ГЛАВА 5

Анализ данных

Если достаточно долго мучить данные, они признаются [в чем угодно].

Рональд Коуз[1]

Следующие три главы посвящены сути аналитической работы: непосредственно анализу данных, целям анализа с позиции компании и тому, как проводить *результативный* анализ данных.

Мы рассмотрим такие аспекты, как виды анализа данных, разработка показателей, извлечение практических выводов, презентация этих выводов, идей и рекомендаций руководителям. В главе 6 мы обсудим разработку показателей и ключевых показателей эффективности деятельности (KPI), а глава 7 посвящена визуализации данных и сторителлингу[2]. В этой главе, первой из трех, речь пойдет непосредственно об анализе данных.

Важно отметить, что мы не будем говорить о том, *как* проводить анализ или статистическое исследование, — на эту тему есть много других более полных источников (см. список дополнительной литературы). Мы сосредоточимся на цели анализа данных: что это означает? К какому результату стремятся аналитики? Какие инструменты входят в их профессиональный набор? Мы вернемся к идее разных уровней аналитики, о которой уже упоминалось в главе 1, и изучим другие точки зрения на виды аналитики.

Наша цель — выделить ряд инструментов статистики и визуализации, которые аналитики могут использовать в своей работе. Дополнительная

[1] Рональд Коуз (1910–2013) — американский экономист, лауреат Нобелевской премии по экономике. *Прим. перев.*

[2] Сторителлинг (от *англ.* storytelling) — маркетинговый прием, использующий медиапотенциал с целью передачи информации и транслирование смыслов посредством рассказывания историй. *Прим. перев.*

цель заключается в том, чтобы стимулировать их применять подходящие инструменты, а при необходимости изучить более сложные инструменты, способные обеспечить более глубокий уровень понимания конкретной проблемы.

Для изготовления деревянного стола опытному столяру требуется качественный исходный материал: древесина красного дерева, набор столярных инструментов, например стамеска и угольник, и профессиональные знания, когда и как пользоваться этими инструментами. Отсутствие хотя бы одного из трех компонентов заметно скажется на качестве конечного продукта. То же самое касается и аналитической работы. Для производства аналитического продукта, имеющего реальную ценность, не обойтись без исходного материала в виде качественных данных, инструментария в формате различных аналитических методов и техник, а также профессиональных знаний, когда и как пользоваться всеми этими инструментами для решения задачи.

Что такое анализ данных?

Уделим немного времени самому термину «анализ». Он происходит от древнегреческого ἀνά [ana] + λύω [luō], что означает «освобождать», «распутывать». В этом есть смысл, но слишком высокопарный, чтобы помочь нам уловить, что это действительно означает. Для целей бизнеса можно воспользоваться определением Марио Фариа из главы 1:

> *Анализ — преобразование данных в выводы, на основе которых будут приниматься решения и строиться действия с помощью людей, процессов и технологий.*

Давайте остановимся на этом подробнее. Надеюсь, из глав 2 и 3 у вас уже сложилось понимание, что такое массив данных, а вот что такое аналитические выводы?

Согласно «Википедии», аналитические выводы — понимание конкретных причин и следствий в конкретном контексте[1]. В английском языке у этого термина (insight) есть несколько сопутствующих значений:

— информация;

— «озарение» — понимание внутренней сути вещей и процессов;

— самоанализ;

[1] URL: https://en.wikipedia.org/wiki/Insight.

— проницательность, способность делать глубокие наблюдения и выводы;

— понимание причин и следствий на основе установления взаимосвязи и поведения в рамках модели, контекста или сценария.

Итак, понимание взаимосвязи причин и следствий, понимание внутренней природы вещей и процессов и так далее. Это будет нам полезно.

Термин «информация»[1], то есть «результат обработки данных для придания им контекста и смысла», часто используется как синоним термина «данные», хотя технически это не одно и то же (см. ниже текст в рамке, а также статью The Differences Between Data, Information and Knowledge («Разница между понятиями “информация”, “данные” и “знания”»)[2].

ДАННЫЕ, ИНФОРМАЦИЯ И ЗНАНИЯ

Данные представляют собой сырые, необработанные факты об окружающем мире. Информация — собранные, обработанные данные, в то время как знания — это набор ментальных моделей и убеждений об окружающем мире, который сформировался на основе информации, полученной на протяжении какого-то периода времени.

Температура на данный момент составляет 6 °C. Это количественный факт. Он существует и соответствует действительности вне зависимости от того, зафиксировал ли его кто-то. К сожалению, этот факт бесполезен (для всех, кроме меня), так как из-за отсутствия контекста (когда? где?) он не позволяет сделать никаких выводов.

В Нью-Йорке 2 ноября 2014 года в 10 утра температура составила 6 °C. У этих данных есть контекст. Однако это по-прежнему лишь констатация факта без интерпретации.

Температура 6 °C гораздо ниже климатической нормы. Это информация. Мы обработали данные и объединили их с другими данными, чтобы определить понятие климатической нормы и оценить, как соотносятся значения.

При температуре 6 °C на улице прохладно, я надену пальто. Вы объединили информацию за какой-то период времени и построили мыслительную модель, что это означает. Это знания. Конечно, все эти модели относительны. Например, житель Аляски может посчитать температуру 6 °C в ноябре не по сезону теплой.

[1] URL: http://foldoc.org/information.

[2] URL: http://www.infogineering.net/data-information-knowledge.htm.

Исходя из глубины информации, мы вновь можем вернуться к подробному определению анализа (рис. 5.1). Хотя в нем по-прежнему остаются такие термины, как «понимание» и «контекст», надеюсь, теперь у вас более четкое представление о том, что такое анализ, по крайней мере концептуально. На этом новом уровне понимания давайте изучим набор инструментов, находящийся в распоряжении аналитиков. Сейчас речь идет не о программных инструментах, таких как Excel или R, а о статистических инструментах и о *видах* анализа данных, которые можно проводить.

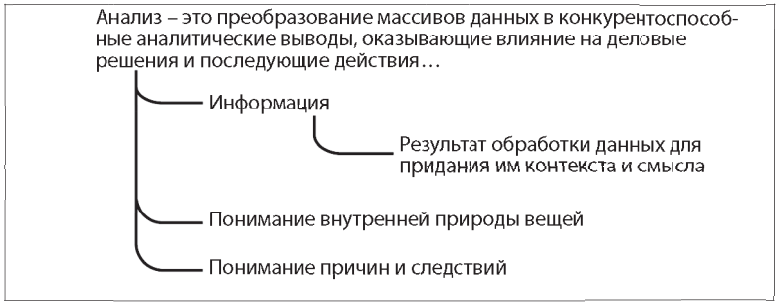

Рис. 5.1. Результат двухуровневого раскладывания определения термина «анализ»

Виды анализа данных

Джеффри Лик, старший преподаватель биостатистики в Университете Джонса Хопкинса, а также один из редакторов блога о статистике[1], выделяет шесть типов анализа данных[2]. Они перечислены далее от простого к сложному:

— описательный (descriptive);

— разведочный (exploratory);

— индуктивный (inferential);

— прогностический (predictive);

— каузальный (причинно-следственный) (causal);

— механистический (mechanistic).

[1] URL: http://simplystatistics.org/.

[2] По крайней мере, он рассматривает эти шесть типов анализа данных в рамках своего курса Data Analysis Course.

