ГЛАВА 3

Решение проблемы

Многие считают это словосочетание, по крайней мере в некоторых отношениях, главным в количественном анализе — ведь именно здесь проводятся аналитические процедуры и проблема получает решение. Конечно, все это очень важно. Но операции на этом этапе более структурированы и точнее определены, чем на предшествующей и последующей фазах. Если у вас нет математической и статистической подготовки, то, скорее всего, вы передадите выполнение этих операций людям с необходимыми навыками и знаниями (см. вставку «Как найти кванта»). Но вне зависимости от ваших личных познаний в математике полезно получить общее представление об основных этапах решения проблемы.

Как найти кванта

Если для решения вашей проблемы требуется количественный аналитик, то существует несколько способов отыскать его.

- Если вы работаете в большой компании, наверняка несколько квантов найдутся в штате. Загляните в отдел маркетинговых исследований, производственную лабораторию, отдел бизнес-аналитики.
- Если ни одного кванта среди сотрудников отыскать не удалось, то можно обратиться к целой армии независимых консультантов.

Проведите интернет-исследование по запросу «консультанты по бизнес-аналитике».

- Если вы хотите привлечь кванта из-за рубежа, то лучше всего обратить внимание на Индию, в частности на компании Mu Sigma, Fractal Analytics и Genpact.
- Возможно, в местном университете удастся найти профессоров или студентов-старшекурсников, специализирующихся на количественном анализе; позвоните заведующему кафедрой статистики, к примеру.
- Если вы считаете необходимым взять кванта на постоянную работу, можно просмотреть объявления на сайтах вакансий, где, как правило, есть соответствующие предложения: например, на сайте Simply Hired есть страница с резюме количественных аналитиков, а на сайте analyticrecruiting.com — с резюме статистиков. Можно обратиться в специализированную рекрутинговую фирму.

Сначала ознакомимся с последовательностью выполняемых этапов. Мы ведь помним, что речь идет об аналитическом проекте, направленном на проверку гипотезы. Сначала мы формулируем проблему (глава 2), затем переходим к моделированию и выбору переменных (первый шаг на этом этапе решения проблемы), а в результате можно будет выдвинуть гипотезу, требующую подтверждения или опровержения. Затем аналитик собирает данные и решает проблему. На каждом из этих шагов необходимо понимать или хотя бы предполагать, как функционирует исследуемый мир, и тогда на основе анализа данных можно будет сделать вывод о том, была ли правильной исходная гипотеза. Однако есть несколько видов анализа, не требующих предварительного выдвижения гипотезы. В интеллектуальном поиске данных и *машинном обучении* (когда модели разрабатываются на основе закономерностей, выявленных в имеющихся данных, с помощью программного обеспечения давая быстрый и оптимальный результат) аналитик просто вводит в компьютер массив данных и запускает поиск закономерностей. Все гипотезы выдвигаются уже потом, на этапе интерпретации и распространения результатов.

Нам не слишком нравится этот подход: в основном потому, что зачастую он дает необъяснимые результаты. А поскольку ни один аналитик не пытался использовать анализ данных для подтверждения своих взглядов на происходящее вокруг, то и комментировать результаты анализа или убеждать в необходимости изменить решение на их основе никто не будет. Однако иногда случаются обстоятельства, в которых подход к анализу как к «черному ящику» может сэкономить немало времени и труда аналитикам. В среде больших данных, где постоянно генерируются колоссальные массивы информации, у аналитика не всегда есть возможность формулировать гипотезы до проведения анализа данных. Например, при размещении рекламы на сайтах издательств решения принимает автоматизированная система в тысячные доли секунды, а компании, занимающиеся этой работой, генерируют несколько тысяч статистических моделей каждую неделю. Очевидно, такой вид анализа не рассчитан на выдвижение гипотез и рассмотрение результатов людьми, поэтому машинная работа здесь абсолютно необходима. Но по большей части в дальнейшем изложении мы будем иметь дело с этапами и методикой анализа на основе проверки гипотез.

Шаг 3. Моделирование (выбор факторов)

Модель — это *преднамеренно упрощенное* представление определенного события или ситуации. Термин «преднамеренно» означает, что модель разрабатывается специально для решения конкретной проблемы. Термин «упрощенно» говорит о том, что следует исключить из рассмотрения все банальные и несущественные детали, выделив важные, полезные и ключевые особенности, определяющие специфику проблемы. Проиллюстрируем процедуру выбора факторов на примере.

3. Моделирование

Модель можно сравнить с карикатурой. Она заостряет внимание на некоторых чертах — носе, улыбке, кудрях, — и на их фоне другие черты теряют выразительность. Хорошая карикатура отличается тем, что отдельные черты выбираются обдуманно и эффективно. Точно так же модель

акцентирует внимание на отдельных особенностях реального мира. При построении любой модели вам придется действовать избирательно. Нужно выбрать именно те особенности, которые имеют отношение к решению вашей проблемы, и пренебречь остальными. Модель носит схематичный характер, чтобы помочь пользователю сфокусироваться на исследуемой проблеме*.

Отсюда следует, что модели не могут быть абсолютно корректными. Знаменитый статистик Джордж Бокс как-то заметил, что «...все модели некорректны, но некоторые при этом полезны»**. Ключевая проблема в том, чтобы определить, когда модель приносит пользу, а когда она некорректна настолько, что искажает реальность. В главе 5 мы подробнее поговорим об этом. А пока заметим, что одним из ключевых является вопрос о выборе факторов для включения в модель.

Каким образом отбираются факторы для модели и прогнозируются их взаимосвязи? По большей части мы в этом вопросе руководствуемся субъективными соображениями. Гипотеза, то есть априори разработанная концепция анализа, представляет собой не более чем наукообразные предположения о том, какие факторы имеют наибольшее значение в каждом конкретном случае. На этом этапе разработка модели требует логического мышления, опыта и знакомства с предшествующими исследованиями. Только в этом случае можно с большой долей уверенности предположить, какие зависимые (те, которые нужно прогнозировать или объяснить) или независимые факторы сыграют основную роль. Можно попытаться протестировать модель — именно это отличает аналитическое мышление от менее точных методов принятия решений вроде интуиции.

Например, если вы социолог и пытаетесь прогнозировать динамику дохода семьи (зависимая переменная), то можно предположить, что независимыми переменными в вашей модели будут возраст, образование, семейный статус и количество работающих постоянно членов семьи. Именно эти переменные имеют смысл при прогнозировании семейного дохода. Впоследствии, в процессе количественного анализа (а точнее, на этапе анализа данных) вы можете обнаружить, что модель недостаточно точно отражает реальную ситуацию, и захотите

---

* *Starfield A., Smith K., and Bleloch A*. How to Model It: Problem Solving for the Computer Age. — New York: McGraw-Hill, 1994. P. 19.

** *Box G. and Drape N*. Empirical Model-Building and Response Surfaces. New York : Wiley, 1987. P. 424.

пересмотреть состав переменных при условии, что по новым переменным можно получить данные.

Даже очень субъективные модели и переменные могут быть полезны для уточнения проблемы. Например, Гарт Сандем, известный популяризатор науки, математики, юморист и писатель на темы гик-культуры, многие жизненные проблемы решал путем анализа субъективно отобранных, но все равно полезных переменных*. В частности, так он подходил к решению вопроса о том, какое именно домашнее животное лучше выбрать и стоит ли его заводить вообще.

Какие переменные человек принимает во внимание, решая, заводить ли домашнее животное? Сандем отобрал следующие:

— Постоянная жизненная потребность в любви (*D*, 1–10, где 10 баллам соответствует жизнь как у начальника тюрьмы днем и честного налогоплательщика ночью).
— Общий уровень ответственности (*R*, 1–10, где 1 балл соответствует убежденности в том, что «дети, налоговый инспектор и дела как-нибудь сами устроятся, если оставить их в покое»).
— Наиболее продолжительная поездка в последние шесть месяцев (*T*, дней).
— Продолжительность сверхурочных (*H*, часов в день).
— Ваша терпимость к проделкам других существ (*M*, 1–10, где 1 балл означает, что вы ведете себя как Стервелла де Виль, а 10 баллов — как доктор Дулиттл).
— Насколько вы заботливы (*N*, 1–10, где 1 балл означает «мой кактус засох»).

Все эти переменные весьма субъективны, но они, по всей видимости, полезны и, уж конечно, забавны. Сандем вывел следующее уравнение (выглядит довольно устрашающе!), где обобщающим показателем является F_{ido} — индекс готовности к заведению домашнего питомца:

$$F_{id0} = \frac{(M+N)^{\sqrt{D}} + HR}{8T^2}.$$

Наиболее важной переменной в этом уравнении является *D* — потребность в любви, которая прямо пропорционально связана с результирующим показателем. Неплохо также, если у вас есть немного

* *Sundem G.* Geek Logik: 50 Foolproof Equations for Everyday Life. New York : Workman, 2006.

