дякуємо Девідові Стілвелу, Тору Ґрейпелу і Майклу Косинські, що надали для деяких прикладів дані про лайки у фейсбуці. Дякуємо Ніку Стріту, що дав дані по цитобласту і що дозволив використати картинки цитобласту в розділі 4. Дякуємо Девіду Мартенсу за допомогу з візуалізацією того, як мобільний телефон визначає місцезнаходження. Дякуємо Крісу Волинські, що надав дані стосовно своєї роботи для Netflix Challenge. Дякуємо Сонні Тамбе за те, що раніше надав нам доступ до результатів своєї роботи з технологій біґ-дати й продуктивності. Дякуємо Патрікові Перрі, що звернув нашу увагу на приклад із дзвінком у банк з розділу 12. Дякуємо Джеффу Веббу, що використав майнингову систему асоціації Magnum Opus.

Найбільша подяка — нашим родинам за їхню любов, терпець і підтримку.

Для підготовки цієї книжки й прикладів із неї було використано багато програмного забезпечення з відкритим кодом. Автори хочуть подякувати розробникам і співавторам програм:

- Python і Perl;
- Scipy, Numpy, Matplotlib, і Scikit-Learn;
- Weka;
- The Machine Learning Repository at the University of California at Irvine (Bache & Lichman, 2013).

І нарешті, ми будемо раді, якщо читачі зазирнуть на наш сайт (*www.data-science-for-biz.com*) і почитають оновлення до матеріалів, нові розділи, списки друкарських помилок, доповнення і слайди до всього цього.

Фостер Провост і Том Фоусетт

Розділ 1

ВСТУП: ДАТА-АНАЛІТИЧНЕ МИСЛЕННЯ

Не мрійте маленькими мріями; вони не здатні рухати серцями мужів.
Йоганн Вольфґанґ Ґете

За останні 15 років у інфраструктуру бізнесу прийшли величезні інвестиції, від чого можливостей збирати дані по всьому підприємству стало куди легше. Майже кожен аспект бізнесу тепер відкритий до збору даних і часто навіть під це підлаштований: операції, виробництво, управління логістикою, поведінка користувачів, ефективність маркетингової кампанії, організація робочого процесу тощо. Водночас інформація широко доступна і поза межами компаній: тренди ринку, новини галузі, дії конкурентів. Через загальну доступність даних зросла цікавість до методів, якими із даних можна дістати корисну інформацію та знання, — царини data science.

Всюдисутність можливостей даних

Тепер, коли стала доступною безмежна кількість даних, компанії в майже кожній галузі зосереджено намагаються експлуатувати її так, щоб отримати конкурентну перевагу. У минулому фірми наймали команди статистиків, розробників моделей і аналітиків, щоб вони вручну досліджували отримані дані, але тепер обсяги й різноманіття цих даних вже вийшло далеко за межі, де їх можна опрацювати вручну. Водночас комп'ютери стали набагато потужнішими, мережева кооперація — поширеним явищем, а алгоритми розвинулися так, що можуть поєднувати бази даних між собою, і так проводити ширший і глибший аналіз, аніж той, що взагалі був раніше можливий. Збіглися два феномени, і відтак у бізнесі почали дедалі активніше використовувати техніки дата-майнингу і принципи науки про дані.

Найширше, мабуть, техніки дата-майнингу використовують у маркетингу — для таргетування, онлайн-реклами і рекомендацій для крос-продажу. Дата-майнинг використовують для управління стосунками з умовним клієнтом, аналізу поведінки клієнта, щоб мати змогу контролювати тертя і зробити споживчу цінність максимально високою. У фінансовій галузі дата-майнингом користуються для того, щоб створювати кредитні рейтинги й торгувати в кредит, а також щоб визначати шахраїв і управляти персоналом. Великі ритейлери, наприклад, Walmart чи Amazon, використовують дата-майнинг у своєму бізнесі всюди: і в маркетингу, і в управлінні логістикою. Багато компаній статегічно диференціювалися за допомогою data science, деякі аж так сильно, що перетворилися на компанії з дата-майнингу.

Головна мета цієї книжки — допомогти вам побачити бізнес-проблеми з погляду даних і зрозуміти, за яким принципом із даних можна видобувати корисну інформацію. У дата-аналітичного мислення є фундаментальна структура і базові принципи, які потрібно розуміти. Подекуди тут також потрібно застосовувати інтуїцію, креативність, здоровий ґлузд і знання проблемної ділянки. Якщо ви дивитиметеся на проблеми з погляду даних, у вас буде структура і принципи, а відтак ви отримаєте шаблон, за яким можна буде систематично аналізувати подібні проблеми. Коли ви краще освоїте дата-аналітичне мислення, то вже почнете інтуїтивно розуміти, де тут можна придумати креативне рішення, а де потрібні знання проблемної ділянки.

У перших двох розділах цієї книжки ми детально обговорюватимемо різноманітні теми й техніки, які стосуються data science і дата-майнингу. Терміни «data science» і «дата-майнинг» часто вважають взаємозамінними, але коли різні компанії та індивідууми почали намагатися заробляти на дата-хайпі, перший термін зажив власним життям. Якщо не вдаватися в подробиці, data science — це набір принципів, якими потрібно керуватися, щоб з даних отримати інформацію. Дата-майнинг — це вилучення інформації з даних, за допомогою технологій, які створені за цими принципами. Сам термін «data science» використовують частіше, ніж «дата-майнинг», але техніки дата-майнингу — це нерідко найкращі ілюстрації принципів data science.

Розуміти data science важливо, навіть якщо ви взагалі не збираєтеся її застосовувати. Із дата-аналітичним мисленням ви зможете оцінювати пропозиції щодо проектів із дата-майнингу. Наприклад, якщо співробітник, консультант або потенційна інвестиційна мета запропонують краще застосування для конкретного бізнесу із використанням інформації, яку можуть дати дані, ви зможете системно оцінити пропозиції і вирішити, чи розважна вона, а чи в ній є недоліки. Це не означає, що ви знатимете, чи спрацює запропоноване — в проектах із дата-майнингом для цього найчастіше потрібно спробувати — але ви зможете помітити явні недоліки, нереалістичні припущення й частини, яких не вистачатиме.