Мы рассмотрим первые пять типов анализа. Механистический тип в большей степени связан с фундаментальной наукой, исследованиями и разработками, и к нему больше подходит термин «моделирование», чем «анализ». Механистическое моделирование и анализ отличаются очень глубоким пониманием системы, которое приходит в результате многолетнего контролируемого изучения стабильной системы посредством большого числа экспериментов. Именно на этом основана моя ассоциация с фундаментальной наукой. Это редкость для большинства компаний, за некоторыми исключениями, такими как научно-исследовательские подразделения фармацевтических компаний и инженерно-проектные подразделения технических компаний. Иными словами, если вы проводите анализ данных на этом уровне, который представляет собой вершину анализа, то практически наверняка вам не требуется читать в этой книге, как его выполнять. Если вернуться к главе 1, то сейчас у вас должен прозвучать звоночек. Ранее мы говорили о восьми уровнях *аналитики*. Сейчас мы говорим о шести типах *анализа* данных, при этом у нас встретилось всего одно общее слово — «прогностический». Что все это значит?

В предыдущем списке перечислены типы статистического анализа. Важно отметить, что они могут относиться к разным уровням аналитики. Например, на основе разведочного анализа данных (о котором шла речь в главе 2) можно подготовить ad hoc отчет (уровень аналитики 2). Также на его основе можно сформулировать бизнес-логику для системы оповещения (уровень аналитики 4), например определить 98-й процентиль в распределении и установить сигнал оповещения, если соответствующий показатель превысит этот уровень.

На рис. 5.2 показана попытка соотнести эти два списка: уровни аналитики (по вертикали) и пять типов анализа данных (по горизонтали). Интенсивность цвета каждой ячейки обозначает примерную оценку усилий или времени, затраченных на проведение этого типа анализа. Например, подготовка стандартных отчетов обычно осуществляется на основе описательного и разведочного типов анализа, при этом крайне маловероятно использование причинно-следственных моделей. С другой стороны, аналитика оптимизации строится на описательном и разведочном анализе, но в первую очередь сосредоточена на прогностическом и, возможно, причинно-следственном анализе.

Необходимо прояснить один момент. Существует множество других видов количественного анализа, например анализ выживаемости, анализ социальных сетей, анализ временны́х рядов. При этом каждый из них связан с конкретной областью профессиональных знаний или типом данных, а применяемые аналитические инструменты и подходы *включают*

в себя шесть более общих аналитических инструментов и подходов. Например, при анализе на основе временны́х рядов можно вычислить период действия явления (описательный анализ), затем определить переменную во времени (разведочный анализ) и, наконец, смоделировать и прогнозировать будущие показатели (прогностический анализ). Вы получаете общую картину. Иными словами, перечисленные шесть классов представляют собой архетипы анализа. Кроме того, есть другие типы качественного анализа. Например, анализ основных причин, метод «Пять “почему”» от Tovota[1] и методология «Шесть сигм». Принимая это во вн лиза.

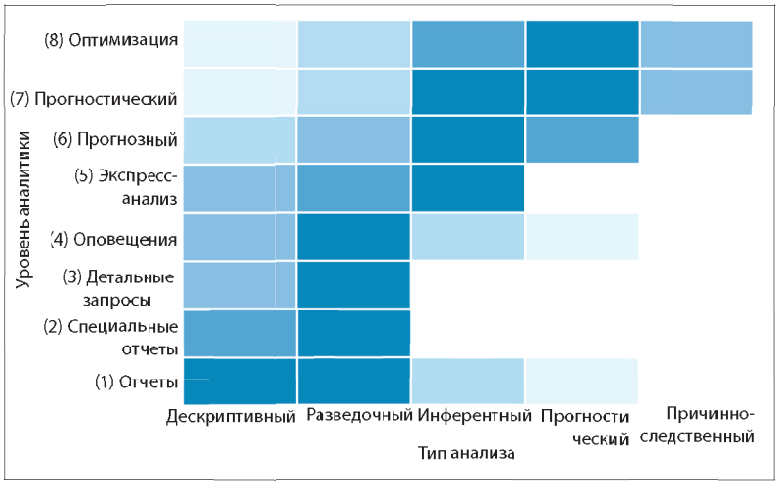

Рис. 5.2. Примерное соотношение между уровнем аналитики (по вертикали) и типом анализа (по горизонтали). Объяснение см. в тексте

СЛОВАРЬ ТЕРМИНОВ

Вы еще не запутались во всех этих «показателях», «переменных», «значениях»? Не переживайте. Эти термины пересекаются, и насчет их определении нет согласия. Ниже представлены мои варианты.

Переменная (Variable)

Показатель, который склонен меняться со временем, пространством или единицами выборки. Например, «Допустим, переменная v = скорость движения автомобиля» или «Пол — категориальная переменная».

[1] URL: https://en.wikipedia.org/wiki/5_Whys.

Измерение (Dimension)

Это переменная. В то время как термин «переменная» чаще используют ученые и программисты, для представителей деловых кругов больше характерно употребление термина «измерение». Измерение — переменная, характеризующая факты и количественные показатели, она может отражать параметр категории или времени, а также рейтинга, рэнкинга или числа. Например, вы можете проанализировать совокупный объем продаж (значение) относительно страны (измерение) или года (измерение) или же рассчитать процент отказов (значение) относительно пола (измерение). В моем представлении измерения, как правило, находятся на оси *x*, а показатели — на оси *y*.

Значение (Measure)

Количественный показатель какого-либо свойства объекта, например длина, или стандартная единица измерения. В области бизнес-аналитики этот термин обычно относится к функции (например, BMI) или агрегированному значению, например минимальное, суммарное или среднее значение количественных данных. Может рассматриваться в виде чистого или производного значения чего-либо.

Показатель (Metric)

Функция от двух или более значений (с точки зрения измерения) или просто значение (в функциональном смысле). Производное значение.

Статистический показатель (Statistic)

Определенный показатель какого-то свойства в выборке значений, например среднее арифметическое = 6,3. Это функция, примененная к набору числовых данных, которая представляет собой отдельное значение. Несколько сбивает с толку, что и сама функция, и итоговое ее значение — статистические показатели.

Ключевые показатели эффективности деятельности (Key performance indicator)

В контексте ведения бизнеса этот показатель связан с целью деятельности и/или некоторыми основными ценностями (подробнее о KPI мы поговорим в следующей главе). То есть этот показатель связан с целью бизнеса или стартовой точкой.

ОПИСАТЕЛЬНЫЙ АНАЛИЗ

Наиболее простой тип анализа данных — описательный (дескриптивный). Он обеспечивает количественное описание набора данных. Важно отметить, что этот тип анализа касается только выборки данных, по которой проводится анализ, и не описывает ту совокупность, из которой он взят. На основании описательного анализа часто формируются данные, которые отображаются в дашбордах, например количество новых пользователей за неделю или размещенных заказов с начала года (см. раздел «Дашборды» в главе 7).

Давайте начнем с одномерного анализа, то есть описывающего одну переменную (ряд или поле) из набора данных. В главе 2 мы уже обсуждали составление пятичисловой сводки, однако есть множество других возможных статистических показателей; их можно условно разделить на меры среднего уровня («середина» данных), меры рассеивания (разброса данных) и формы распределения. Ниже перечислены **показатели, относящиеся к числу простейших**, но при этом наиболее важных.

Размер выборки

Количество единиц (записей) в выборке данных.

Далее перечислены **меры среднего уровня**.

Среднее значение

Чтобы найти среднее арифметическое, нужно сложить все значения и разделить на их количество.