свободного времени (*H*), чтобы проводить его с питомцем, и вы ответственный человек *(R)*. Эти две переменные также прямо пропорционально влияют на F_{ido}. Но если вам приходится много ездить, значение вашего индекса существенно снизится. В зависимости от итогового результата Сандем предлагает выбрать одно из следующих домашних животных:

— если F_{ido} менее 1, то даже морские рачки будут слишком обременительны;
— если F_{ido} составляет от 1 до 2, попробуйте завести золотых рыбок;
— если F_{ido} составляет от 2 до 3, можно завести кошку;
— если F_{ido} превышает 3, то можно взять собаку.

Джин Хо подставил собственные значения в это уравнение и получил значение индекса готовности к заведению домашнего питомца 0,7, а значит, ему не стоит рисковать даже с кактусом.

Конечно, кто-то может сказать, что слишком большая точность расчетов при решении данного вопроса не требуется, но так или иначе этот пример показывает, что даже очень субъективные и банальные решения можно оценить количественно и смоделировать.

Какие переменные отобрать, а какие отбросить — зависит от цели разработки модели и того, связана ли переменная непосредственно с решением проблемы. Например, если вы рисуете карту Нью-Йорка, то расстояния между точками имеют большое значение и должны быть пропорциональны реальным расстояниям. Однако если вы рисуете схему нью-йоркского метро, то расстояния между станциями на карте совсем не обязательно должны быть пропорциональны расстояниям на местности. Ведь главная цель схемы метро — это показать, как можно добраться от одной станции до другой.

Еще один прекрасный пример важности тщательного выбора переменных модели — это спор по поводу того, кто является автором серии опубликованных в 1861 году писем. Десять писем, подписанных Квинтусом Куртиусом Снодграссом, появились в New Orleans Daily Crescent. В них мистер Снодграсс (ККС) описывал свои военные приключения во времена службы в Национальной гвардии Луизианы. Сразу после публикации письма не привлекли особого внимания. Они впервые попали в поле зрения широкой публики лишь в 1934 году, то есть спустя семьдесят три года после выхода из печати. О них в своей книге Mark Twain, Son of Missouri упомянула Минни Брашер.

В частности, она привела текст одного из писем, пересказала содержание трех других и сделала смелый вывод о том, что «письма ККС имеют огромное значение в качестве свидетельства становления Марка Твена как юмориста; именно Марка Твена следует признать их автором, а некоторые различия в стиле можно объяснить его стремлением выработать свой собственный литературный стиль»*. Оставшиеся шесть писем ККС опубликовал и проанализировал Эрнст Лейзи в 1946 году**. Проведенный им тщательный анализ аналогий позволил утверждать, что письма действительно написаны Твеном, но кое-кто из литературных исследователей до сих пор считает, что у них был другой автор.

В русле исследований вопроса о том, действительно ли Шекспир был автором всех приписываемых ему произведений, Томас Менденхолл в конце двадцатого века опубликовал две статьи, в которых изложил статистический подход к проблеме определения авторства. Топ-менеджер нефтяной компании Клод Бринегар, имевший хорошее университетское образование и увлекавшийся коллекционированием первых изданий книг Марка Твена, изучил историю вопроса и применил метод Менденхолла, впоследствии получивший название *стилометрии,* или количественного анализа литературного стиля, к письмам ККС.

Этот метод основан на предположении о том, что, хочет он того или нет, каждый автор чаще использует одни слова, чем другие, и сохраняет одинаковый литературный стиль, по крайней мере в долгосрочной перспективе. С позиций количественного анализа это означает, что доля слов определенной длины будет постоянной во всех текстах, написанных данным автором. Если доля слов определенной длины в двух разных текстах существенно отличается, это можно считать подтверждением того, что тексты написаны разными авторами. В качестве переменных для анализа писем ККС выбирались слова различной длины, и их удельный вес сравнивался с аналогичными показателями из работ, определенно принадлежавших перу Твена. Для проверки авторства проводился тест по критерию согласия. Результаты тестирования показали, что расхождения по набору переменных слишком велики, чтобы считать их случайными, — поэтому вряд ли

---

* *Brashears M.* Mark Twain: Son of Missouri. Whitefish, MT : Kessinger Publishing, 2007.

** *Leisy E. (ed.).* The Letters of Quintus Curtius Snodgrass. Irving, TX : University Press of Dallas, 1946.

Марк Твен является автором этого произведения (подробности см. на сайте книги)*.

Далее в этой главе мы еще поговорим об анализе текстов (в противоположность анализу чисел), а пока отметим, что Бринегар в процессе анализа перевел слова в числа.

Шаг 4. Сбор данных (измерения)

На следующем шаге анализа проводится сбор данных и измерения выбранных переменных. Измерение — это определение значения переменной; массив данных — это набор таких значений. Существуют разные способы измерения переменных (см. вставку «Способы измерения переменных»). Сформулированная проблема сначала представляется в виде набора переменных в процессе моделирования, а затем приобретает вид массива данных в результате измерения.

Таким образом, массив данных организован с учетом переменных, выбранных на предыдущем шаге.

Методы измерения данных

Известны три основных метода измерения данных.

Двоичные переменные. Такие переменные имеют только два значения, и для целей статистического анализа лучше определять их как наличие или отсутствие определенного фактора со значениями 0 и 1. В качестве примера можно привести данные о поле респондентов, когда возможен выбор двух значений: женщина или мужчина (в первом случае переменная приобретает значение 1, во втором — 0), или о наличии гражданства США (либо гражданин, либо нет).

Категориальные (также называемые номинальными) переменные. В этом случае переменная может приобретать одно из нескольких заранее определенных значений. Так измеряются цвет глаз, вкус мороженого, штат или район проживания. Поскольку перевод таких

* *Brinegar C.* Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship // Journal of the American Statistical Association. 1963, no. 58.

значений в количественную форму представляет определенные сложности, существует отдельное направление статистики, занимающееся анализом категориальных данных.

Ординальные переменные. Эти переменные имеют упорядоченные количественные значения, причем чем оно больше, тем сильнее выражен соответствующий признак. Таким образом, у этих переменных разница между 1 и 2 — это не то же самое, что разница между 5 и 6. Типичный пример ординальных переменных — шкала Ликерта, получившая название в честь автора, социолога Ренсиса Ликерта. Обычно применяется в опросах и включает такие значения, как «полностью согласен», «отчасти согласен», «не могу выразить отношение», «отчасти не согласен», «не согласен». Несколько ординальных переменных, сведенных вместе, носят название шкалы Ликерта.

Количественные (интервальные и рациональные) переменные. Значения этих переменных выражены числами, обычно в стандартных единицах: вес в фунтах или килограммах, рост в дюймах или сантиметрах. Чем больше значение, тем сильнее выражен соответствующий параметр. Количественные переменные хорошо подходят для традиционных видов статистического анализа, например корреляционного или регрессионного.

Если значения нужных вам переменных часто собирает и анализирует кто-то еще (иногда такие факты всплывают во время изучения предыдущих поисков решения), то этот этап будет несложным. Можно просто позаимствовать результаты измерений, полученные вашими предшественниками. Однако в некоторых случаях приходится вести работу самостоятельно. Нужно помнить, что даже субъективные события можно систематически измерять.

4. Сбор данных

Предположим, что вам нужно собрать данные по волнующей в наше время многих (если судить по телевизионной рекламе)

проблеме мужской потенции. Оказывается, что вам повезло: на эту тему уже проводился сбор данных, которые вполне подходят для ваших целей. Однако если бы вы были первопроходцем в этой области, то пришлось бы проводить сбор данных самостоятельно.

В 1990-е годы Р. С. Розен и его коллеги разработали компактный, надежный и простой для изучения критерий потенции, чувствительный к изменениям в состоянии здоровья пациентов в результате лечения*. О проблемах с потенцией можно узнать только от самого пациента. Объективных диагностических тестов не существует, и это весьма усложняет жизнь практикующим врачам. Розен и его коллеги определили, что ключевыми переменными для анализа проблемы мужской потенции являются:

- регулярность эрекции
- сила эрекции
- частота возбуждения
- способность к половому акту
- удовлетворение

В их разрезе был организован сбор информации с использованием вопросов, приведенных в табл. 3.1.

Вопрос о том, возможно ли ответы на них перевести в диагноз, решается довольно просто. Каждому варианту ответа присваивается балл от 5 до 25. Проблему с потенцией классифицировали по пяти степеням: серьезная (5–7), умеренная (8–11), от умеренной до незначительной (12–16), незначительная (17–21) и отсутствие проблемы (22–25). Этот простой в применении диагностический тест называется IIEF-5 (вариант Международного индекса эректильной функции из пяти вопросов) и прекрасно иллюстрирует способы сбора субъективной информации.