У цій книжці ми опишемо низку фундаментальних принципів data science, і проілюструємо кожен із них щонайменше однією технікою дата-майнингу, в якій буде використано цей принцип. Зазвичай кожен із принципів застосовується в багатьох техніках, тож у цій книжці ми вирішили скоріше говорити про базові принципи, аніж про конкретні техніки. А отже, ми не будемо сильно наполягати на різниці між data science і дата-майнингом, якщо тільки це не буде напряму впливати на пояснення принципів.

Розгляньмо два невеликих кейси з аналізу даних, де потрібно отримати прогнозовані сценарії.

Приклад: ураган Френсіс

Розгляньмо приклад зі статті 2004 року в New York Times.

> Ураган Френсіс на повній швидкості мчав через Карибське море і загрожував ударити просто по атлантичному узбережжю Флориди. Мешканці узбережжя втекли туди, де було повище, але управління магазинів Wal-Mart у Бентонвілі вирішили, що в цій ситуації буде дуже доречно використати їхню найновішу зброю… технології передбачення.
>
> За тиждень до того, як ураган мав дістатися землі, керівниця інформаційного управління Wal-Mart Лінда Ділман наказала своїм працівникам створити прогнози погоди на основі того, що сталося, коли за кілька тижнів до того налетів ураган Чарлі. На основі трильйонів байтів історії покупок, які були у сховищі даних Wal-Mart, вона вирішила, що компанія може «почати передбачати, що станеться, а не чекати, коли це станеться», сказала вона. (Гейс, 2004)

Подумайте, *чому* передбачення на основі даних в такому сценарії може бути корисним. Воно може бути корисним, тому що так можна передбачити, що люди, які тікатимуть від урагану, будуть купувати більше води у пляшках. Можливо, але це трішки очевидно, та й хіба потрібні дані, щоб це зрозуміти? Воно може бути корисним, щоб спрогнозувати, *на скільки саме підніметься продажі* через ураган, щоб закупити в місцеві Wal-Mart потрібну кількість товару. Можливо, дата-майнинг покаже, що через ураган зростають продажі певних DVD — але можливо, вони того тижня розпродалися у Wal-Mart по всій країні, не тільки там, куди сунув ураган. Передбачення могло бути корисним, але, мабуть, загальнішим, аніж те, яке хотіла отримати міс Ділман.

Було би куди цінніше пошукати пов'язані з ураганом патерни, які були б не такі очевидні. Щоб це зробити, аналітики могли би оцінити величезний обсяг даних Wal-Mart із попередніх, подібних ситуацій (наприклад, коли був

ураган Чарлі), та ідентифікувати незвичайні товари, на які під час урагану в цьому районі піднімається попит. Із таких патернів компанія могла би визначити, на які *незвичні* товари зросте попит, і наповнити ними склади до того, як налетить ураган.

Насправді, так і сталося. У The New York Times (Гейс, 2004) написали: «…експерти промайнили дані і дізналися, що магазинам знадобляться запаси конкретних товарів — і не тільки ліхтариків, як можна було подумати. "Ми раніше не знали, що полуничні «поп-тартс» перед ураганом починають продаватися набагато краще, приблизно в сім разів краще, ніж зазвичай, — сказала міс Ділман у нещодавньому інтерв'ю. — А найкраще перед ураганом продавалося пиво"»*.

Приклад: передбачення плинності клієнтів

Як проводять такий аналіз даних? Уявімо на секунду більш типовий бізнес-сценарій і як із ним можна повестися, якщо дивитися на все з погляду даних. Ця проблема буде нам за приклад для багатьох питань, які порушуватимуться в цій книжці, своєрідним критерієм.

Припустімо, ви щойно отримали прекрасну роботу аналітика в MegaTelCo, одній із найбільших телекомунікаційних фірм у Сполучених Штатах. У них велика проблема: клієнти відмовляються від їхніх бездротових послуг. У середньоатлантичному регіоні 20 % клієнтів мобільних телефонів ідуть, коли у них закінчується контракт, а шукати нових стає дедалі важче. Оскільки ринок мобільних телефонів насичений, величезні темпи росту бездротового ринку впали. Комунікаційні компанії тепер борються за клієнтів одна одної і водночас намагаються втримати своїх. Коли клієнти переходять від компанії до компанії, це називається «плинністю», і це дорого, як не крути: одна компанія повинна вкладатися в засоби заохочення, щоб привабити клієнта, а інша втрачає прибуток, коли він іде.

Вас узяли на роботу, щоб ви розібралися з проблемою і допомогли придумати рішення. Приваблювати нових клієнтів набагато дорожче, ніж утримувати старих, тож велика частка маркетингового бюджету спрямована на те, щоб уникати плинності. Маркетинг-відділ уже придумав спеціальну пропозицію, щоб їх утримати. Ваше завдання — розробити конкретний покроковий план, які дата-спеціалістам використати обширні дата-ресурси MegaTelCo, щоб вирішити, яким клієнтам запропонувати особливі умови до того, як у них закінчаться контракти.

Добре подумайте, які ви можете використати дані та як саме. Зокрема, як MegaTelCo вибрати користувачів, які отримають пропозицію, щоб яко-

* Ну звісно! Що краще підходить до полуничного печива, ніж хороше холодне пиво?

мога більше скоротити плинність з урахуванням бюджету? На це питання відповісти куди складніше, ніж може здатися спочатку. У книжці ми неодноразово повертатимемося до цього завдання, і в силу того, як ми розумітимемо принципи data science, наше рішення ставатиме дедалі елегантнішим.

У реальності технології дата-майнингу переважно і використовували, коли розбиралися з відтоком клієнтів — особливо в телекомунікаційному та фінансовому бізнесах. Вони одними з найперших почали широко використовувати технології дата-майнингу, з причин, які ми обговорюватимемо пізніше.

Data science, програмування і ухвалення рішень на основі даних

Data science — це принципи, процеси й техніки, потрібні для розуміння феноменів через (автоматичний) аналіз даних. У цій книжці ми говоритимемо про прийняття рішень як про головну мету data science, оскільки зазвичай саме це напряму цікавить бізнес.