Среднее геометрическое

Этот показатель применяется для определения среднего значения при наличии мультипликативного эффекта, например сложных процентов со ставкой, меняющейся из года в год. Чтобы найти среднее геометрическое, нужно перемножить все значения и извлечь из них корень. Степень корня определяется количеством значений. Если вы получили 8% в первый год, а затем по 6% следующие три года, средняя процентная ставка составит 6,5%.

Среднее гармоническое

Средним гармоническим называется число, обратное среднему арифметическому их обратных. Например, если вы доехали

до магазина со скоростью движения 80 км/ч, а на обратной дороге попали в пробку и скорость вашего движения составила 32 км/ч, ваша средняя скорость составит не 56, а 47 км/ч.

Медиана

Медиана — 50-й процентиль.

Мода

Наиболее часто встречающееся значение.

К **мерам рассеяния** относятся следующие.

Минимум

Наименьшее значение в выборке (0-й процентиль).

Q1

25-й процентиль. Значение выборки такое, что одна четвертая остальных значений выборки меньше него.

Q3

75-й процентиль. Значение выборки такое, что одна четвертая остальных значений выборки больше него.

Максимум

Максимальное значение в выборке (100-й процентиль).

Межквартильный размах

Центральные 50% данных, разность между третьим и первым квартилями.

Размах

Разница между максимумом и минимумом.

Стандартное отклонение

Наиболее распространенный показатель рассеивания значений случайной величины относительно ее математического ожидания. Вычисляется как квадратный корень из дисперсии. Измеряется в тех же единицах, что и сама случайная величина.

Дисперсия

Мера разброса значений случайной величины относительно ее математического ожидания. Вычисляется возведением стандартного отклонения в квадрат. Измеряется в квадратах единицы измерения случайной величины.

Стандартная ошибка

Вычисляется путем деления стандартного отклонения на квадратный корень размера выборки. Показывает ожидаемое стандартное отклонение среднего значения выборки, если бы мы повторно получали выборки такого же размера из того же источника генеральной совокупности.

Коэффициент Джини

Количественный показатель, изначально разработанный, чтобы показать степень неравенства при распределении доходов. Тем не менее его можно использовать более широко. Он равен половине ожидаемой абсолютной разницы между доходами двух случайно выбранных людей, деленной на средний доход.

Меры формы включают следующие.

Коэффициент асимметрии

Величина, характеризующая асимметрию распределения. Коэффициент асимметрии положителен, если правый хвост распределения длиннее левого, и отрицателен в противном случае. Число фолловеров среди пользователей сервиса Twitter характеризуется положительным коэффициентом асимметрии (см., например, отчет An In-Depth Look at the 5% of Most Active Users[1] и статью Tweets loud and quiet[2]).

Коэффициент эксцесса

Мера остроты пика распределения случайной величины. У распределения с высоким коэффициентом эксцесса[3] острый пик и плоские хвосты. На это стоит обратить внимание при инвестировании, так как это означает вероятность более резких колебаний по сравнению с переменной с нормальным распределением.

---

[1] URL: https://www.sysomos.com/2009/08/05/exploring-twitters-most-active-users/.

[2] URL: https://www.oreilly.com/ideas/tweets-loud-and-quiet.

[3] URL: https://en.wikipedia.org/wiki/Kurtosis.

Кроме того, мне кажется, что тип распределения также можно назвать полезной описательной статистикой. Например, нормальное распределение (распределение Гаусса), логарифмически нормальное распределение, экспоненциальное распределение и унимодальное распределение — обычные. Зная тип, а следовательно, и форму распределения, можно узнать его потенциальные характеристики (например, что в нем могут быть редкие, но сильно отклоняющиеся значения), понять логику процесса генерации данных, а также определить, какие еще показатели требуется собрать. Например, если распределение представляет собой ту или иную форму экспоненциального закона, как распределение фолловеров в Twitter, очевидно, что следует вычислить отрицательный показатель экспоненты, который представляет собой важный критерий.

Не все переменные — непрерывные. Например, пол и продуктовая линейка относятся к категориальным переменным. Таким образом, описательный анализ может включать таблицы частотности для разных категорий или факторные таблицы, подобные следующей.

Объем продаж по регионам					
Пол	**Западный**	**Южный**	**Центральный**	**Восточный**	**Итого**
Мужской	3485	1393	6371	11 435	22 684
Женский	6745	1546	8625	15 721	32 637
Итого	10 230	2939	14 996	27 156	**55 321**

На этом уровне анализа проводящий его специалист должен знать, по какому критерию следует группировать данные, и понимать, когда какие-то данные выделяются из общей массы и представляют интерес. Например, в предыдущей таблице интересно, почему настолько велика доля женщин, совершающих покупки, в западном регионе.

При работе с двумя переменными описательный анализ может включать меры ассоциации, например вычисление коэффициентов корреляции и ковариации.

Цель описательного анализа состоит в числовом описании основных характеристик выборки. Он должен прояснять основные значения, отражающие распределение данных, кроме того, он может описывать взаимоотношения между переменными с показателями, описывающими ассоциации, или в сводных таблицах.

Некоторые из этих простых показателей могут оказаться весьма ценными сами по себе. Возможно, вам потребуется узнать и отследить среднее число заказов или наибольшую длительность их выполнения для

разрешения практического вопроса с клиентом. Таким образом, этих данных может быть достаточно для составления стандартного и ad hoc отчетов, запроса или оповещения (уровни аналитики 1–4), и это может принести пользу компании. Кроме того, вы можете убедиться в качестве данных. Например, если максимальный возраст игрока, который зарегистрировался на сайте игры — «стрелялки» от первого лица, указан как 115 лет, то либо пользователь ошибся при вводе этой информации, либо в графе с датой рождения была установлена дата по умолчанию 1900 (ну, или это *реально* крутая бабушка). Помочь это определить могут простые минимум и максимум, размах выборки и гистограммы.

Наконец, описательный анализ обычно бывает первым шагом — возможностью познакомиться с данными — к более глубокому анализу.

РАЗВЕДОЧНЫЙ АНАЛИЗ

Описательный анализ — важный первый шаг. При этом просто итоговых цифр может быть недостаточно. Одна из проблем заключается в том, что большое число значений сводится к нескольким итоговым цифрам. А потому не стоит удивляться, что одни и те же итоговые статистические показатели могут описывать разные выборки с разным распределением данных, формами и свойствами.

На рис. 5.3 представлены две выборки с одинаковым средним значением, равным 100, но очень разным распределением.

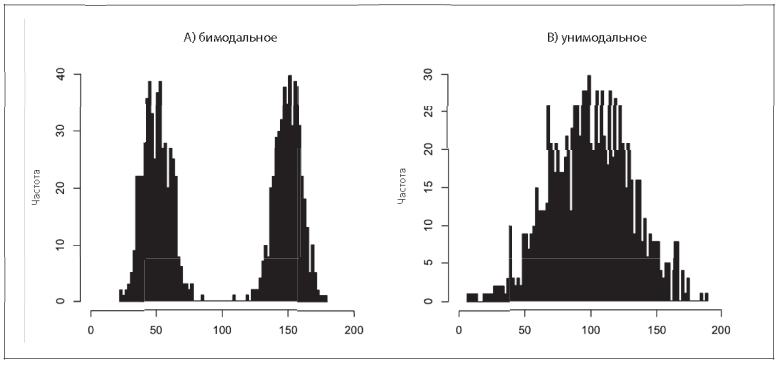

Рис. 5.3. A) бимодальное распределение и B) унимодальное распределение. В обоих случаях среднее значение одинаковое, примерно равно 100

Теперь это кажется не таким удивительным. У нас имеется простой итоговый статистический показатель — среднее значение одной

переменной. Существует множество потенциальных «решений», или выборок, которым может соответствовать это значение.