Неважно, каким объемом данных вы располагаете, — всегда остаются возможности собрать еще больше или расширить круг показателей, по которым собирались данные. После начала работы над проектом обычно выявляется, что тех показателей, которые были отобраны на этапе идентификации проблемы, недостаточно. Талантливый квант Рама Рамакришнан, о котором мы уже говорили в главе 2, в своем блоге

* *Rosen R. et al.* The International Index of Erectile Function (IIFF): A Multidimensional Scale for Assessment of Erectile Function // Urology. 1997. Vol. 49, no. 6. P. 822–830; *Rosen R. et al.* Development and Evaluation of an Abridged, 5-item Version of the International Index of Erectile Function (IIEF-5) as a Diagnostic Tool for Erectile Dysfunction // International Journal of Impotence Research. 1999. Vol. 11. P. 319–326.

описал интересный способ улучшить качество данных: «Одно из моих любимых занятий — улучшать качество данных. Это означает не увеличивать их количество, а, скорее, получать новые по характеру данные по сравнению с теми, которые использовались до этого момента. Если у вас имеются демографические данные, добавьте данные об объемах закупок. Если у вас и те и другие, попробуйте добавить функцию их свободного просмотра. Если у вас есть количественные данные, добавьте к ним текстовые (кстати говоря, в последней работе мы получили весьма обнадеживающие результаты, добавив к традиционным данным об объемах продаж и сбытовых мероприятиях текстовые данные о покупателях с целью их персонификации и моделирования потребительского поведения)»*.

Таблица 3.1

Ключевые переменные
для диагностирования эректильной дисфункции

За последние шесть месяцев					
Оцените вашу уверенность в том, что вы можете ощутить и поддерживать эрекцию	Очень низкая — 1 балл	Низкая — 2 балла	Средняя — 3 балла	Высокая — 4 балла	Очень высокая — 5 баллов
В тех случаях, когда с помощью сексуальной стимуляции удавалось добиться эрекции, была ли она достаточной?	Никогда или почти никогда — 1 балл	Изредка (гораздо меньше половины случаев) — 2 балла	Иногда (примерно в половине случаев) — 3 балла	В большинстве случаев (гораздо чаще, чем в половине случаев) — 4 балла	Всегда или почти всегда — 5 баллов

* *Ramakrishnan R.* Three Ways to Analytic Impact // The Analytic Age (blog), July 26, 2011. URL: http://blog.ramakrishnan.com/.

Продолжение таблицы 3.1

За последние шесть месяцев					
Насколько часто в процессе полового акта вам удавалось поддерживать эрекцию после введения?	Никогда или почти никогда — 1 балл	Изредка (меньше половины случаев) — 2 балла	Иногда (примерно в половине случаев) — 3 балла	В большинстве случаев (гораздо чаще, чем в половине случаев) — 4 балла	Всегда или почти всегда — 5 баллов
Насколько сложно вам поддерживать эрекцию до завершения полового акта?	Очень сложно	Довольно сложно	Сложно	Есть некоторые затруднения	Затруднений не испытываю
Как часто половой акт приносит вам удовлетворение (от общего числа попыток)?	Никогда или почти никогда — 1 балл	Изредка (гораздо меньше половины случаев) — 2 балла	Иногда (примерно в половине случаев) — 3 балла	В большинстве случаев (гораздо чаще, чем в половине случаев) — 4 балла	Всегда или почти всегда — 5 баллов

Специалист по интеллектуальному поиску данных Ананд Раджараман также писал в своем блоге о возможностях улучшения качества анализа за счет включения новых данных.

> Я веду курс по интеллектуальному поиску данных в Стэнфордском университете. Студентам поручают выполнить аналитический проект, включающий нетривиальный вариант интеллектуального поиска данных. Многие

из них пытались разработать более совершенную методику подбора рекомендаций по поводу кино, чем в проекте Netflix Challenge.

Это яркий пример того, как действует конкуренция. Netflix предоставляет огромный массив данных о рейтингах 18 тысяч фильмов, выставленных почти полумиллионом посетителей сайта. Основываясь на этой информации, надо спрогнозировать рейтинги, которые выставят пользователи тем фильмам, которые они еще не оценивали. Первая группа аналитиков, которой удастся разработать методику, работающую лучше, чем Netflix Challenge, получит миллион долларов!

Студенты в моей группе пытались применить разные подходы для решения этой проблемы, причем одна команда использовала уже известные алгоритмы, а вторая — новые идеи. Их результаты позволяют взглянуть на проблему шире. Первая команда предложила очень сложный алгоритм, основанный на имеющихся данных. Вторая использовала довольно простой алгоритм, но зато на основе не только имеющихся, но и новых данных, которых в базах Netflix не было. Их позаимствовали из онлайновой базы данных о фильмах (Internet Movie Database). Какая из команд, по вашему мнению, добилась лучших результатов? Представьте себе, вторая! Ее результаты оказались почти так же хороши, как и результаты лучших участников конкурса Netflix!*

В том же посте Раджараман отмечает, что появившийся недавно источник информации — гипертекстовые ссылки — стал отличительной чертой поискового механизма Google по сравнению с прочими поисковиками, использовавшими только текст на веб-страницах. В своем высокорентабельном алгоритме AdWords, предназначенном для размещения рекламы, Google также использовал дополнительные данные, которыми на тот момент не интересовался ни один из конкурентов — коэффициент эффективности баннеров (отношение числа щелчков к общему числу показов), рассчитывавшийся для каждого баннера рекламодателей.

Раджараман и Рамакришнан в один голос утверждают, что больший объем и лучшее качество данных почти в любом случае важнее, чем лучший алгоритм расчетов. Оба ссылаются на опыт розничного бизнеса и электронной коммерции, но и в других областях существует

* *Rajaraman A.* More Data Usually Beats Better Algorithms // Datawocky (blog), March 24, 2008. URL: http://anand.typepad.com/datawocky/2008/03/more-data-usual.html.

множество подобных примеров. Топ-менеджер команды НБА Houston Rockets Дэррил Морей является одним из лучших аналитиков в профессиональном баскетболе (мы вспомним о нем в главе 6). Он считает, что «реальное преимущество обеспечивают лишь эксклюзивные данные», и держит в штате нескольких квантов, анализирующих действия соперников в защите в каждой игре*. Кроме того, Морей стал одним из первых менеджеров в НБА, которые начали анализировать видеозаписи отдельных матчей.

В страховом бизнесе одним из факторов, долгое время отличавших компанию Progressive от менее склонных к аналитике компаний, стала ее уникальная база данных. Компания первой стала использовать кредитный рейтинг агентства FICO (этот пример рассматривается в главе 4) в качестве одной из переменных в модели страховых тарифов, а также в течение долгого времени использовала гораздо больше данных и переменных в анализе клиентского риска и расчете страховых тарифов, чем ее конкуренты. Progressive выступила первопроходцем в сборе данных о манере вождения автомобилей клиентами (конечно, с разрешения последних) и расчете страховых тарифов в зависимости от их водительских привычек (эту программу компания сейчас называет Snapshot). Вы можете не захотеть сообщать страховой компании такие сведения, но если проявите себя осторожным водителем, то получите скидку по страховке.

Ценность вторичных данных

Многие аналитики самостоятельно собирают, а затем анализируют данные. Но иногда можно воспользоваться данными, собранными кем-то другим (так называемыми *вторичными данными*), и существенно сэкономить время. Обычно вторичные данные получают из результатов переписей, опросов, внутренней документации и других подобных источников. Таких данных везде очень много, и они просто ждут, когда аналитики обратят на них внимание.

Иногда вторичные данные помогают получить очень важные результаты. Достаточно вспомнить, например, работу астронома Иоганна Кеплера. Он родился в бедной семье, но ему повезло получить очень

* *Morey D.* Success Comes from Better Data, Not Better Analysis // Harvard Business Review (blog). August 8, 2011. URL: http://blogs.hbr.org/cs/2011/08/success_comes_from_better_data.html.

точные вторичные данные о движении астрономических объектов, тщательно собиравшиеся в течение нескольких десятилетий. Необыкновенный математический талант и удача помогли ему разгадать тайны планет.

Данные достались Кеплеру в основном от датского дворянина и блестящего астронома Тихо Браге (1546–1601), который сумел сделать точные астрономические наблюдения при помощи уникальных инструментов еще до изобретения телескопа. При поддержке датского короля Браге построил исследовательский центр, получивший название Ураниборг (Небесный замок), и разместил в нем лучшую на тот момент в Европе обсерваторию. Он сам разработал и изготовил высокоточные измерительные инструменты, откалибровал их и каждую ночь в течение более чем двадцати лет вел астрономические наблюдения.