На рис. 1.1 data science поставлена в контекст різноманітних пов'язаних із даними та близьких до неї процесів у організації. За ним можна розі-

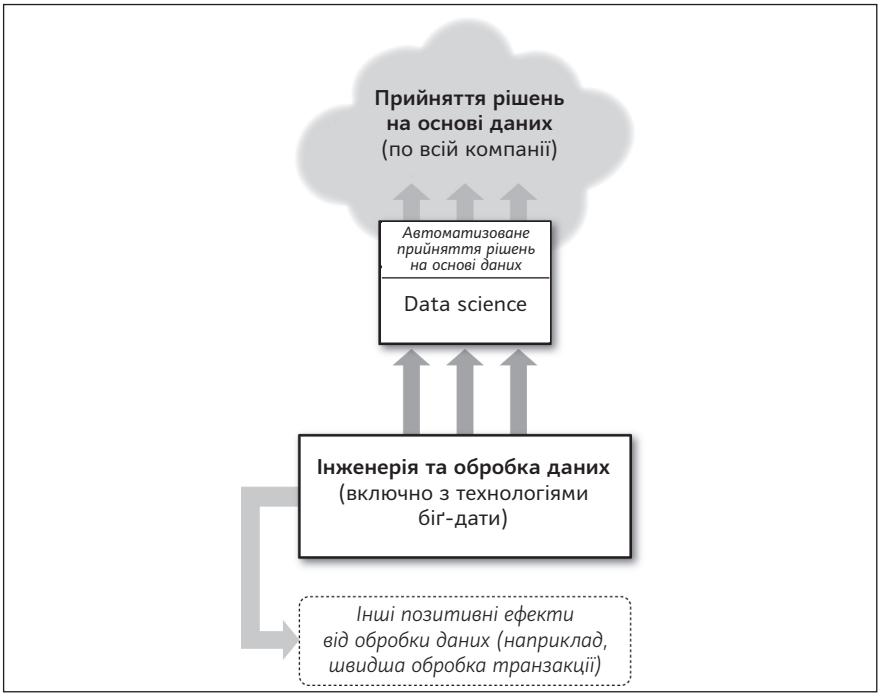

Рис. 1.1. Data science у контексті різних пов'язаних із даними процесів у організації

братися, чим data science відрізняється від обробки даних, до якої в бізнес-середовищі зараз зростає інтерес. Почнімо згори.

Ухвалювання рішення на основі даних — це основувати рішення на аналізі даних, а не лише на інтуїції. Наприклад, маркетолог може вибрати рекламу, засновуючись суто на своєму великому досвіді роботи в галузі та інтуїції до прийомів, які працюють. Або ж на аналізі даних про те, як споживачі реагують на різні реклами. Можна також використати обидва підходи. Ухвалення рішень на основі даних — не такий підхід, де або все, або нічого, і в різних компаніях до цього вдаються де більше, а де й менше.

Переваги ухвалення рішень на основі даних підтверджені беззаперечно. Економіст Ерік Бринолфссон та його колеги з MIT та Ліги Плюща дослідили, як прийняття рішень на основі даних впливає на продуктивність компаній (Brynjolfsson, Hitt, & Kim, 2011). Вони розробили оцінку такого підходу, який ранжував компанії за тим, наскільки активно вони використовують дані, коли приймають рішення в компанії. Дослідження показало, що за статистикою, що більше в компанії орієнтуються на дані, то продуктивніша вона — вона навіть може контролювати велику кількість розбіжних факторів. І різниця немаленька. Одне стандартне відхилення вгору по шкалі прийняття рішень на основі даних — це зростання продуктивності на 4–6 %. Ухвалення рішень на основі даних корелюється також із вищим прибутком на активи, рентабельністю капіталу, використанням наявних ресурсів і ринковою цінністю, і схоже, що ці фактори між собою пов'язані.

Рішення, які цікавитимуть нас у цій книжці, можна поділити на два типи: 1) рішення, для яких у межах даних потрібно робити «відкриття», 2) рішення, які повторюються, особливо коли масштаб величезний, і навіть якщо прийняття рішень на основі аналізу даних стане трішечки точнішим, це може позитивно вплинути на процес прийняття рішень загалом. Наведений вище приклад із Wal-Mart — ілюстрація проблеми першого типу: Лінді Ділман потрібна була інформація, як допомогти Walmart підготуватися до неминучого урагану Френсіс.

У 2012 році магазин Target, конкурент Walmart, потрапив у новини зі своєю історією про прийняття рішень на основі даних. Це теж була проблема першого типу (Duhigg, 2012). Target, як і переважна більшість роздрібних торговців, переймається купівельними звичками людей, тим, що їх підштовхує купувати і як на них можна вплинути. Зазвичай у покупців звички досить-таки усталені й змінити їх дуже складно. Однак люди, які приймали рішення в Target, знали, що один момент, коли купівельні звички суттєво змінюються, існує. Це момент, коли в людей з'являється дитина. «Якщо ми зробимо так, що вони купуватимуть у нас підгузки — вони купуватимуть у нас і все інше». Більшість роздрібних магазинів про це знають і конкурують

один з одним — намагаються продати новоявленим батькам дитячі товари. Доступ до більшості записів про народження можна отримати легко, тому магазини збирають інформацію про новонароджених дітей і надсилають батькам свої спеціальні пропозиції.

Однак Target хотів отримати перевагу в цій конкуренції. Їм стало цікаво, чи можуть вони передбачити, що люди чекають на дитину. Якби вони могли — то могли би отримати перевагу перед конкурентами, адже надсилали би свої пропозиції раніше, ніж це робили би вони. Target використали техніки data science і проаналізували історію даних про клієнтів, про яких пізніше стало відомо, що вони вагітні. І тепер вони могли отримати інформацію про те, що клієнти вагітні. Наприклад, вагітні жінки часто міняють раціон, гардероб, починають купувати інші вітаміни тощо. Ці індикатори можна було дістати з історії даних, зібрати в моделі передбачення й використати у маркетингових кампаніях. Далі в книжці ми дуже детально обговоримо прогностичні моделі. Поки досить розуміти ось що: у прогностичній моделі більшість складнощів світу відкидається, і натомість увага зосереджується на конкретному наборі індикаторів, які певним чином корелюють із величиною інтересу (хто з клієнтів піде, хто зробить покупку, хто вагітний і т.д.). Важливо, що і в прикладі з Walmart, і в прикладі з Target аналіз даних не обмежився звичайним тестуванням простих гіпотез. Натомість дані досліджували, сподіваючись відкрити щось корисне*.