Сейчас я покажу вам гораздо более удивительный пример. Предположим, у вас четыре набора данных с двумя переменными со следующими характеристиками.

Характеристика	**Значение**
Размер выборки в каждом случае	11
Среднее значение переменной *x* в каждом случае	9
Дисперсия переменной *x* в каждом случае	11
Среднее значение переменной *y* в каждом случае	7,5
Дисперсия переменной *y* в каждом случае	4,122 или 4,127
Корреляция между *x* и *y* в каждом случае	0,816
Прямая линейной регрессии в каждом случае	$y = 3{,}00 + 0{,}500x$

Это система с жесткими заданными ограничениями. Значит, графики этих четырех наборов данных с идентичными статистическими характеристиками должны быть достаточно похожими, не так ли? А вот рис. 5.4 показывает, что это далеко не так.

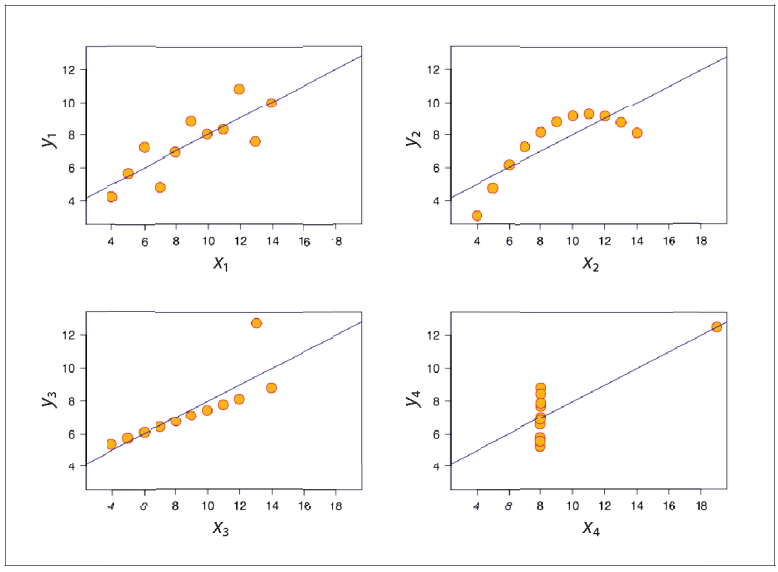

Рис. 5.4. Квартет Энскомба. В каждом из четырех наборов данных идентичны среднее значение x, среднее значение y, дисперсия x, дисперсия y, корреляция и прямая линейной регрессии (до двух знаков после запятой)

Источник: https://en.wikipedia.org/wiki/Anscombe's_quartet

Это так называемый квартет Энскомба[1], названный по имени математика и статистика Фрэнсиса Энскомба, который составил его в 1973 году. Энскомб выступил против существовавшей на тот момент доктрины в области статистических вычислений, которая гласила, что:

1) числовые данные точные, а графики — приблизительные;
2) для каждого конкретного вида статистических данных существует только один набор вычислений, обеспечивающий правильный статистический анализ;
3) выполнение сложных расчетов — единственно верный путь, изучение данных только вводит в заблуждение.

Энскомб утверждал:

> Большинство статистических вычислений строятся на предположениях относительно поведения данных. Эти предположения могут оказаться неверными, и тогда результаты вычислений тоже будут содержать ошибку. Всегда следует пытаться проверять, являются ли предположения верными. А если они ошибочны, мы должны быть способны понять, что с ними не так. В этом весьма полезны графики.

Применение графиков для визуализации и изучения данных получило название разведочного анализа данных. Наибольшую известность он приобрел благодаря продвижению американским математиком Джоном Тьюки в книге Exploratory Data Analysis (Pearson), опубликованной в 1977 году. При правильном подходе графики помогают видеть более масштабную картину, а также отмечать очевидные или необычные закономерности (это врожденное свойство человеческого мозга). Нередко аналитические выводы и понимание данных начинают формироваться именно на этом этапе. Почему у этой кривой такое отклонение? В какой момент наступает снижение возврата на маркетинговые расходы?

Разведочный анализ позволяет опровергнуть или подтвердить наши предположения относительно данных. Поэтому, когда в главе 2 шла речь о качестве данных, я рекомендовал использовать команду `pairs()` в среде R. Часто у нас сформированы обоснованные ожидания, что может быть не так с качеством данных, в отличие от ожиданий, какими должны быть достоверные данные.

[1] Anscombe F. J. Graphs in statistical analysis, American Statistician 27 (1973): 17–21.

По мере того как мы набираемся опыта и знаний в профессиональной области, у нас развивается интуитивное понимание, какие факторы и возможные отношения могут быть задействованы. Разведочный анализ, с его широким набором способов рассмотреть данные и их взаимоотношения, предлагает набор «луп» для изучения системы.

Это, в свою очередь, помогает специалисту по анализу данных выдвинуть новые гипотезы относительно того, что может произойти, если вы понимаете, какие переменные находятся под вашим контролем и какими рычагами вы можете воспользоваться для движения показателей, например выручки или конверсии, в нужном направлении. Кроме того, разведочный анализ способен показать пробелы в наших знаниях и определить, что можно сделать для их ликвидации.

Для одномерных непрерывных (действительные числа) или дискретных данных (целые числа) обычно строят диаграмму «стебель-листья» (рис. 5.5), гистограммы (рис. 5.6) и диаграммы размаха, или коробчатые диаграммы (рис. 5.7).

Стебель	Листья
1	0 3 6
2	1 6 7 8
3	5 5 6
4	1 1 5 6 9
5	0 3 6 8

Рис. 5.5. Диаграмма «стебель-листья»

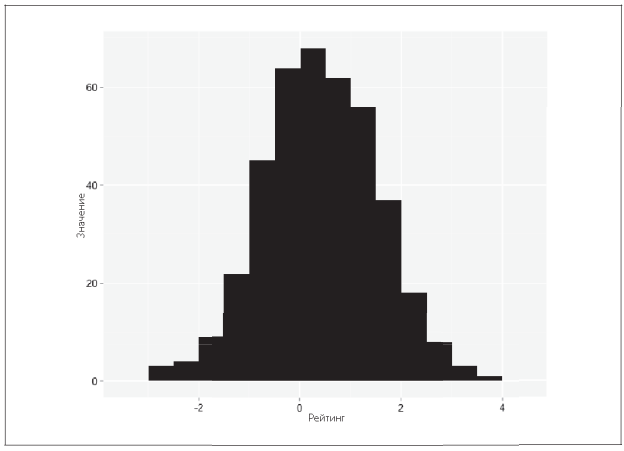

Рис. 5.6. Гистограмма

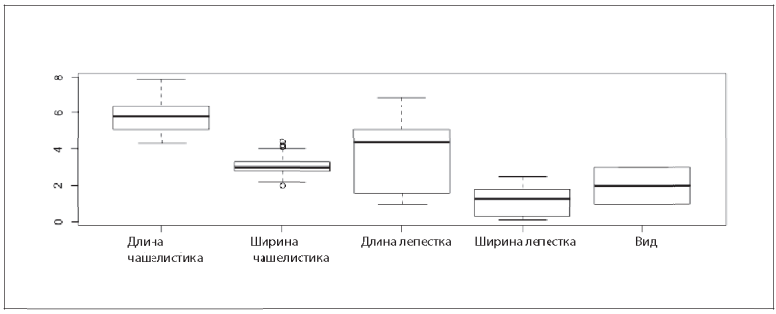

Рис. 5.7. Коробчатая диаграмма

Если гистограмма строится в таком масштабе, что ее площадь равна 1, это функция плотности распределения вероятностей.