В 1600 году Браге пригласил Кеплера, блестящего, но бедного учителя, в помощники. Они не очень-то ладили: сказывалась разница в характерах и жизненном опыте. Браге опасался, что его умный молодой помощник со временем затмит его и станет лучшим астрономом своего времени. В следующем, 1601 году Браге внезапно заболел и умер. Разгорелся спор о его наследстве, и Кеплер понял, что если не будет действовать быстро, то навсегда потеряет возможность воспользоваться данными, собранными учителем. Он немедленно забрал результаты наблюдений (по его выражению, узурпировал их) и уже не выпустил из рук. Через два дня после похорон Браге Кеплер был назначен на его должность придворного математика. Наконец-то уникальная коллекция записей об астрономических наблюдениях была полностью в его распоряжении! Анализируя их, Кеплер сделал вывод, что орбиты планет имеют форму эллипса, а затем сформулировал свои знаменитые законы движения планет*.

Конечно, можно привести массу более современных примеров использования вторичных данных. Например, источник вторичных данных компании Recorded Future прекрасно известен: интернет.

* Tycho Brahe // Wikipedia. URL: http://en.wikipedia.org/wiki/Tycho_Brahe; *Fowler M*. Tycho Brahe. URL: http://galileoandeinstein.physics.virginia.edu/1995/lectures/tychob.html; *Koestler A*. The Watershed: A Biography of Johannes Kepler. Doubleday, 1960; Johannes Kepler // Wikipedia. URL: http://en.wikipedia.org/wiki/Johannes_Kepler; Johannes Kepler // Encyclopædia Britannica Online Academic Edition. URL: http://www.britannica.com/EBchecked/topic/315225/Johannes-Kepler.

Основатель компании — консультант по аналитике Кристофер Альберг, а основной вид деятельности — анализ информации в интернете на предмет частоты упоминания и классификации тех или иных событий и субъектов. Особое внимание компания уделяет подсчету предсказаний — упоминаний о будущем. Данные и аналитика пользуются спросом у государственных разведывательных служб, интерес которых к частоте упоминания террористических актов и войн легко объясним. Среди клиентов есть и финансовые компании, которые интересуются данными, отражающими настроения инвесторов и потребителей.

Первичные данные

Но если вам не так повезло, как Кеплеру или Recorded Future, и не досталось ценных вторичных данных (а может быть, данных, имеющих отношение к вашей проблеме, пока просто не существует), то вам придется собрать их самостоятельно (это *первичные данные*). Существует несколько методов получения первичных данных: опрос, включающий разработку анкет и проведение интервью; наблюдения, в ходе которых наблюдатель открыто или скрытно фиксирует информацию; тщательно спланированные и контролируемые «сумасшедшие» эксперименты, предназначенные для изучения специфических проблем. Выбор метода сбора данных зависит от особенностей сформулированной проблемы и включенных в анализ переменных.

Структурированные и неструктурированные данные. В течение долгого времени почти все количественные аналитики работали со *структурированными* данными: данными в числовой форме, которые легко можно представить в табличном виде. Независимо от того, проводится ли анализ с помощью электронных таблиц, мощной статистической программы или старомодного калькулятора, все равно данные структурируются при помощи строк и столбцов (обычно в строках отражаются события или наблюдения, а в столбцах — значения соответствующих переменных). Все, что вам оставалось выяснить, это сколько наблюдений следует сделать и сколько знаков после запятой показывать в таблице.

Но положение дел стало меняться с распространением в последние годы XX века анализа текстов. На примере истории с письмами

Марка Твена мы показали, что в тексте можно искать не только числа, но и логические закономерности. Типичный вопрос: как часто повторяется в тексте то или иное слово? Текст представляет собой пример неструктурированных данных. Поскольку он состоит из определенной последовательности слов, его трудно разложить по строкам и столбцам таблицы. Однако лишь после 2000 года резко возросли объем и разнообразие неструктурированных данных. Именно этот год стал началом массированного использования интернета, когда компании вроде Recorded Future приступили к анализу огромных массивов данных в виде текста, изображений и щелчков мышки. Телекоммуникации и социальные медиа поставляют огромные объемы информации социальной направленности. Объем аудио- и видеоданных, которые хотели проанализировать организации, рос в геометрической прогрессии. Революция в генетике привела к необходимости анализировать большие объемы сведений о генах.

Сейчас мы официально вступили в век больших данных, когда обработка нескольких петабайт информации стала для организаций рутинным делом. (1 петабайт равен 1000 терабайт, или 10^{15} байт, то есть 1 000 000 000 000 000 единиц информации.) Например, хранилище информации eBay имеет объем более чем в 40 петабайт. Каждое ваше нажатие на изображение видеокамеры или украшенной цветочным орнаментом вазы фиксируется в общей базе данных.

Анализ данных такого рода имеет существенные отличия от анализа структурированных количественных данных, особенно на первых шагах. Во многих случаях, прежде чем приступить к подсчету, требуется провести тщательную фильтрацию и классификацию, а также другие подготовительные операции. *Специалист по базам данных* — это человек, глубоко разбирающийся не только в анализе данных, но и в процедурах их подготовки к проведению анализа. Такие программные инструменты, как Hadoop и MapReduce, получают все большее распространение в организациях, сталкивающихся с необходимостью анализа больших данных. Они предназначены для такой фильтрации и классификации данных, которая позволит применять количественные методы анализа. Видео- и аудиоинформация также требует серьезной обработки, прежде чем можно будет ее анализировать количественными методами. Во многих случаях после подготовки организация будет анализировать

эти массивы данных при помощи традиционных статистических приложений.

Билл Франкс из компании Teradata в своем посте в блоге Международного института аналитики подчеркивает*:

> Неструктурированные данные в последнее время очень популярный предмет для обсуждения, поскольку слишком многие распространенные источники больших данных предоставляют их в неструктурированном виде. Но зачастую забывают об очень важном обстоятельстве: никакая аналитика не имеет дела напрямую с большими данными. Последние могут стать толчком к проведению анализа, но когда дело доходит до собственно аналитических процедур, то неструктурированные данные не обрабатываются. «Как же так?» — спросите вы. Позвольте объяснить.
>
> Вот пример: отпечатки пальцев. Если вы любите сериалы вроде «CSI: полиция Майами», то постоянно видите, как эксперты идентифицируют их. Отпечатки пальцев представляют собой неструктурированные данные, причем довольно большого объема — если изображение высококачественное. Когда полицейские — в сериале или в жизни — сравнивают их, то есть ли смысл накладывать одно изображение на другое? Нет. Сначала они определяют несколько ключевых точек на каждом отпечатке. Затем по этим точкам формируется карта (многоугольник). Именно по этим картам производится сравнение. Особое значение имеет тот факт, что карта представляет собой структурированные данные, к тому же небольшого объема, даже если исходное изображение «весило» много. Как видите, хоть неструктурированные данные и необходимы для начала анализа, но в самом процессе обрабатываются не они, а полученные из них структурированные данные.
>
> Всем понятный пример такого рода — анализ текстов. В общедоступных средствах массовой информации в последнее время принято вести смысловой анализ множества сообщений. Но можно ли непосредственно анализировать твиты, посты в Facebook и прочие посты и комментарии в соцсетях на предмет их смысловой оценки?
>
> В действительности — нет. Текст необходимо разбить на фразы или слова. Затем определенным фразам и словам присваивается определение «положительный» или «отрицательный». В простом случае фразе или слову, определенному как «положительное», присваивается значение 1,

* *Franks B.* Why Nobody Is Actually Analyzing Unstructured Data // International Institute for Analytics (blog post). March 9, 2012. URL: http://iianalytics.com/ 2012/03/why-nobody-is-actually-analyzing-unstructured-data/.

«отрицательному» — 1, а «нейтральному» — 0. Смысл сообщения оценивается по сумме значений входящих в него слов или фраз. Таким образом, оценка ведется на основе структурированных количественных данных, полученных из первоначально неструктурированного источника — текста. Любой дальнейший анализ тенденций или стандартных моделей полностью основывается на структурированном, количественном выражении текста, но не на самом тексте.

Так же как в ситуациях, приведенных Франксом в качестве примера, многие приложения для обработки больших данных первоначально предназначались для обработки неструктурированных данных, но после того как те проходят через такие приложения, как Hadoop и MapReduce, можно их анализировать как структурированные данные с использованием статистических программ или инструментов визуализации.

Шаг 5. Анализ данных

Поскольку сами по себе данные ни о чем не говорят, нужно проанализировать их и определить значения и взаимосвязи. Анализ данных включает выявление устойчивых моделей, или взаимосвязей между переменными, значения которых введены в массив данных. Если удается выявить взаимосвязи, тогда можно объяснить динамику переменных. Тогда будет легче решить проблему.