Один із прикладів плинності демонструє проблему ухвалення рішень на основі даних другого типу. У MegaTelCo — сотні мільйонів клієнтів, і кожен із них — кандидат на виліт. Щомісяця в десятків мільйонів користувачів спливають терміни контрактів, тож ризик, що кожен із них у найближчому майбутньому піде, підвищений. Якби ми навчилися краще приблизно вираховувати вірогідність по кожному клієнтові — наскільки вигідно для нас зосередити зусилля саме на цій людині — то могли би надзвичайно плідно застосувати це вміння для мільйонів своїх користувачів. Ту саму логіку можна застосувати в багатьох галузях, де data science і дата-майнинг застосовуються найширше: прямий маркетинг, онлайн-реклама, рейтинги кредитоспроможності, торгівля на фінансових ринках, управління службою техпідтримки, виявлення шахраїв, пошукове ранжування, рекомендації продуктів і т.д.

Діаграма на малюнку 1.1 показує, як data science допомагає приймати рішення на основі даних, але також — як ці дві речі перетинаються. Тут варто звернути увагу на факт, який часто з уваги випускають: бізнес-рішення дедалі частіше ухвалюють автоматично, і роблять це комп'ютерні системи.

---

* У Target вийшло настільки добре, що навіть порушили питання, чи етично застосовувати такі техніки. Говорити про етику й приватність цікаво і дуже важливо, але цю розмову ми лишимо на інший час і місце.

У різних галузях автоматичне ухвалення рішень набуло різних масштабів. Галузі фінансів і телекомунікацій почали застосовувати це раніше, переважно тому, що мережі даних у них розвивалися не за роками, і вони проводили обрахування величезного масштабу. Вони могли об'єднувати дані й створювати великомасштабні моделі, а також застосовувати отримані моделі у прийнятті рішень.

У 1990-х автоматичне ухвалення рішень до невпізнаваного змінило банківські системи й споживчі кредити. Банки й телекомунікаційні компанії в 1990-х також застосовували системи великих масштабів, щоб мати змогу приймати рішення стосовно контролю за шахраями на основі даних. Роздрібні системи комп'ютеризувалися дедалі більше, рішення щодо мерчандайзингу вже були повністю автоматизовані. Відомі приклади — програми винагород у казино Harrah's і автоматичні рекомендації Amazon і Netflix. Зараз ми з вами спостерігаємо революцію в рекламі, оскільки тепер користувачі величезну частку часу проводять в інтернеті, і в онлайні рішення щодо реклами можна ухвалювати за (буквально) долю секунди.

Обробка даних та «біґ-дата»

Зараз важливо відхилитися й обговорити ще один момент. В обробці даних є багато такого, що не стосується data science — хоча, якщо судити по медіа, то можна подумати інакше. Інженерія й обробка даних критично важливі для існування data science, але це більш загальні поняття. Наприклад, сьогодні багато вмінь, систем і технологій з обробки даних помилково називають data science. Але щоб розуміти data science і компанії, які ухвалюють рішення на основі даних, важливо розуміти цю відмінність. Для data science потрібен доступ до даних, і правильна інженерія може тільки піти на користь, але ці технології — не технології власне data science. Як показано на рис. 1.1, вони підтримують data science, але можуть бути корисні ще багато для чого. Технології обробки даних дуже важливі для багатьох задач у бізнесі, для яких потрібні дані, але де не потрібно вміти діставати з даних корисну інформацію або ухвалювати на їх основі рішення. Це, наприклад, ефективна обробка транзакції, підтримування роботи сучасної веб-системи і рекламна онлайн-кампанія.

Останнім часом досить багато уваги медіа привернули технології біґ-дати (Hadoop, Hbase, і MongoDB). По суті, «біґ-дата» означає, що для традиційних технологій обробки даних ці набори даних завеликі, і потрібні нові технології. Технології біґ-дати використовують для багатьох задач, включно з інженерією даних — як і традиційні технології. Інколи технології біґ-дати застосовують, щоб імплементувати техніки майнингу даних. Однак набагато частіше так широко відомі технології біґ-дати використо-

вують для обробки даних на підтримку технік майнингу даних та інших пов'язаних із даними задач, як показано на рис. 1.1.

Раніше ми обговорювали дослідження Бринолфссона, яке показує, чим добре ухвалювати рішення на основі даних. В іншому дослідженні, яке зробив економіст Парсанна Тамбе зі Школи Штерна, перевірили, до якої міри технології біґ-дати насправді допомагають фірмам (Tambe, 2012). Він виявив, що після того, як інші можливі фактори, які могли би завадити продуктивності, взяті під контроль, використання біґ-дати суттєво підвищує рівень продуктивності в компанії. Точніше кажучи, стандартне відхилення, пов'язане з активнішим застосуванням біґ-дати в роботі середньої компанії, — зростання продуктивності на 1–3 %. Водночас стандарте відхилення, пов'язане з меншим застосуванням біґ-дати в роботі середньої компанії — зниження продуктивності на 1–3 %. За таких умов на екстремальних точках продуктивність компаній буде різнитися дуже сильно.

Від біґ-дати 1.0 до біґ-дати 2.0

Один зі способів зрозуміти, в якому зараз статусі технології біґ-дати — згадати, як компанії починали користуватися інтернет-технологіями. У часи Web 1.0 бізнеси активно займалися тим, що закупали базові інтернет-технології: їм потрібно було створити присутність в інтернеті, вибудувати процес електронної комерції і покращити ефективність операцій. Можемо уявити, що ми зараз живемо в еру Біґ-дата 1.0. Компанії активно набираються вмінь обробляти біґ-дату, переважно для того, щоб підтримувати ті операції, що вони вже проводять — наприклад, щоб робити їх ефективнішими.

Коли компанії вже добре освоїлися з технологіями Web 1.0 (а в процесі ціни на потрібні технології сильно впали), вони почали шукати далі. Вони почали запитувати, що мережа може для них зробити, як їм почати робити те, що вони роблять, краще — і ми ввійшли в еру Web 2.0. Нові системи й компанії почали користуватися перевагами інтерактивної природи мережі. Це змінило сам підхід, і зміни ці проникли всюди. Найочевидніший із прикладів — те, яку роль почали відігравати соціальні мережі і наскільки гучнішим став «голос» користувача (і простої людини).

Найімовірніше, після фази Біґ-дата 1.0 настане фаза 2.0. Коли фірми навчаться добре обробляти дані, вони почнуть питати: «Що я тепер можу такого, чого раніше не могла, або принаймні чи можу я щось тепер робити краще?». Найімовірніше, це буде золота доба data science. Принципи і техніки, про які йдеться в цій книжці, застосовуватимуться куди ширше й активніше, ніж зараз.