Еще один полезный способ представить те же самые данные — составить интегральную функцию распределения.

Это может выделить интересные точки распределения, включая основные опорные точки.

На рис. 5.8, 5.9, 5.10 представлены основные графики для одномерных категориальных (качественных) переменных.

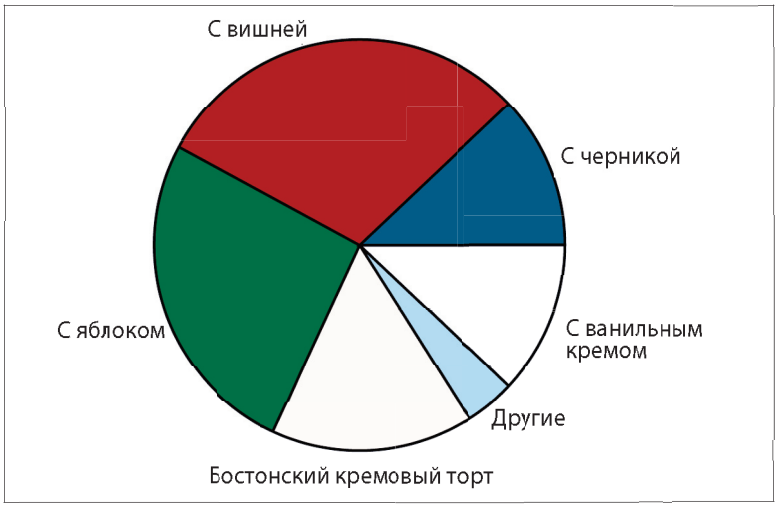

Рис. 5.8. Круговая диаграмма

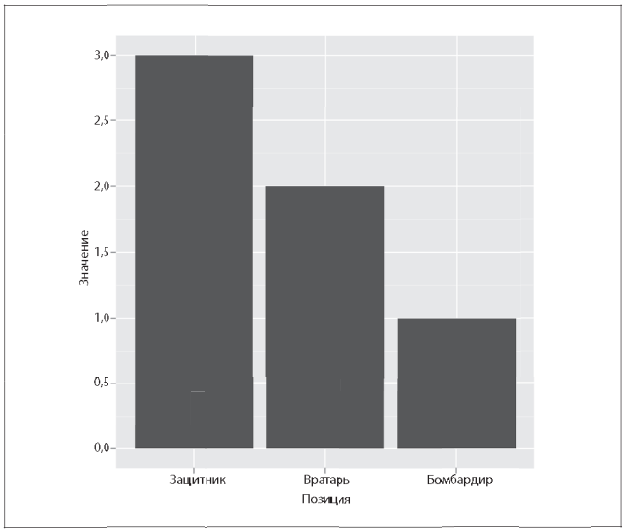

Рис. 5.9. Столбиковая диаграмма

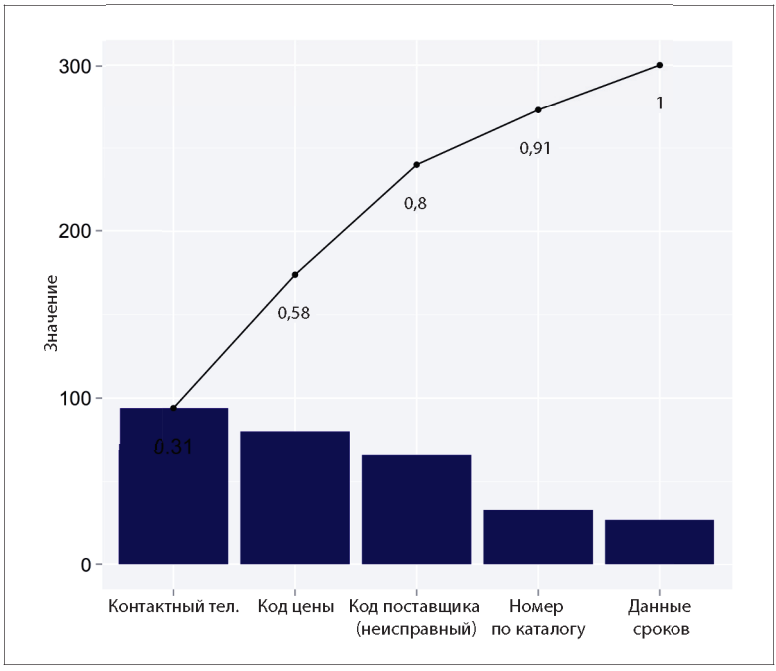

Рис. 5.10. Диаграмма Парето

Для визуализации двух переменных можно воспользоваться разными типами графиков.

	Непрерывная или дискретная переменная	**Категориальная переменная**
Категориальная переменная	Коробчатая диаграмма (box plot) Диаграмма с областями (area chart) Гистограмма распределения (range chart) Таблица (table chart)	Лепестковая диаграмма (spider/radar chart) Составная столбиковая диаграмма (stacked bar chart) Воронкообразный график (funnel chart)
Непрерывная или дискретная переменная	Диаграмма рассеяния (scatter plot) Линейный график (line graph) Карты и диаграмма Вороного (maps & Voronoi diagram) График плотности (density plot) Контурная диаграмма (contour plot)	Такие же, как в левом верхнем углу

(См. также рис. 7.5.)

Есть целый набор графиков для одновременного изучения трех переменных. Некоторые из них более общие и привычные (график поверхности (surface), пузырьковая диаграмма (bubble plots), 3D-диаграмма рассеивания (3D scatter)), а некоторые применяются для особых целей (см. the D3 gallery[1]).

В случае, когда одна из переменных — время (например, годы) или категориальная переменная, также можно использовать подход небольших множеств (small multiples), при котором создается решетка из одномерных или двумерных графиков (рис. 5.11).

Не ограничивайтесь использованием одного или двух типов диаграмм. Каждый из этих типов диаграмм выполняет свою задачу. Изучите их преимущества и недостатки и применяйте те из них, которые лучше всего отражают интересные сигналы, тренды или образцы. (Мы еще вернемся к некоторым из этих аспектов в главе 7.)

[1] URL: https://github.com/d3/d3/wiki/Gallery.

Там, где возможно, пользуйтесь командами, например `pairs()`, при автоматическом создании графиков и диаграмм для различных комбинаций переменных, которые вы можете быстро просмотреть в поисках интересных деталей или странностей, заслуживающих дополнительного внимания.