Предположим, что мы собрали данные по выборке избирателей относительно их намерения голосовать за того или иного кандидата. Метод сбора данных — опрос по телефону. Но в процессе анализа мы пытаемся выявить, каким образом регион проживания, образование, уровень дохода, пол, возраст и партийная принадлежность способны повлиять на выбор того или иного кандидата. Для обнаружения зависимостей в данных можно использовать целый ряд методов, начиная с достаточно простых — графиков, расчета удельного веса и средних значений переменных — и заканчивая сложными статистическими исследованиями.

5. Анализ данных

Параметры массива данных и сложность предстоящего анализа подскажут, какими именно методами лучше воспользоваться. В главе 2 мы привели примеры таких методов. Если вы просто описываете сложившуюся ситуацию, то достаточно составить отчет или разработать набор графиков, показать, сколько анализируемых событий случилось в каждом временно́м интервале, и прокомментировать эту информацию. Обычно приходится приводить сведения о некоторых *показателях, отражающих основную тенденцию*, в частности о средних значениях — медианах.

Исходя из этих условий, потребуется программное обеспечение, ориентированное на составление отчетов. Сбалансированные системы показателей, сводные таблицы, тревожные сигналы — это все формы отчетов. Во вставке «Основные поставщики аналитического программного обеспечения» мы перечислили ключевых поставщиков программного обеспечения, обеспечивающего визуальное представление результатов анализа.

Основные поставщики аналитического программного обеспечения

ПРОГРАММЫ — ГЕНЕРАТОРЫ ОТЧЕТОВ

- BOARD International
- IBM Cognos
- Information Builders WebFOCUS
- Oracle Business Intelligence (including Hyperion)
- Microsoft Excel/SQL Server/SharePoint
- MicroStrategy
- Panorama
- SAP BusinessObjects

ИНТЕРАКТИВНАЯ ВИЗУАЛЬНАЯ АНАЛИТИКА

- QlikTech QlikView
- Tableau
- TIBCO Spotfire

КОЛИЧЕСТВЕННЫЕ МЕТОДЫ
И СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

- IBM SPSS
- R (свободно распространяемое программное обеспечение)
- SAS

У всех этих поставщиков программного обеспечения есть программы для графического представления данных, но некоторые из них специализируются именно на *интерактивной визуальной аналитике*, то есть визуальном представлении данных и отчетов. Иногда такие программы используют для простого построения графиков, иногда для исследования данных: законов распределения данных, позволяющих идентифицировать *выбросы* (точки с нетипичными значениями) и визуальную взаимосвязь между переменными. Таких поставщиков мы выделили в отдельный список.

Кроме того, в перечне выделена группа поставщиков, специализирующихся на еще одной категории аналитических программ — *количественных методах и статистическом моделировании*. В них статистика используется для выявления взаимосвязи между переменными и переноса закономерностей выборки на генеральную совокупность. Его формы — предсказательная аналитика, рандомизированное исследование и различные формы регрессионного анализа. Программное обеспечение для статистического моделирования и для генерирования различных отчетов разрабатывается разными группами поставщиков, хотя со временем они начинают смешиваться между собой.

Например, самая распространенная в мире аналитическая компьютерная программа Microsoft Excel (хотя большинство пользователей считает ее всего лишь электронной таблицей) способна решать некоторые задачи статистического анализа (и визуальной аналитики), равно как и генерировать отчеты. Однако если вам необходимо обработать большой массив данных или построить сложную статистическую модель, то возможностей Excel не хватит. Поэтому к данной категории программного обеспечения она не относится. В корпоративной среде для решения аналитических задач в дополнение к Microsoft

Excel часто используют и другие программы Microsoft, в том числе SQL Server (главным образом предназначенную для работы с базами данных и решения некоторых аналитических задач) и SharePoint (обеспечивает совместную работу над проектом и решение некоторых аналитических задач).

Типы моделей

Аналитики и компании для решения аналитических задач и принятия решений на основе анализа используют множество типов моделей. Мы не собираемся учить читателей статистике, но считаем, что им было бы полезно знать, какие критерии применяют количественные аналитики, выбирая наиболее адекватную модель. Это поможет читателям сделать первые шаги в бизнес-аналитике и твердо усвоить ее основы. Если мы хотим знать, какие типы моделей лучше всего подойдут в том или ином случае, надо оценить специфику ситуации с точки зрения тех, кто принимает решения (или их аналитиков).

— Чтобы правильно выбрать модель, надо ответить на три основных вопроса.

— Сколько переменных подлежат анализу? Возможны такие варианты ответа: одна переменная (*одномерная модель*), две переменные (*двумерная модель*), три и более переменных (*многомерная модель*). Последний вариант ответа достаточен для решения любой проблемы.

— Требуется ли нам описание решения проблемы или просто ответы на поставленные вопросы? *Описательная статистика* просто описывает имеющиеся данные и не пытается делать выходящих за их рамки обобщений. Средние значения, медианы и стандартные отклонения — вот классический пример описательной статистики. Они весьма полезны, но не слишком интересны с математической или статистической точки зрения. *Индуктивная статистика* исследует выборку из какой-либо совокупности и распространяет выводы о средних характеристиках ее объектов на всю совокупность. Примеры такой статистики — корреляционный и регрессионный анализ (см. далее): они включают оценку вероятности того, что взаимосвязи, выявленные на основе выборки, характерны и для всей совокупности. Статистики и количественные аналитики обычно

отдают предпочтение индуктивной статистике по сравнению с описательной.

— Насколько точно можно оценить значения интересующих переменных? Некоторые методы оценки описаны во вставке «Методы измерения данных».

Конкретный тип используемой вами (или вашими квантами) модели зависит от того, какого вида ваш аналитический проект и какого типа данные. Некоторые характеристики проектов и массивов данных, а также моделей, выбранных для их обработки, описаны ниже. Мы рассмотрели далеко не все типы моделей, но из тех, которые изо дня в день используются организациями для аналитики, здесь представлены примерно 90 процентов.

Модели с двумя числовыми переменными. Если требуется установить взаимосвязь между двумя числовыми переменными, то проще всего это сделать с помощью *корреляционного анализа*. Это один из простейших видов статистического анализа. В типичном случае с его помощью можно установить, меняется ли одна переменная с изменением другой. Для примера возьмем рост и вес человека. Можно ли утверждать, что вес человека увеличивается с увеличением его роста? Как правило, так и бывает, поэтому можно утверждать, что эти две переменные коррелируют между собой. Поскольку корреляционный анализ является одним из методов индуктивной статистики, существуют способы определить: может ли определенный уровень корреляции быть случайным? Если вам, например, говорят, что «статистическая значимость связи равна 0,05», то это означает, что в пяти случаях из ста наблюдается согласованное изменение анализируемых показателей.

Две категориальные переменные или больше. Если вы используете данные опросов и они представлены *номинальными категориями* (например, мужской и женский пол; молодой, средний или пожилой возраст), то вам понадобится ряд аналитических процедур для анализа категориальных данных. Результаты этого вида анализа часто оформляют в виде таблицы, в ячейках которой указано количество наблюдений. Например, если вы устанавливаете связь между полом и продолжительностью жизни, то обнаружите, что численность мужчин и женщин в молодом и среднем возрасте примерно одинакова, но поскольку женщины обычно живут несколько дольше, чем мужчины,

то в старшем возрасте их численность будет выше. Если эта или подобная закономерность присутствует в вашем массиве данных, то таблица покажет значимую (то есть вряд ли случайную) взаимосвязь в соответствии со значением такого статистического критерия, как хи-квадрат. Взаимосвязь может быть значимой при уровне значимости 0,05 или 0,01. Такие бинарные категориальные переменные, как пол, можно также обрабатывать с помощью регрессионного анализа, используя при этом фиктивные переменные: то есть такие, которые получают значение 0 при отсутствии признака (например, мужского пола), и 1 при его наличии.

Более чем две количественные переменные. Если количественных переменных более двух, то проводится углубленный анализ корреляционной связи, называемый *регрессионным анализом*: иногда множественной регрессией (если для объяснения динамики одной переменной используются несколько других переменных), а иногда линейной регрессией (если взаимосвязь между переменными остается стабильной (линейной) во всех интервалах их значений). Регрессия представляет собой метод подбора уравнения (или линии, если речь идет о графическом выражении), описывающего совокупность собранных в прошлом данных. Если вам это удалось, то с помощью уравнения регрессии можно прогнозировать поведение переменных в будущем. В регрессионной модели каждой независимой переменной приписывается определенный коэффициент, отражающий (или прогнозирующий) ее «вес» в модели.