Важлива примітка: в еру Web 1.0 деякі особливо добре розвинені компанії почали застосовувати ідеї Web 2.0 задовго до того, як це почали робити всі інші. Прекрасний приклад — Amazon, який раніше за інших почав підключати «голос» клієнта: рейтинги продуктів, відгуки про продукти (і навіть більше — рейтинги відгуків про продукти). І так само зараз ми вже бачимо, як деякі компанії застосовують Біґ-дату 2.0. Amazon знову попереду всіх: вони дають рекомендації на основі величезної кількості даних. Є й інші приклади. Онлайн-рекламодавці повинні обробляти неймовірно величезну кількість даних (мільярди реакцій на рекламу на день — звичайна річ), і повинні дуже швидко відробляти отримувану інформацію (системи відкритих торгів у режимі реального часу приймають рішення за десятки мілісекунд). Потрібно стежити за цими та подібними галузями і шукати натяки на те, як саме інші галузі пізніше застосовуватимуть біґ-дату і data science.

Вміння працювати з даними і data science як стратегічне надбання

З усього сказаного вище можна вивести один із фундаментальних принципів data science: дані і здатність діставати з них корисну інформацію треба вважати ключовими стратегічними надбаннями. Є аж надто багато компаній, які вважають, що дата-аналітики просто перетворюють якісь дані, що вже існують, на гроші. Часто вони небагато уваги приділяють тому, чи є взагалі у компанії компетентні дата-аналітики. Якщо вважати все це надбаннями, можна буде максимально чітко зрозуміти, скільки ви готові в ці надбання інвестувати. Часто у нас немає правильних даних, щоб ухвалити рішення якнайкраще, і/чи правильного спеціаліста, який би якнайкраще допоміг ухвалити рішення за допомогою даних. А отже, якщо ми почнемо думати про ці надбання, то зрозуміємо, що вони доповнюють одне одного. Без потрібних даних навіть найкращі дата-спеціалісти не дадуть великих результатів; рішення не стануть набагато кращими, якщо у вас будуть ідеальні дані, але не буде людини, яка зможе з ними якнайкраще впоратися. У ці надбання часто потрібно інвестувати — як і в будь-які інші. Зібрати команду екстра-класу з data science — завданнячко не з простих, але на ухвалення рішень це може неабияк вплинути. Ми детально обговоримо стратегічні ходи щодо data science у розділі 13. У наступному кейсі розберемо, як ідея добре подумати, як інвестувати в надбання з даних, може повернутися сторицею.

Така історія сталася з маленьким банком Signet у 1990-х. До цього, у 1980-х, data science змінила бізнес споживчих кредитів. Моделювання вірогідності невиконання обов'язків змінило індустрію: від персональної оцінки вірогідності невиконання обов'язків банки перейшли до статегій великих масштабів і ринкової частки, а це повело за собою економіки масштабу. Зараз

це може здаватися дивним, але тоді правила виплат у всіх кредиток були однакові, з двох причин: 1) у компаній не було адекватних інформаційних систем, щоб управляти різноманітними виплатами у великих масштабах, і 2) управління банків вважали, що клієнти не потерплять дискримінації за цінами. Близько 1990 року два прогресивних стратеги (Річард Фейрбенкс і Найджел Морріс) зрозуміли, що в інформаційних технологій вистачає потужності, щоб робити складніші прогностичні моделі — якщо використовувати техніки, які ми обговорюватимемо в цій книжці — і почали пропонувати різні умови (сьогодні це: відсоткові ставки, кредитні ліміти, відсоткові канікули, кешбеки, бали лояльності і так далі). Переконати великі банки взяти їх консультантами і дати спробувати це зробити ці два чоловіки не змогли. Нарешті, коли всі великі банки вони вже обійшли, вони змогли зацікавити один маленький регіональний банк у Вірджинії: банк Signet. Менеджмент банку Signet повірив, що моделювання рентабельності, а не тільки вірогідності невиконання обов'язків — це правильна стратегія. Вони розуміли, що маленька частка клієнтів насправді приносить більше ніж 100 % прибутку банку за операціями по кредитках (тому що решта виходить в нуль або на них вони втрачають гроші). Якщо вони зможуть моделювати рентабельність, вони зможуть робити найкращі пропозиції для найкращих клієнтів і «збирати вершки» з клієнтів великих банків.

Але в банку Signet була велика проблема, яка заважала застосувати цю стратегію. У них не було потрібних даних, щоб змоделювати рентабельність і запропонувати різним клієнтам різні умови. Їх ні в кого не було. Банки видавали кредити за конкретними умовами і за конкретною моделлю вірогідності невиконання обов'язків, і в них були тільки дані, щоб змоделювати рентабельність 1) щодо умов, які вони пропонували раніше, і 2) для тих клієнтів, яким кредит уже пропонувався (тобто тих, яким за наявною моделлю його і так можна було видати).

То що міг зробити банк Signet? Вони застосували фундаментальну стратегію data science: заплатити свою ціну за потрібні дані. Якщо ми розглядаємо дані як надбання бізнесу, потрібно подумати і про те, чи готові ми в неї інвестувати і скільки. У випадку з Signet дані можна було зібрати за рентабельністю клієнтів, яким у межах різних експериментів пропонували би різні умови. Ці різні умови пропонували випадковим клієнтам. Якби це відбувалося не в контексті дата-аналітики, це здавалося би дурістю: ви ж просто втратите гроші! Це правда. У цьому випадку втрата грошей — ціна отриманих даних. Якщо думати дата-аналітично, потрібно розуміти, чи будуть витрати на дані того варті.

То що ж сталося із банком Signet? Як і можна було очікувати, коли вони почали пропонувати людям випадково вибрані умови, кількість поганих рахунків просто злетіла. До цього Signet був лідером у галузі з кількості

амортизованих боргів (не виплачувалося 2,9% заборгованостей), тепер відсоток зріс до 6%. Ці втрати тривали кілька років, поки дата-спеціалісти працювали над прогностичною моделлю, оцінювали її і застосовували, щоб підняти рентабельність. Компанія вважала ці втрати інвестицією в дані, тому продовжувала гнути своє, хоча акціонери були незадоволені. Зрештою операція з кредитками в Signet показала себе і стала такою прибутковою, що її навіть довелося відділити від інших банківських операцій, тому що вони опинилися в тіні успішних споживчих кредитів.