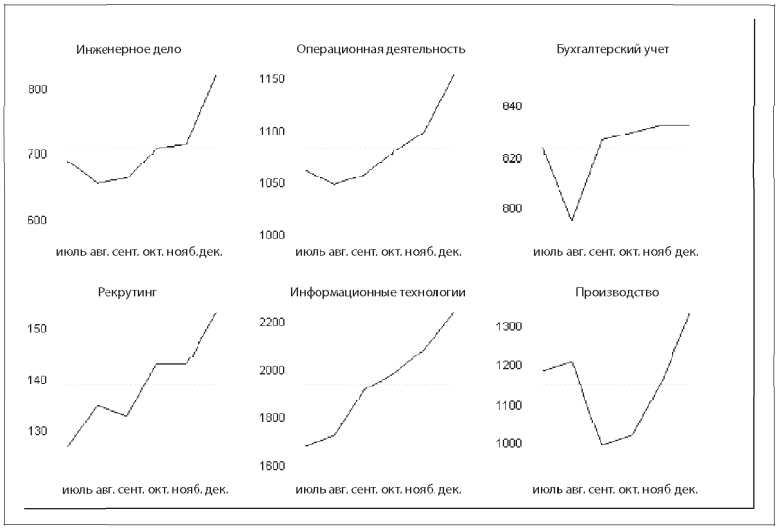

Рис. 5.11. Пример маленьких множеств

Источник: https://en.wikipedia.org/wiki/Small_multiple

ИНДУКТИВНЫЙ АНАЛИЗ

Описательный и разведочный виды анализа выступают под широкой зонтичной структурой описательной статистики: они *описывают* характеристики предлагаемого набора данных. Далее мы перейдем к другому основному направлению — статистическим исследованиям. Их цель заключается в *логическом извлечении* информации (параметры, распределение или взаимосвязи) о более широкой генеральной совокупности, из которой был взят набор данных. Кроме того, они обеспечивают основу для тестирования гипотез, на основе которых можно разрабатывать и проводить эксперименты для анализа нашего понимания внутренних механизмов и процессов.

Поскольку наша книга не учебник по статистике, в этом разделе мы лишь поверхностно проведем обзор вопросов, которые могут возникнуть, типов практических выводов, которые можно сформулировать, а также дополнительной ценности, которую можно получить благодаря

применению индуктивного анализа. Если вам требуется более подробная вводная информация по теме, настоятельно рекомендую ознакомиться с бесплатным ресурсом OpenIntro Statistics[1].

Зачем нужны статистические выводы? Как правило, мы делаем выводы обо всей генеральной совокупности на основе взятой из нее выборки, так как полный сбор данных бывает слишком дорогим, непрактичным, а иногда и просто невозможным. Возьмем, например, опрос граждан на выходе с избирательных участков, так называемый экзитпол. Невозможно опросить 125 млн избирателей, но вместо этого можно постараться получить качественную репрезентативную выборку и сделать точное умозаключение, каким мог быть результат, если бы были опрошены все избиратели. Также если вы обеспечиваете проверку качества производимой продукции и проводите испытания с *разрушением* опытного образца, очевидно, что вы не сможете протестировать подобным образом абсолютно всю продукцию, иначе вам просто нечего будет продавать.

Еще одна причина применения индуктивного анализа заключается в обеспечении объективности оценки расхождений и результатов. Предположим, вы решили провести кампанию для поощрения лояльности своих клиентов[2] и выбрали тысячу клиентов на основе общего критерия: например, каждый из них совершил не менее двух покупок за прошедший год и участвует в программе лояльности. Половине из отобранных клиентов (тестовая группа) вы отослали небольшой подарок с сообщением: «Просто потому, что мы любим своих клиентов, мы хотим преподнести вам этот скромный подарок». Вторая половина из отобранных клиентов (контрольная группа) не получила ничего. В течение следующих трех месяцев вы оцениваете число совершённых покупок, и описательный анализ показывает, что участники тестовой группы ежемесячно тратят на покупки в среднем на 3,36 долл. больше, чем участники контрольной группы. Что это означает? Очевидно, что это хорошо, но насколько надежны эти цифры? Получили бы мы похожий результат при повторном проведении эксперимента, или это просто случайность? Может быть, все объясняется тем, что один покупатель сделал крупный заказ? Статистические выводы позволяют оценить вероятность того, что это повышение покупательского спроса было просто случайностью, если при этом не наблюдалось реальных изменений внутренних образцов покупательского поведения.

---

[1] URL: https://www.openintro.org/stat/textbook.php.

[2] URL: http://brainsonfire.com/2013/02/12/7-awesome-examples-of-surprise-and-delight-that-will-blow-your-mind/.

Представьте, что вы отчитываетесь о результатах перед руководителем. На основе описательного анализа вы можете только констатировать результат: «Мы обнаружили разницу в объеме 3,36 долл./месяц, вектор движения правильный, и, кажется, это результаты кампании по поощрению лояльности клиентов». Однако на основе индуктивного анализа ваши выводы могут быть более убедительными: «Мы обнаружили разницу в объеме 3,36 долл./месяц, и вероятность того, что мы получили бы подобный результат без реального изменения в поведении покупателей, составляет всего 2,3%. Данные убедительно свидетельствуют, что это эффект от проведения кампании по поощрению лояльности клиентов». Или наоборот: «Мы обнаружили разницу, но при этом вероятность того, что этот результат случаен, составляет 27%. Вероятнее всего, кампания не была эффективной, по крайней мере, для данного конкретного показателя». Как с позиции аналитика, так и с позиции руководителя можно утверждать, что индуктивный анализ имеет бóльшую ценность и оказывает более значительное влияние на деятельность компании.

Статистические выводы обеспечивают ответы на приведенные ниже типы вопросов (но не ограничиваются ими).

Стандартная ошибка, доверительный интервал, статистическая погрешность

Насколько можно быть уверенным в этом среднем выборочном или в доле выборки? Насколько будет отличаться значение, если провести эксперимент повторно?

Математическое ожидание по одной выборке

Насколько полученное среднее выборочное отличается от ожидаемого значения?

Разница средних значений по двум выборкам

Насколько сильно отличаются средние значения по двум выборкам? (Говоря более техническим языком, какова вероятность, что мы бы наблюдали эту разницу средних значений или выше, будь верна нулевая гипотеза про отсутствие разницы между средними значениями по генеральной совокупности по двум выборкам?)

Вычисление размера выборки и анализ статистической мощности

Каким должен быть минимальный размер выборки, учитывая, что мне уже известно о процессе, чтобы достигнуть определенного

уровня уверенности в качестве данных? Эти типы статистических инструментов важны для планирования A/B-тестирования (подробнее об этом в главе 8).

Распределение данных

Соответствует ли распределение значений в этой выборке нормальному (конусообразному) распределению? Вероятно ли, что у этих двух выборок будет одинаковое исходное распределение?

Регрессия

Предположим, я провел тщательно разработанный эксперимент, в котором системно изменял одну (независимую) переменную, контролируя при этом максимально возможное число других факторов, после чего я построил прямую регрессии. Насколько я могу быть уверен в этой прямой? Насколько высока вероятность ее изменения (угол наклона и точка пересечения) при многократном повторении эксперимента?

Критерий соответствия и ассоциированности

В случае с категориальной переменной (например, категория продукта), соответствует ли частота или число (например, покупок) ожидаемой относительной частоте? Наблюдается ли взаимосвязь между двумя переменными, одна из которых категориальная?

Несмотря на краткость приведенного обзора, надеюсь, вы смогли разглядеть потенциальную ценность того набора инструментов, с помощью которого делаются статистические выводы. Он позволяет разрабатывать эксперименты и получать более объективный анализ данных, снижая количество ложноположительных результатов, происходящих из-за чистой случайности.

ПРОГНОСТИЧЕСКИЙ АНАЛИЗ

Делать прогнозы чрезвычайно сложно,
особенно относительно будущего.
приписывается Нильсу Бору

Прогностический анализ строится на индуктивном анализе. Цель в том, чтобы изучить взаимосвязи между переменными на основе существующего набора данных и разработать статистическую модель,

способную прогнозировать значения для новых, неполных или будущих точек данных.