В качестве примера множественной линейной регрессии можно привести случай из практики экономиста из Принстона Орли Ашенфельтера. Он использовал регрессионный анализ для прогнозирования аукционных цен на марочные французские вина. Его прогноз аукционных цен основывался на погоде в период сбора урожая вин этого года — и вызвал шок в среде экспертов по винам и даже привел их в ярость. (Газета New York Times опубликовала на первой странице статью об этом прогнозе под названием «Уравнение цены на вино вывело из строя многие носы»*.) Если у вас есть хорошее уравнение, то зачем вам эксперты?

---

* *Passell P.* Wine Equation Puts Some Noses Out of Joint // New York Times. March 4, 1990.

Большинство экспертов сходятся в том, что хорошее вино получается в том случае, если предшествующая зима была дождливой, в сезон созревания винограда стояла теплая погода, а в сезон его сбора — сухая. Таким образом, Ашенфельтер выбрал три независимые переменные, относящиеся к погоде и влияющие на качество винограда: средняя температура воздуха в период созревания и количество осадков в период сбора винограда, а также количество осадков в предшествующую зиму. Кроме того, поскольку вкус вина, как правило, зависит от его выдержки, еще одной независимой переменной стала продолжительность выдержки в годах.

Качество сбора винограда влияет на цену зрелого вина, которая и становится зависимой переменной, которую Ашенфельтер пытался предсказывать. Он собрал информацию о ценах на лондонском аукционе за шесть бутылок бордо шато в 1960–1969 годы. Этот период был выбран потому, что вина, сделанные из урожая сборов этих лет, уже созрели, а в их качестве не было сомнений. Данные о значениях независимых переменных предоставило бюро прогнозов погоды из района выращивания винограда.

Ашенфельтер составил регрессионное уравнение логарифма цены вина, включающее показатели возраста вина и параметров погоды. Он получил такое выражение:

> Качество вина = 12,145 (константа) + 0,0238 × Возраст вина + 0,616 × Средняя температура периода созревания 0,00386 × Количество осадков в период сбора урожая + 0,0017 × Количество осадков предшествующей зимой.

Как показывают значения коэффициентов при переменных, возраст вина, умеренная температура в период созревания и количество осадков в течение предшествующей зимы оказывают прямое положительное влияние на цену вина. Осадки в период сбора урожая оказывают негативное влияние на качество вина. *Коэффициент детерминации R-квадрат* (подробнее см. во вставке «Основные статистические концепции и аналитические приемы») для этого уравнения составляет 0,828, что означает, что включенные в уравнение переменные на 83 процента объясняют отклонения в ценах на вино. Коротко говоря, эти переменные в совокупности играют определяющую роль в процессе установления цен. Легко понять, почему эксперты сочли эти результаты до некоторой степени спорными и менее интересными,

чем бесконечные разговоры о *терруаре*[*], дубовых бочках и переспевшем винограде.

Во вставке «Основные статистические концепции и аналитические приемы» мы описали наиболее часто встречающиеся индуктивные статистические модели (мы уже говорили, что описательные и ориентированные на отчеты модели полезны, но не слишком интересны с точки зрения количественного анализа). Конечно, написано множество книг на эту тему, поэтому мы сделаем только краткий обзор.

Основные статистические концепции и аналитические методы[**]

Дисперсионный анализ (ANOVA). Статистический тест на равенство средних значений двух и более групп.

Причинно-следственная связь. Взаимосвязь между двумя событиями (причиной и следствием), когда второе событие считается последствием первого. В типичном случае причинно-следственная связь — это зависимость между рядом факторов (причинами) и результирующим фактором (следствие). Наличие причинно-следственной связи требует соблюдения трех условий:

- Событие-причина должно предшествовать событию-следствию во времени и пространстве.
- При наличии причины наступает следствие.
- При отсутствии причины следствие не наступает.

Кластеризация, или кластерный анализ. Распределение результатов наблюдений (записей в базе данных) по группам (кластерам)

[*] Микроклимат в контексте виноделия, то есть совокупность природных факторов (тип почв, количество солнца, средняя температура воздуха и другие особенности местности), которые могут повлиять на качество вина (букет и даже вкус). *Прим. ред.*

[**] Определения взяты из Википедии, учебника Хайнца Кохлера (Heinz Kohler) Statistics for Business and Economics (2002), «шпаргалки» от компании Dell по аналитике (2012, табл. 6 и 8). Рекомендуем для быстрого ознакомления со статистической терминологией и методами книгу: Минько А. А. Статистика в бизнесе. Руководство менеджера и финансиста. М. : Эксмо, 2008. *Прим. ред.*

таким образом, что результаты в одной группе имеют сходные черты, в то время как результаты разных групп отличны друг от друга. Кластеризация является основной задачей интеллектуального поиска данных и стандартным приемом анализа статистических данных в самых разных областях.

Корреляция. Степень зависимости двух или более переменных друг от друга. Степень зависимости выражается коэффициентом корреляции, принимающим значения в интервале от 1,0 до –1,0.

Если коэффициент корреляции равен +1 (полная положительная корреляция), то это означает, что обе переменные пропорционально изменяются в одинаковом направлении.

Коэффициент корреляции равен 0 — между переменными нет связи.

Если коэффициент корреляции равен –1 (полная отрицательная корреляция), то это означает, что при возрастании одной переменной вторая уменьшается.

Наличие корреляции не обязательно означает, что имеется причинно-следственная связь. Иначе говоря, корреляция является необходимым, но не достаточным условием причинности.

Факторный анализ. Статистический метод, раскрывающий взаимосвязь между многими переменными или объектами. Это позволяет объединить взаимосвязанные переменные в группы, называемые факторами. Такой прием часто используется для структурирования и/или сокращения количества видов данных. Например, если исследователю предстоит проанализировать более сотни переменных, факторный анализ позволит объединить их в десяток комбинированных показателей, каждый из которых отражает динамику десятка исходных переменных.

Зависимая переменная. Переменная, значение которой неизвестно и подлежит прогнозированию или объяснению. Например, если вы хотите предсказать качество вина урожая определенного года на основе среднегодовой температуры периода созревания, количества осадков в период сбора урожая и в предшествующую зиму, то качество вина будет зависимой переменной. Иногда используются еще термины «объясняемая переменная» и «результирующий фактор».

Независимая переменная. Переменная, значение которой известно и применяется для прогнозирования или объяснения динамики зависимой переменной. Например, если вы хотите предсказать качество вина на основе исследования различных переменных (средняя температура в период созревания, количество осадков в период сбора и предыдущей зимой, возраст вина), то эти переменные и будут независимыми. Иногда их называют еще объясняющими переменными, переменными регрессии, фактор-аргументами.

Регрессия. Статистический метод, позволяющий построить уравнение для оценки неизвестного значения зависимой переменной через известные значения одной или более независимых переменных. Простая регрессия означает, что для оценки зависимой переменной используется одна независимая переменная. Множественная регрессия означает, что для прогнозирования зависимой переменной используются несколько независимых переменных. Логическая регрессия использует несколько независимых переменных для прогнозирования бинарной категориальной зависимой переменной (то есть переменной вида да/нет, за/против, покупать/не покупать).

R-квадрат (R2). Наиболее популярный показатель для оценки степени совпадения рассчитанной регрессии с данными выборки, по которой произведен расчет. R-квадрат отражает также степень изменчивости зависимой переменной по сравнению с рассчитанной линией регрессии. Его значение колеблется в интервале от 0 до 1, и если оно равно, например, 0,52, то это означает, что 52 процента вариации зависимой переменной объясняется независимыми переменными, включенными в уравнение регрессии. В общем случае чем выше значение R2, тем более адекватной считается модель.

Проверка гипотез. Системный подход к проверке исходного предположения об окружающей реальности. Он включает сопоставление исходной гипотезы или утверждения с доказательствами истинности и на этом основании принятие решения о том, следует ли признать ее истинной или ложной. Гипотезы можно разделить на два вида: нулевая гипотеза и альтернативная гипотеза. Суть нулевой гипотезы

(H0) состоит в том, что между результатами приведенных наблюдений не существует статистически значимой связи*.

Альтернативная гипотеза (Ha или H1) исходит из предположения о наличии такой связи. Проверка гипотез включает в себя сравнение эмпирически выявленных закономерностей в выборке с теоретически предполагаемыми (то есть предполагаемыми для случая, если нуль-гипотеза верна). Например, если вы хотите предсказать качество вина на основе его возраста, то нулевая гипотеза будет звучать следующим образом: «Возраст вина не влияет на его качество», в то время как альтернативная гипотеза такова: «Возраст вина существенно влияет на его качество». Данные собираются и анализируются с целью установления соответствия H0. Редкие или нестандартные результаты наблюдений (часто определяемые по p-значению ниже определенного уровня) являются показателем того, что H0 ложная; это означает, что существует статистически значимая вероятность того, что альтернативная гипотеза истинна.