Фейрбенкс і Морріс стали головою й CEO та президентом і COO, і продовжили застосовувати data science у роботі компанії — вони не тільки залучали нових клієнтів, а й утримували старих. Коли людина телефонувала й просила підібрати їй кращу пропозицію, базовані на даних моделі рахували потенційну вірогідність різних дій (різні пропозиції, включно з варіантом, щоб узагалі нічого не змінювати), і на комп'ютері представника служби підтримки висвітлювалася найкраща з можливих пропозицій.

Про маленький банк Signet ви, мабуть, не чули. Але якщо ви читаєте цю книжку, то мали чути про їхню дочірню компанію: Capital One. Нова компанія Фейрбенкса і Морріса виросла в одного з найбільших емітентів кредитних карт у галузі з одним із найменших відсотків боргів, які не виплачуються. У 2000 році, за звітами банку, таких «наукових тестів», як вони їх називали, проводилося 45 000*.

Дослідження із конкретними кількісними даними про цінність надбань даних знайти важко, переважно тому, що фірми не дуже полюбляють розголошувати інформацію стратегічної цінності. Є виняток — дослідження Мартенса і Провоста (2011), де оцінюється, як дані за певними транзакціями клієнтів банку можуть покращити моделі вирішення, які саме продукти запропонувати. Банк створив моделі на основі даних, щоб вирішити, кому які продукти запропонувати. У дослідженні вивчалися багато різних типів даних та їхній вплив на ефективність прогнозів. За допомогою соціодемографічних даних можна було відмінно моделювати типажі клієнтів, які куплять той чи інший продукт. Але це все, на що здатні соціодемографічні дані; коли кількість даних доходить до певної межі, збільшення цієї кількості перестає приносити користь. Натомість деталізовані дані з індивідуальних транзакцій клієнтів (анонімно) дуже суттєво піднімали продуктивність, порівняно із соціодемографічними даними. Зв'язок тут очевидний, він просто вражає, і — що суттєво для теми, про яку тут ідеться — що більше використовується даних, то кращі результати показують прогностичні моделі. Продуктивність підвищувалася в усьому, що досліджували Мартенс і Провост, і ознак спадання ніде не було. Тут є важливий глибинний сенс:

* Можете більше почитати про історію Capital One (Clemons & Thatcher, 1998; McNamee 2001).

банки з більшими надбаннями даних можуть мати важливу стратегічну перевагу над меншими конкурентами. Якщо ці тренди поширяться і банки зможуть застосовувати складну аналітику, то ті банки, у яких даних буде більше, краще визначатимуть, для якого клієнта який продукт буде ідеальним. У результаті або люди почнуть більше користуватися продуктами банку, або впаде вартість нового клієнта, або відбудеться і те, й інше.

Ідею даних як стратегічного надбання однозначно можна застосувати не тільки в Capital One, та й не тільки в банківській сфері. В Amazon досить рано з'явилася можливість збирати дані щодо онлайн-покупців, від чого з'явилися суттєві витрати на переключення: клієнтам були важливі рейтинги і рекомендації, які пропонував Amazon. Відтак, Amazon було легше втримати клієнтів, і вони навіть змогли брати платню за преміум-акаунти (Brynjolfsson & Smith, 2000).

Казино Harrah's відомі тим, що інвестували у збір та майнинг даних про гравців і з маленького казино, яким вони були в середині 90-х, виросли до покупки Caesar's Entertainment у 2005-му і стали найбільшою у світі компанією з азартних ігор. Величезна ціна Facebook стала такою тому, що у них є величезні та унікальні запаси даних (Sengupta, 2012) — інформація про людей та їхні вподобання, а також інформація про структуру соціальної мережі. Інформація про структуру мережі виявилася важливою для прогнозування, і відмінно допомогла моделювати, хто купуватиме певні продукти (Hill, Provost, & Volinsky, 2006). Абсолютно ясно, що надбання даних у Facebook ні з чим неможливо порівняти. Але чи є у них правильні стратегії data science, щоб використати потенціал цих даних на повну — питання відкрите.

Далі у книжці ми ще поговоримо детально про фундаментальні концепції, які стоять за цими історіями успіху, коли досліджуватимемо принципи дата-майнингу і дата-аналітичного мислення.

Дата-аналітичне мислення

Аналіз таких явищ, як плинність, допомагає краще навчитися підходити до проблем «дата-аналітично». Розповісти, що так можна робити, — основна мета цієї книжки. Коли перед вами стоїть бізнес-проблема, ви маєте вміти оцінити, чи можна її вирішити продуктивніше за допомогою даних, і як це зробити. Ми обговоримо декілька фундаментальних принципів і концепцій, які допоможуть усе ретельно обдумувати. Розробимо шаблони й структуруємо аналіз, щоб робити це можна було систематично.

Як уже писалося вище, розуміти data science важливо, навіть якщо ви не збираєтеся її застосовувати, тому що сьогодні аналіз даних для бізнес-стратегії критично важливий. Дата-аналітика дедалі частіше стає на чолі

компаній, тож якщо ви вмієте компетентно взаємодіяти з такими компаніями і працювати в них — це велика професійна перевага. Розуміти фундаментальні концепції і мати шаблони, щоб організувати дата-аналітичне мислення, допомагає не тільки компетентно співпрацювати, але й бачити, де можна ухвалювати рішення на основі даних краще, і помічати базовані на даних загрози від конкурентів.

Компанії в багатьох традиційних галузях намагаються отримати перевагу над конкурентами і досліджують нові та вже наявні ресурси даних. Вони беруть на роботу дата-спеціалістів, щоб ті підключали новітні технології, піднімали їм прибутки і знижували витрати. Окрім того, багато нових компаній із першого дня використовують майнинг даних як ключовий стратегічний компонент. Facebook, Twitter і багато інших компаній із «Digital 100» (Business Insider, 2012) коштують так багато переважно через те, що у них є дані, які вони зібрали або створили самі*. Менеджерам дедалі частіше треба управляти роботою аналітиків і аналітичними проектами, маркетологам — організовувати і розуміти, як працюють кампанії на основі даних, венчурним інвесторам потрібно робити мудрі інвестиції в компанії, які мають важливі дані, а бізнес-стратегам потрібно розробляти плани, де дані будуть використовуватися.