На первый взгляд это кажется магией вуду, не меньше. В конце концов, мы не имеем ни малейшего представления, когда следующее мощное землетрясение разрушит Сан-Франциско (сроки имеющегося предсказания уже прошли), где и когда в следующем сезоне образуются ураганы или сколько будут стоить акции Apple в понедельник утром (если бы я мог сделать такой прогноз, то не писал бы сейчас эту книгу). Реальность такова, что мы не в состоянии точно предсказать какие-то неожиданные события и катастрофы, так называемых черных лебедей[1]. При этом во многих аспектах бизнеса и других областях знаний есть достаточные сигналы, с обработкой которых прогностический анализ отлично справляется. Например, в 2008 году Нейту Сильверу удалось предсказать результаты выборов в Сенат и победителей в 49 штатах из 50.

В сфере розничной торговли могут наблюдаться устойчивые закономерности. На рис. 5.12 приводится четкая и предсказуемая кривая (синяя сверху) ежегодных продаж солнечных очков, которая достигает пика в июне-июле и находится на спаде в ноябре и январе (предположительно небольшой ее рост наблюдается в декабре во время сезонной распродажи). Похожая кривая, но со смещением на шесть месяцев, отражает ежегодные продажи перчаток: ее пик приходится на декабрь. Таким образом, на основе результатов прогностического анализа можно разработать планы, когда производить или покупать товары, какой объем товаров производить или покупать, когда организовать доставку в магазины и так далее.

Помимо временны́х рядов прогностический анализ также способен делать прогнозы, к какому *классу* может относиться объект анализа. Например, на основе информации о размере заработной платы, истории покупок, оплаченных кредитной картой, истории оплаты (или неоплаты) счетов того или иного человека можно вычислить степень кредитного риска. Или на основе записей в Twitter, содержащих краткую оценку фильма, каждый из которых был отмечен пользователем положительно («фильм понравился») или отрицательно («отвратительный фильм»), можно разработать модель, прогнозирующую эмоциональную окраску — положительную или отрицательную — новых записей,

[1] Taleb N. N. The Black Swan. The Impact of the Improbable (New York: Penguin Press, 2007). Издана на русском языке: Талеб Н. Черный лебедь. Под знаком непредсказуемости. М. : Азбука-Аттикус : КоЛибри, 2016. *Прим. ред.*

например, таких как «спецэффекты в фильме просто классные», кото-

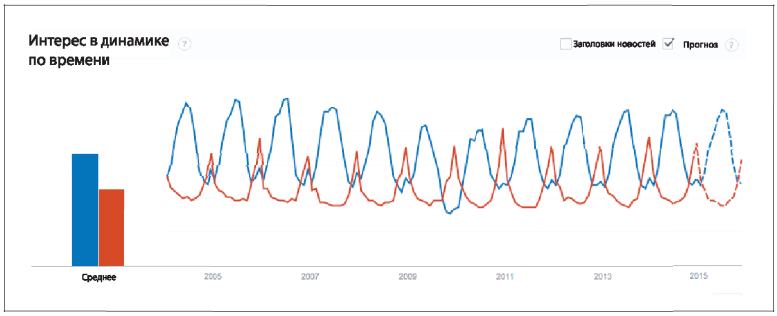

Рис. 5.12. Инструмент Google Trends отражает предсказуемую сезонную закономерность интереса к солнечным очкам (верхняя синяя кривая) и перчаткам (нижняя красная кривая) в период 2004–2014 годов и прогноз на год, до 2015-го

Существует множество приложений, использующих прогностическую аналитику, и они весьма заметны на рынке. Ниже приведено несколько примеров.

Прогнозы, формирующие основу сервиса как такового

Приложения для знакомств

Качественные приложения для поиска новых знакомых могут повысить степень удовлетворенности потребителей.

Приложения для игры на бирже (на риск пользователя!)

Отслеживая движение цен на акции и определяя закономерности, с помощью специальных алгоритмов можно попытаться покупать на спаде, продавать на пике и максимизировать рентабельность вложенных средств.

Прогнозы, обеспечивающие более высокий уровень обслуживания для клиентов

Спам-фильтры

Обнаружение и фильтрация спама («Купите “Виагру” онлайн») от не спама («Запланированная встреча с генеральным директором») делает работу с электронной почтой более эффективной, а пользователя — более счастливым.

Рекомендации по контенту

Качественные рекомендации, что можно посмотреть (Netflix), гарантируют возврат пользователей и снижают количество пользователей, отказавшихся от услуг.

Общение в социальных сетях

Сервис LinkedIn «Люди, которых вы можете знать» повышает эффективность пользования социальной сетью и обеспечивает более высокую ценность для пользователей и более ценные данные для социальной сети.

Прогнозы, способные обеспечить более высокий уровень конверсии и размер корзины

Кросс-продажи и увеличение объема покупки

Даже самые простые рекомендации, основанные на ассоциациях, например «Пользователи, которые купили DVD “Холодное сердце”, также покупают “Русалочку”» (Amazon), увеличивают объем продаж, а некоторым пользователям значительно облегчают и ускоряют процесс совершения покупок.

Рекламные объявления и купоны

Изучение истории покупок пользователя, а также прогнозирование его потенциальных интересов или намерений, может способствовать более релевантному отображению рекламных объявлений или более эффективному предложению купонов (например, от компании Tesco, далее мы поговорим об этом подробнее).

Прогнозы, способствующие улучшению стратегии

Одобрение от банка

Прогноз, у кого из заемщиков потенциально могут возникнуть трудности с выплатой взятых на себя обязательств, можно включить в процесс одобрения кредитных заявок, что снизит риск невозврата кредита.

Прогнозирование в работе органов правопорядка

Можно делать прогнозы относительно того, где могут вспыхнуть беспорядки, и принимать решения, куда и когда отправить полицейские наряды.

Прогнозирование активности пользователей

Благодаря прогнозированию наплыва или активности пользователей, например, что во время «Суперкубка» может произойти резкое увеличение количества сообщений в Twitter, можно заранее расширить технические мощности, чтобы предотвратить сбой в работе сервиса.

Политические кампании

Качественное прогнозирование намерений избирателей (голосовать / не голосовать, за демократов / за республиканцев / не определился) и ежедневное обновление данных привело к повышению эффективности в работе со СМИ, во взаимодействии с избирателями и в сборе пожертвований на проведение избирательной кампании, что в значительной мере обеспечило успех президентской кампании Барака Обамы.

Это всего лишь несколько примеров. Для получения более подробного обзора по теме прогностического анализа я рекомендую книгу Джона Сигела Predictive Analytics (John Wiley & Sons), в частности табл. 1–9.

Итак, как проводится прогностический анализ? Для этого существует целый ряд инструментов и подходов. Самая простая из возможных моделей — прогнозировать, что завтра будет таким же, как сегодня. Этот подход может сработать в случае медленно изменяющихся явлений, например, когда речь идет о погоде в Южной Калифорнии, но не в случае с волатильными системами, например такими, как цена на акции. Регрессия — самая обширная семья статистических инструментов. Для работы с разными характеристиками данных применяют разные виды регрессии (лассо-регрессию, гребневую, робастную и так далее). Особенный интерес представляет логистическая регрессия, которую можно применять для прогнозирования классов. Например, если раньше для определения категории спам / не спам использовалась модель наивного байесовского классификатора, то сегодня чаще применяется логистическая регрессия. К другим техникам и так называемому машинному обучению относятся нейронные сети, деревья решений и регрессии, алгоритм машинного обучения «Случайный лес», метод опорных векторов, метод *k* ближайших соседей.