P-значение. В процессе проверки гипотез p-значение показывает вероятность подтверждения данными истинности нулевой гипотезы. Невысокое p-значение указывает на небольшое количество или нестандартный характер данных, подпадающих под нулевую гипотезу, что, в свою очередь, говорит о ее ложности (отсюда можно сделать вывод, что истинна альтернативная гипотеза). При тестировании гипотез мы «отбрасываем нулевую гипотезу», если p-значение меньше, чем уровень значимости α (альфа греческого алфавита), который обычно равен 0,05 или 0,01. Если нулевая гипотеза отбрасывается, то результат считается статистически значимым.

Уровень значимости альфа (α). Уровнем значимости называется такое максимальное отношение количества нетипичных выборочных значений (выбросов) ко всему объему выборки, что нулевая гипотеза отклоняется**.

---

* Строго говоря, нулевой гипотезой может быть любое предположение о генеральной совокупности. Предположение, что между наблюдениями не существует значимой связи, только одна из возможных гипотез, которая не обязана быть нулевой. *Прим. ред.*

** Здесь авторы пытаются на пальцах дать определение критической области, то есть той части выборочного пространства, которая приводит к отклонению нулевой гипотезы. *Прим. ред.*

Иными словами, уровень значимости показывает количество нетипичных наблюдений (выборочных значений), необходимых для признания ложности нулевой гипотезы. Обычно уровень значимости задается как 5 процентов (0,05), но в ситуациях, когда предъявляются особенно строгие требования к доказательству истинности альтернативной гипотезы, этот показатель может быть задан и на более низком уровне, например 1 процент (0,01). Значение α, равное 5 процентам, означает, что для отбрасывания нулевой гипотезы как ложной достаточно наличия менее 5 процентов нетипичных данных от их общего количества (при условии истинности нулевой гипотезы). На практике это требование часто проверяется путем расчета p-значения. Если p-значение меньше, чем α, то нулевая гипотеза признается ложной, а альтернативная гипотеза — истинной.

Ошибка первого рода, или ошибка α. Эта ошибка возникает, когда нулевая гипотеза истинна, но тем не менее отбрасывается. В традиционной проверке гипотез нулевая гипотеза отбрасывается в том случае, если p-значение меньше, чем α. Таким образом, вероятность ошибочного отбрасывания нулевой гипотезы как ложной равняется α, почему эта ошибка и называется ошибкой α.

Тест (статистический критерий) χ-квадрат. Статистический тест, отражающий соответствие данных выборки определенному типу распределения. Измерение этого критерия обычно показывает расхождение между фактическим распределением событий и ожидаемым исходя из некоего заданного распределения. Наиболее часто используется для проверки соответствия фактического распределения заданному.

t-тест, или t-критерий Стьюдента. Метод статистической проверки гипотез путем проверки равенства средних значений двух выборок или проверки равенства среднего значения одной выборки некоторому заданному значению.

Изменение модели

Нетрудно понять, что ни одну модель нельзя использовать неограниченно долго. Если мир в своих основных проявлениях изменился, то очень вероятно, что и модель больше не является его адекватным

отражением. Мы уже говорили о том, насколько важны исходные допущения в моделях, а также о том, что проверять их нужно так, чтобы все заинтересованные лица знали, можно ли еще их применить (более подробно об этом поговорим в следующих главах). Достаточно сказать, что любая организация или частное лицо, использующие количественные модели, должны их регулярно пересматривать, чтобы убедиться, что они по-прежнему имеют экономический смысл и соответствуют данным. Если же это не так, то их следует модифицировать. Под словом «*регулярно*» мы имеем в виду ежегодно, если только нет причин делать это чаще.

В некоторых случаях модели следует пересматривать с еще меньшей периодичностью. Например, если на основании модели вы определяете стратегию торговли ценными бумагами, то придется пересматривать их очень часто. Владелец компании Renaissance Technologies Джеймс Симонс управляет одним из крупнейших в мире хеджевых фондов и занимается пересмотром моделей постоянно. Он приглашает на работу профессоров, хакеров, интересующихся статистикой инженеров и ученых. С момента основания в марте 1988 года материнская компания Симонса Medallion Fund, располагающая капиталом в 3,3 миллиарда долларов и продававшая все, начиная с фьючерсов на соевые бобы и до французских государственных облигаций, обеспечила ежегодную доходность в размере 35,6 процента. За полных одиннадцать лет, до декабря 1999 года, кумулятивная доходность Medallion Fund достигла ошеломляющей величины в 2478,6 процента. В 2008 году Симонс получил рекордную прибыль в сумме 2,5 миллиарда долларов, а чистая стоимость его компании достигла 8,7 миллиарда. Журнал Forbes поставил Симонса на восьмидесятое место в списке богатейших людей планеты и на двадцать девятое место в списке богатейших людей США. В 2006 году Financial Times назвала его самым умным миллиардером планеты»*.

Симонс понимал, что выгодные возможности по своей природе невелики и непостоянны. На одном из семинаров он так высказался по этому поводу: «Эффективная теория рынка права в том, что в глобальном смысле рынок действительно эффективен. Тем не менее мы видим незначительные и краткосрочные аномалии. Мы делаем прогноз. Вскоре после этого мы еще раз оцениваем ситуацию и пересматриваем прогноз, а также инвестиционный портфель. Мы тратим

* Alternative Rich List // FT.com. September 22, 2006.

на это целый день. Мы всегда считаем и пересчитываем, считаем и пересчитываем. Именно благодаря нашей активности мы и зарабатываем деньги». Чтобы сохранять позиции, Симонс еженедельно меняет свои модели.

Мир вокруг меняется, и именно способность приспосабливаться к этим изменениям сделала Симонса столь успешным бизнесменом. Он говорит: «Временной горизонт статистических прогнозов охватывает несколько лет — может быть, пять или десять. Вам приходится постоянно внедрять что-то новое, потому что рынок играет против нас. Если вы не совершенствуетесь, значит, вы становитесь хуже».

Пример аналитического мышления: модель ценообразования опционов Блэка и Шоулза

Фишер Блэк и Майрон Шоулз решили проблему ценообразования ценных бумаг*, долгое время доставлявшую неудобства инвесторам. Блэк получил степень PhD по прикладной математике в Гарвардском университете, затем работал в консалтинговой фирме Arthur D. Little, Inc. Получив степень по экономике в Чикагском университете, Шоулз недавно приступил к работе на кафедре финансов в MIT.

Терминология по ценообразованию опционов в значительной степени специализированная. *Опцион* — это ценная бумага, дающая право, но не обязывающая купить или продать определенный вид активов на установленных условиях в течение указанного времени. Цена, уплачиваемая за актив в момент исполнения опциона, называется *ценой исполнения, или страйк-ценой*. Последний день, когда возможно исполнение опциона, называется *сроком погашения*. Простейший вид опциона, часто называемый *колл-опционом*, представляет собой право на покупку обычных акций компании. *Премия за риск* — это сумма, уплачиваемая инвестором за акции или другие виды активов сверх цены аналогичных безрисковых активов.

В целом чем выше цена акций, тем больше будет цена опциона. Если цена акций намного превышает цену исполнения опциона, то опцион наверняка будет исполнен. С другой стороны, если цена акций

* *Black F. and Scholes M*. The Pricing of Options and Corporate Liabilities // Journal of Political Economy. 1973. Vol. 81, no. 3. P. 637–654; Black–Scholes // Wikipedia/ URL: http://en.wikipedia.org/wiki/Black–Scholes; The Prize in Economics 1997 // Пресс-релиз, Nobelprize.org. URL: http://nobelprize.org/nobel_prizes/economics/laureates/1997/press.html.

намного ниже цены исполнения опциона, владелец вряд ли будет его исполнять, и тогда его цена стремится к нулю. Если срок погашения опциона очень отдален во времени, то цена опциона приблизительно равна цене акций на текущий момент. Обычно цена опциона падает по мере приближения срока его погашения даже при том условии, что цена самих акций может и не изменяться. Но размер премии за риск предугадать трудно.

Определение и формулирование проблемы. Необходимое условие эффективного управления рисками, связанными с опционами и другими деривативами, это корректное установление цены на них. Предыдущие попытки разработать эффективную модель ценообразования на деривативы по целому ряду причин оказались неудачными. Возник вопрос о поиске нового метода — научно обоснованного и подкрепленного фактическими данными.