Ось кілька прикладів. Якщо консультант запропонує вам намайнити дані, щоб покращити роботу бізнесу, вам потрібно буде оцінити, чи є в цьому взагалі сенс. Якщо конкурент оголосить про нових партнерів по даних, вам потрібно буде розуміти, чи не ставить це вас стратегічно у невигідне положення. Або, скажімо, вас беруть на роботу у фірму венчурних інвестицій і ваш перший проект — оцінити потенціал інвестиції в рекламне агентство. Його засновники висунули переконливий аргумент, що вони будуть збирати унікальні дані, які матимуть велику цінність, і через це хочуть, щоб їхня компанія була оцінена дещо вище, ніж мала би. Чи є в цьому сенс? Якщо ви розумітимете основи data science, то зможете придумати кілька запитань і перевірите, чи справді це сильні аргументи.

Якщо говорити менш великими, але, мабуть, ближчими до тіла масштабами, то проекти з дата-аналітики стосуються всіх працівників компанії. Людям потрібно співпрацювати з дата-спеціалістами. Якщо у них не буде фундаментального розуміння принципів дата-аналітичного мислення, вони не розумітимуть по-справжньому, що відбувається в компанії. А якщо вони цього не розумітимуть — для проектів з data science це буде куди більш шкідливо, ніж для інших технологічних проектів, тому що завдяки data science можна приймати кращі рішення. У наступному розділі ми будемо

* Звісно, це не новий феномен. Amazon і Google — компанії, які міцно стоять на ногах і коштують неймовірних грошей, тому що мають надбання даних.

говорити про те, що для цього дата-спеціалісти і люди, які відповідають за прийняття рішень у компанії, повинні тісно співпрацювати. Компанії, де такі люди не розуміють, чим займаються дата-спеціалісти, сильно програють — вони втрачають час і сили, або гірше: можуть зрештою прийняти неправильні рішення.

Потреба в менеджерах, які знають дата-аналітику

За підрахунками консалтингової фірми McKinsey and Company, «організаціям не вистачатиме кадрів, які потрібні для того, щоб скористатися перевагами біґ-дати. До 2018 року тільки у Сполучених Штатах не вистачатиме 140 000–190 000 людей з глибокими знаннями в аналітиці, а також 1,5 млн менеджерів та аналітиків, які би знали, як використовувати аналіз біґ-дати для ухвалення ефективних рішень» (Manyika, 2011). Чому менеджерів і аналітиків потрібно вдесятеро більше, ніж людей із глибокими знаннями в аналітиці? Ясна річ, що дата-спеціалістами не так складно управляти, щоб на одного такого потрібно було аж 10 менеджерів! Причина в тому, що бізнес може приймати кращі рішення з допомогою дата-спеціалістів у різних сферах. Однак, як зазначали McKinsey, щоб справді отримати максимальну перевагу, менеджерам у цих сферах потрібно розуміти основи data science.

Ця книжка

У цій книжці йдеться про основи data science і майнингу даних. Це набір принципів, концепцій і технік, які структурують мислення й аналіз. Із ними можна на диво глибоко зрозуміти процеси і методи data science, при цьому не потрібно розбиратись у величезній кількості алгоритмів майнингу даних.

Книжок, у яких добре розписано алгоритми і техніки майнингу даних, багато — і практичних посібників, і теорії з математики та статистики. Але в цій книжці йтиметься про фундаментальні концепції і те, як вони можуть допомогти розв'язати проблему, якщо можна застосувати майнинг даних. Це не означає, що ми ігноруватимемо техніки майнингу даних; багато алгоритмів — це якраз утілення цих базових концепцій. Але, за кількома винятками, ми не говоритимемо про технічні деталі того, як ці техніки працюють. Ми спробуємо пояснити все так, щоб ви розуміли, що ці техніки роблять і як саме вони основані на фундаментальних принципах.

Майнинг даних і data science, новий погляд

Чимало уваги в цій книжці приділено тому, як дістати корисні (нетривіальні, і сподіваємось, такі, які можна використати) патерни або моделі з великої кількості даних (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), і фундаментальним принципам data science, які лежать в основі такого майнингу даних.

У нашому прикладі з прогнозуванням плинності ми братимемо дані про попередні плинності і шукатимемо патерни, наприклад патерни поведінки, які будуть нам корисні — це може допомогти прогнозувати, які клієнти можуть у найближчому майбутньому піти, або допомогти покращити саму послугу.

Фундаментальні концепції data science зібрані із багатьох сфер, де вивчають дата-аналітику. Такі концепції ми обговоримо у книжці, але щоб ви уявляли, про що мова, коротко поговорімо про декілька з них зараз. Ми докладніше розповімо про них і про інші концепції далі у книжці.

Фундаментальна концепція: до видобування корисної інформації з дати для вирішення завдань бізнесу можна підходити систематично, якщо дотримуватися процедури з мотивовано чітко визначеними етапами. Міжгалузевий стандартний процес для майнингу даних, абревіатура CRISP-DM (CRISP-DM Project, 2000) — одна із кодифікацій цього процесу. Якщо тримати такий процес у голові, ви матимете шаблон структури підходу до дата-аналітичних завдань. Наприклад, на практиці ми регулярно бачимо аналітичні «рішення», які не основані на уважному аналізі проблеми або не оцінені належним чином. Для структурованого підходу до аналітики ці часто недооцінені аспекти того, як дані можуть допомагати ухвалювати рішення, дуже важливі. За такого структурованого підходу також одразу чітко видно, де потрібно застосувати людську креативність, а де — потужні аналітичні інструменти.

Фундаментальна концепція: у великому обсязі даних за допомогою інформаційних технологій можна знайти інформативні описові атрибути об'єктів, які нас цікавлять. У нашому прикладі з плинністю об'єкт, який нас цікавить — клієнт, і кожного клієнта можна описати великою кількість атрибутів: частота користування, історія спілкування із службою техпідтримки, і багато інших факторів. Що із цього може дати нам інформацію, що клієнт може піти від компанії, коли в нього закінчиться контракт? Скільки це інформації? Інколи вважають, що цей процес зводиться просто до того, щоб знайти змінні, які «корелюють» із плинністю (цю точку зору ми обговоримо окремо). Бізнес-аналітик може висунути декілька гіпотез і протестувати їх; існують інструменти, які можуть допомогти в цьому експерименті (див. «Інші техніки й технології аналітики», с. 59). Як варіант, аналітик може застосувати інформаційну технологію й автоматично визначити інформативні атрибути — по суті, це буде автоматизований експеримент великого масштабу. Пізніше, як ми побачимо, цю концепцію можна застосувати повторно, щоб створити прогностичні моделі плинності на основі декількох атрибутів.