Прогностический анализ весьма эффективен, но не обязательно сложен. Наиболее сложное в нем — получить качественный набор данных. При разработке классификатора часто это означает ручной контроль над данными, например маркировку набора сообщений в Twitter как положительных или отрицательных, что может быть особенно трудоемко.

Однако при наличии этих данных с хорошей библиотекой, такой как scikit-learn[1], для составления базовой модели потребуется буквально несколько строк кода. При этом для получения *хорошей* модели часто требуется приложить больше усилий, провести больше итераций, а также процесс генерирования признаков (feature engineering). Признаки — вводные данные для модели. Они могут включать основные собранные данные, например количество заказов, простые производные переменные, такие как «Заказ был сделан в выходные? Да/нет», а также более сложные абстрактные признаки, такие как «коэффициент похожести» двух фильмов. Генерация признаков — это и искусство, и наука, и она зависит от степени владения профессиональными знаниями.

Наконец, для проведения прогностического анализа не требуется большого объема данных. Объем базы данных, на основе которой Нейт Сильвер составлял прогнозы по итогам предвыборной кампании 2008 года, был всего 188 тыс. единиц (см. презентацию Оливера Гризела, в которой подтверждаются эти цифры и приводится хороший краткий обзор прогностического анализа[2]). Основную роль сыграло то, что Сильвер располагал множеством самых разных источников и данных опросов, каждый из которых в чем-то был ошибочным и необъективным, тем не менее в совокупности они относительно точно отразили действительность. Подтверждено на практике, по крайней мере для определенных классов проблем, что большой объем данных позволяет обходиться простыми моделями[3] (см. приложение А).

Резюмируя сказанное, прогностический анализ — мощный инструмент в арсенале компании с управлением на основе данных.

КАУЗАЛЬНЫЙ (ПРИЧИННО-СЛЕДСТВЕННЫЙ) АНАЛИЗ

Вероятно, каждый из нас знает утверждение: «Корреляция не подразумевает причинно-следственных отношений»[4]. Если вы проведете сбор данных, а затем разведочный анализ, чтобы выявить интересные взаимосвязи между переменными, то, скорее всего, что-нибудь обнаружите. Однако даже если между двумя переменными наблюдается очень сущес-

[1] URL: http://scikit-learn.org/stable/.

[2] URL: https://speakerdeck.com/ogrisel/predictive-analytics.

[3] Fortuny E. J. de, Martens D. and Provost F. Predictive Modeling with Big Data: Is Bigger Really Better? Big Data 1, no. 4 (2013): 215–226. URL: http://online.liebertpub.com/doi/full/10.1089/big.2013.0037.

[4] Если не верите, проверьте ложные корреляции (например, объем потребления сыра в США коррелирует с количеством людей, умерших от того, что запутались в собственном постельном белье). URL: http://www.tylervigen.com/spurious-correlations.

твенная корреляция, это не означает, что одна из них обусловливает другую. (Например, уровень холестерина-ЛПВП обратно пропорционален вероятности развития сердечно-сосудистых заболеваний: чем выше уровень этого «хорошего» холестерина, тем лучше. При этом препараты, повышающие уровень холестерина-ЛПВП, никак не влияют на предотвращение сердечно-сосудистых заболеваний. Почему? Потому что холестерин-ЛПВП представляет собой побочный продукт нормальной сердечной деятельности, а не ее причину.) Таким образом, у подобного апостериорного анализа есть серьезные ограничения. Если вы действительно хотите понять систему и точно узнать, какими рычагами влияния на фокусные переменные и показателями вы обладаете, тогда вам требуется разработать причинно-следственную модель.

Основная идея похожа на ту, что была в описанном ранее примере с поощрением лояльности клиентов: провести один или серию экспериментов с изменением одного параметра и контролем максимального количества всех остальных. Например, можно провести эксперимент с электронной рассылкой клиентам, в которой вы протестируете тему сообщения. При прочих равных условиях (то же самое содержание, время отправки и так далее) с единственной разницей в теме, если вы отметите, что уровень просмотра сообщения с другой темой гораздо выше, у вас есть все основания сделать вывод, что именно тема сообщения — причина интереса к нему.

У этого эксперимента есть свои ограничения, так как, несмотря на то что он подтверждает влияние фактора темы сообщения, неясно, какое именно слово или фраза вызвали отклик пользователей. Чтобы это выяснить, требуется проведение дополнительных экспериментов. Рассмотрим более количественный пример: время отправки сообщения может оказать серьезное влияние на уровень просмотра. Чтобы это проверить, можно провести контролируемый эксперимент с вариантами (сделать отправку электронной рассылки по частям в 8, 9, 10 часов утра и так далее) и проанализировать, как время отправки сообщения повлияло на уровень просмотра. Так вы сможете прогнозировать (интерполировать) предполагаемый уровень просмотра сообщения, отправленного в 8:30 утра.

ЧТО ВЫ МОЖЕТЕ СДЕЛАТЬ?

Рекомендация аналитикам. Вам стоит стремиться действовать в двух направлениях — «точить топор» и расширять арсенал инструментов. Вы станете более эффективным и ценным специалистом, кроме того, это будет инвестицией в себя и в развитие вашей карьеры. Оцените статистические навыки и навыки визуализации данных,

которыми вы сейчас пользуетесь. Как вы можете их улучшить? Например, если вы освоите среду R, поможет ли это вам быстрее и эффективнее проводить разведочный анализ? Окажет ли более глубокий аналитический подход более важное влияние на ваш проект? Что вам необходимо, чтобы овладеть новым навыком?

Рекомендация руководителям. Обращайте особое внимание на ситуации, в которых применение дополнительных видов аналитической работы способно обеспечить более глубокие выводы и повлиять на эффективность деятельности компании. Если отсутствие товара на складе становится проблемным местом цепочки поставок, можно ли исправить эту ситуацию с помощью прогнозных моделей? Можно ли проводить больше экспериментов, которые углубят институциональные знания причинных факторов? Стимулируйте специалистов по работе с данными, чтобы они повышали квалификацию, и всячески их в этом поддерживайте. Позвольте им опробовать новые программные средства, которые могут облегчить их работу и сделать ее более эффективной.

Подобные эксперименты обеспечивают более глубокое понимание системы и причинно-следственных взаимосвязей, что можно использовать при составлении прогнозов и планировании кампаний и других изменений, цель которых — улучшить и без того хорошие показатели, которых кто-то только стремится достичь. На их основе также можно строить имитационные модели, которые можно применять для оптимизации системы. Например, можно смоделировать цепочку поставок и изучить, как разные варианты схемы и условий пополнения склада влияют на дефицит товаров на складе или на совокупные расходы на транспортировку и хранение товаров. Этот вид деятельности отражен в правом верхнем углу матрицы Дэвенпорта в табл. 1.2. Это наивысший уровень аналитики. Принимая во внимание контролируемый, научный характер сбора данных на протяжении определенного периода, а также высокую эффективность подобных каузальных моделей, они становятся, по словам Джеффри Лика, «золотым стандартом» анализа данных.

С точки зрения ведения бизнеса вся эта бурная деятельность по анализу данных и разработке моделей проводится не ради самой деятельности и не по прихоти высшего руководства. Ее цель — поддержка основных показателей, таких как уровни просмотров, конверсии, наконец, показатель выручки. Поэтому критически важно, чтобы эти основные показатели были правильными и были качественно разработаны. В противном случае вы будете оптимизировать не то, что надо. Учитывая важность качественной разработки показателей, подробнее остановимся на этом вопросе в следующей главе.