Изучение предыдущих поисков решения. Ценообразование на деривативы имеет долгую историю, начиная с 1900 года. В большинстве случаев речь шла об установлении цены на так называемые варранты (колл-опционы, выпускаемые компаниями и предоставляющие владельцу право выкупить у компании акции по определенной цене), причем методики расчета цены базировались на аналогичных формулах. Эти формулы, как правило, включали в себя один или более произвольно выбранный параметр, вследствие чего отличались неполнотой и страдали одним и тем же глубоким недостатком: отсутствием объективной методики расчета премии за риск. К сожалению, модели ценообразования на ценные бумаги в условиях равновесия рынка, которая была бы основана на адекватной методике расчета премии за риск, просто не существовало. Блэк и Шоулз впервые в истории попытались вывести формулу цены опциона исходя из условия равновесия рынка.

Моделирование (выбор переменных). Было установлено, что на цену опциона влияют пять переменных, в том числе:

- срок погашения
- спот-цена соответствующего актива (цена, по которой в данное время и в данном месте продаются реальный товар или ценные бумаги на условиях немедленной поставки)
- цена исполнения опциона

- ставка процента по безрисковым ценным бумагам
- волатильность доходности соответствующего актива (показатель, характеризующий изменчивость цены).

Отметим, что среди переменных отсутствовало отношение инвесторов к риску. Блэк и Шоулз внесли существенный вклад в развитие темы, по сути дела, показав, что нет необходимости учитывать премию за риск при установлении цены на опцион. Это не значит, что премия за риск вообще отсутствует, но ее величина уже учтена в текущей цене акций.

Сбор данных (измерения). Модель Блэка и Шоулза основана на некоторых технических допущениях и признании взаимосвязей между переменными. На этапе разработки модели никаких измерений не проводилось. Однако Блэк и Шоулз провели эмпирические тесты своей теоретической модели на большом массиве данных о колл-опционах и опубликовали результаты в статье The Pricing of Options and Corporate Liabilities*.

Анализ данных. Блэк и Шоулз вывели дифференциальное уравнение с частными производными на основе некоторых технических допущений и теоретических предположений (с использованием методов дифференциального исчисления, а не статистики). Решением этого уравнения и стала формула Блэка и Шоулза, показывающая, каким образом можно рассчитать цену колл-опциона как функцию ставки процента по безрисковым ценным бумагам, вариации цен на базовый актив и параметров опциона (страйк-цены, срока погашения и рыночной цены базового актива). Формула основана на том предположении, что чем выше текущая цена акций и ее волатильность, а также ставка процента по безрисковым ценным бумагам и чем дольше период до погашения опциона, тем выше будет его цена. Аналогично этому рассчитывается цена и других деривативов.

Результаты и необходимые меры. Блэк и Шоулз пытались опубликовать результаты своих исследований, отправив их сначала в Journal of Political Economy, но редакция отклонила статью. Будучи уверенными в ценности своих изысканий, они послали работу в журнал Review of Economics and Statistics, где ее постигла та же участь. Большинству

* *Black F. and Scholes M*. The Pricing of Options and Corporate Liabilities // Journal of Political Economy. May 1973. Vol. 81, no. 3.

экспертов мысль о том, что можно математически рассчитать цену опциона, не учитывая при этом отношение инвесторов к риску, казалась неприемлемой и слишком неординарной. Изучив развернутые высказывания нескольких знаменитых экономистов по этому поводу, Блэк и Шоулз опять отправили статью в Journal of Political Economy, и на этот раз там ее приняли. Через некоторое время профессор MIT Роберт Мертон опубликовал статью, развивавшую некоторые математические аспекты модели Блэка и Шоулза.

Несмотря на проблемы с публикацией, основные выводы статьи получили широкое распространение во всем мире среди тысяч трейдеров и инвесторов, применявших их для рутинных расчетов цены опционов. Модель проста в расчетах и подробно раскрывает взаимосвязи между всеми входящими в нее переменными. Она обеспечивает полезную аппроксимацию, особенно при анализе направленности движения цен на опционы в критических точках. Даже если результаты нельзя считать абсолютно точными, их можно использовать в качестве первого приближения, а затем уточнить.

Модель Блэка и Шоулза стала незаменимой не только при прогнозировании цен на опционы, но и при решении многих других проблем экономики. Ее можно назвать самой успешной экономической концепцией во всей экономической теории. Мертон и Шоулз в 1997 году получили Нобелевскую премию по экономике за развитие новых методов определения цены деривативов. Хотя умерший в 1995 году Блэк не смог стать нобелевским лауреатом, но его заслуги были специально отмечены Академией наук Швеции.

Пример аналитического мышления: подозрительный муж

В 1973 году в разделе «Советы читателям» газеты Dear Abby появилась такая заметка*:

> Dear Abby, в вашей колонке написано, что женщина вынашивает ребенка 266 дней. Кто вам это сказал? Я вынашивала своего ребенка десять месяцев и пять дней; в этом не может быть сомнений, поскольку я точно знаю, когда он был зачат. Мой муж — флотский офицер, и ребенок не мог

* *Larsen R. and Marx M.* An Introduction to Mathematics Statistics and Its Applications. Englewood Cliffs, NJ : Prentice-Hall, 1981. P. 159. Эту заметку впоследствии процитировали во многих учебниках по статистике и курсах лекций.

> быть зачат в другой день, поскольку я видела своего мужа всего лишь в течение часа и в следующий раз мы встретились уже после рождения ребенка. Я не пью и не гуляю с мужчинами, поэтому отцом ребенка может быть только мой муж. Пожалуйста, напечатайте опровержение этой заметки насчет 266 дней, иначе у меня будут большие неприятности.
>
> Читательница из Сан-Диего

В ответной заметке газета постаралась ободрить читательницу, но о сроках беременности было написано немного.

> Дорогая читательница! Средний период беременности действительно составляет 266 дней. В некоторых случаях дети рождаются недоношенными, а в некоторых — переношенными. В вашем случае ребенок родился переношенным.

Если бы газета уделила больше внимания количественной стороне вопроса, то в ответной заметке содержалось бы больше чисел. Последние всегда более убедительны, а в данном случае речь идет об относительно простой проблеме, связанной с теорией вероятности. Рассмотрим ее в рамках стандартного шестишагового подхода к проблеме количественного анализа.

Формулирование проблемы. В данном случае вопрос не в том, что ребенок родился переношенным, это и так понятно. Десять месяцев и пять дней — это примерно 310 дней, что существенно больше среднего срока беременности в 266 дней, о котором упоминала газета. Вопрос в том, насколько нетипичен этот случай (или какова его вероятность). Достаточно ли он нетипичен, чтобы заподозрить женщину во лжи?

Изучение предыдущих поисков решения. Мы можем с уверенностью предположить, что распределение продолжительности беременности является нормальным (то есть график распределения напоминает колокол). Вероятность того, что беременность будет продолжаться 310 дней, легко рассчитать с помощью *Z-критерия* (количество стандартных отклонений от среднего значения) для нормального распределения, что является азбукой статистических расчетов.

Моделирование (выбор переменных). Вероятность того, что беременность может длиться по крайней мере 310 дней.

Сбор данных (измерения). Имеющиеся данные позволяют сделать вывод о том, что среднее значение продолжительности беременности составляет 266 дней со стандартным отклонением 16 дней.

Анализ данных. Если средняя продолжительность беременности составляет 266 дней со стандартным отклонением 16 дней, то вероятность ее продолжительности в 10 месяцев и пять дней (300 и более дней) составляет 0,003 (если принять, что распределение нормальное).

Результаты и необходимые меры. Это значит, что три ребенка из тысячи рождаются более чем через 300 дней после зачатия. Казалось бы, вероятность очень невелика, но только не в случае больших чисел. В Америке ежегодно рождается около четырех миллионов детей. Соответственно, около двенадцати тысяч из них рождаются с таким большим опозданием. Видимо, Dear Abby стоило написать нечто вроде следующего: «Каждый год в США со столь большим запозданием рождаются примерно двенадцать тысяч детей, и одним из них стал ваш ребенок». Это успокоило бы не только читательницу, но и ее мужа.

В статистическом тестировании гипотез рассчитанное выше значение вероятности 0,003 называется *p*-значением, что равно вероятности получения данного значения критерия (в данном случае *Z*-значения, равного 2,75) в предположении, что нулевая гипотеза истинна. В данном случае нулевая гипотеза (H_0) звучит следующим образом: «Отцом ребенка является муж». В стандартной методике проверки гипотез нулевая гипотеза отбрасывается как ложная, если *p*-значение меньше уровня значимости. В данном случае *p*-значение равно 0,003, а это значит, что нулевая гипотеза будет отброшена, даже если уровень значимости составит 1 процент. Вообще говоря, мы должны были бы отбросить гипотезу об отцовстве мужа читательницы. Как можно объяснить этот ошибочный результат проверки гипотезы? Это типичный пример ошибки первого вида (или ошибки альфа), когда отклоняется нулевая гипотеза (H_0) при ее истинности. Этот пример показывает, что жизнь может не укладываться в рамки теории вероятности.