Фундаментальна концепція: якщо надто уважно придивитися до набору даних, можна щось знайти — але за межами тих даних, на які ви дивитесь, це може не справджуватися. Це називається перепідгонкою набору даних.

Техніки майнингу даних можуть бути дуже потужними, і вміння помітити перепідгонку й уникнути її — одна із найважливіших концепцій, які треба знати, коли починаєте застосовувати майнинг даних до реальних задач. Концепція перепідгонки та її уникнення знайшла своє відображення у процесах, алгоритмах і методах оцінки data science.

Фундаментальна концепція: щоб сформулювати рішення й оцінити результати майнингу даних, потрібно добре подумати про контекст, в якому ці рішення будуть застосовані. Якщо вам потрібно отримати потенційно корисні знання, як сформулювати, що таке корисні? Це повністю залежить від того, де вони будуть застосовуватись. У нашому прикладі із плинністю — як саме ми використаємо патерни, які дізнаємося з даних? Чи потрібно брати до уваги цінність клієнта, а не тільки вірогідність, що він піде? І більш загально — чи допоможе цей патерн ухвалити кращі рішення, ніж якась мотивована альтернатива йому? Наскільки хороше можна було б ухвалити рішення з розумною «стандартною» альтернативою?

Це тільки чотири з усіх фундаментальних принципів data science, які ми розглянемо. У книжці ми детально обговоримо десяток таких фундаментальних концепцій, і проілюструємо, як вони допомагають структурувати дата-аналітичне мислення і розуміти техніки й алгоритми майнингу даних, а також застосування data science взагалі.

Пробірки — не суть хімії: data science і робота дата-спеціаліста

Перш ніж говорити далі, потрібно ненадовго повернутись до інженерної сторони data science. На момент, коли це пишеться, у розмовах про data science усюди згадуються не тільки аналітичні вміння й техніки для розуміння даних, а й популярні інструменти, які при цьому використовуються. У визначеннях дата-спеціалістів (і оголошеннях про пошук таких) говориться не тільки про те, на яких галузях людина повинна знатися, а й про конкретні мови програмування та інструменти. Часто з'являються вакансії, де згадуються техніки майнингу даних (наприклад, випадковий ліс або метод опорних векторів), конкретні сфери, де робота спеціаліста буде застосовуватись (рекомендаційні системи, організація з розміщення реклами), а також популярне програмне забезпечення для обробки біґдати (Hadoop, MongoDB). Мало хто відрізняє саму науку і технології, які допомагають управлятися з великими обсягами даних.

Ми повинні зазначити, що data science, як і комп'ютерні науки, — молода сфера. Деякі пов'язані з data science питання з'явилися зовсім нещодавно, а загальні принципи тільки починають формуватися. Зараз data science приблизно в такому самому стані, як була хімія в середині XIX ст., коли

теорії і загальні принципи тільки формулювалися і сфера була переважно експериментальна. Кожен хороший хімік повинен був добре вміти працювати в лабораторії. Так само складно уявити, щоб дата-спеціаліст не знав ідеально деякі інструменти із програмного забезпечення.

Отже, у цій книжці йтиметься про науку, а не про технології. Ви тут не знайдете інструкцій, як найкраще запустити великий майнинг даних у кластерах Hadoop, або навіть що таке Hadoop, або чому вам варто навчитися в ньому працювати*. Тут ми зосереджуємось на загальних принципах data science, які вже сформулювались. Найімовірніше, через 10 років головні зараз технології зміняться або настільки покращають, що все, що ми про них зараз скажемо, втратить актуальність. Але основні принципи лишилися такими самими, як і 20 років тому, і в наступні десятиліття, найімовірніше, не особливо зміняться.

Підсумки

Це книжка про видобування корисної інформації і знань із великого обсягу даних з метою ухвалювати в бізнесі кращі рішення. Величезні набори даних стали доступними в будь-якому секторі індустрії та будь-яким співробітникам компаній, і можливості майнити ці дані теж стали доступні. В основі величезної кількості технік майнингу даних лежить набагато менший набір фундаментальних концепцій, з яких складається data science. Ці концепції загальні і переважно передають суть майнингу даних і бізнес-аналітики.

Сьогодні, коли бізнес орієнтований на дані, потрібно думати, як ці фундаментальні концепції можна застосувати до реальних завдань бізнесу — як думати дата-аналітично. Наприклад, у цьому розділі ми поговорили про те, що дані потрібно вважати надбанням бізнесу, і якщо ми вже почали думати так, то потрібно подумати і про те, чи готові ми інвестувати в дані (і скільки). Отже, розуміти ці фундаментальні концепції потрібно не тільки дата-спеціалістам, а й усім, хто працює з ними, бере їх на роботу, інвестує в дата-компанії, або управляє в компанії тим, як буде застосована аналітика.

Думати дата-аналітично допомагають концептуальні шаблони, які ми будемо обговорювати по ходу книжки. Наприклад, автоматичне видобування патернів із даних — це процес, де етапи чітко визначені, ми говоритимемо про нього в наступному розділі. Якщо розуміти процес і етапи — це допоможе структурувати дата-аналітичне мислення, зробити його систематичнішим, а отже, помилок і пропущених моментів буде менше.

---

* Ну гаразд: Hadoop — це відкрита архітектура для високопаралелізованих обрахувань, дуже популярна. Це одна із сьогоденних технологій біґ-дати, яка обробляє набори даних, що більші за ті, з якими можуть упоратися бази даних реляційного типу. Hadoop працює на базі шаблону паралельної обробки MapReduce, який розробив Google.

Це переконливий доказ, що ухвалення рішень на основі даних і технології біґ-дати суттєво підвищують продуктивність бізнесу. Data science допомагає ухвалювати рішення на основі даних — й інколи переводить ухвалення рішень на автоматичний рівень — і залежить від технологій зберігання й інженерії біґ-дати, але принципи в них різні. Принципи data science, про які йде мова в цій книжці, відрізняються також і від інших важливих технологій (і доповнюють їх), наприклад, тестування статистичних гіпотез і постановка запитів для баз даних (для них є свої книжки і курси). У наступних розділах ми поговоримо про деякі із цих відмінностей детальніше.