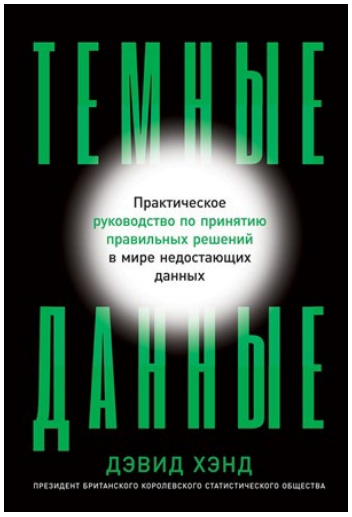
Дэвид Хэнд. Темные данные. Практическое руководство по принятию правильных решений в мире недостающих данных

Человечество научилось собирать, обрабатывать и использовать в науке, бизнесе и повседневной жизни огромные массивы данных. Но что делать с данными, которых у нас нет? Допустимо ли игнорировать то, чего мы не замечаем? Британский статистик Дэвид Хэнд считает, что это по меньшей мере недальновидно, а порой крайне опасно. В своей книге он выделяет 15 влияющих на наши решения и действия видов данных, которые остаются в тени. Например, речь идет об учете сигналов бедствия, которые могли бы подать жители бедных районов, если бы у них были смартфоны, о результатах медицинского исследования, которые намеренно утаили или случайно исказили, или о данных, ставших темными из-за плохого набора критериев для включения в выборку. Хэнд также рассказывает о том, какие меры могут сгладить эффект темных данных и как их можно обратить себе на пользу.

Дэвид Хэнд. Темные данные. Практическое руководство по принятию правильных решений в мире недостающих данных. – М.: Альпина Паблишер, 2023. – 366 с.



Купить книгу в издательстве Альпина Паблишер, цифровую книгу в ЛитРес, бумажную книгу в Ozon

Часть I. Темные данные. Происхождение и последствия

Глава 1. Темные данные. Незримая сила, которая формирует наш мир

Каждый год корь убивает почти 100 000 человек. К счастью, для Соединенных Штатов это редкое заболевание — например, в 1999 г. было зарегистрировано всего 99 случаев. Именно поэтому, когда родителям рекомендуют делать детям прививку от кори — заболевания, которого они и в глаза не видели, которым не болели ни их друзья, ни соседи и которое Центр по контролю и профилактике заболеваний признал не эндемичным для Соединенных Штатов, — они принимают такой совет с изрядной долей скепсиса.

Однако риск заражения все-таки существует, причем такой же реальный, как и раньше. Просто информация и данные, которые нужны родителям для принятия решений, отсутствуют, и риски становятся неочевидными. Для многочисленных видов отсутствующих данных я использую обобщающий термин «темные данные». Темные данные скрыты от нас, и этот факт означает, что мы рискуем недооценить опасность, сделать неправильный вывод и принять неверное решение. Иначе говоря, наше неведение становится причиной ошибок.

Темные данные ведут себя аналогично темной материи: мы не видим их, они не обнаруживаются, но все же способны оказывать существенное влияние на наши выводы, решения и действия. И, как я покажу на дальнейших примерах, если не осознать саму вероятность существования чего-то неизвестного, то последствия такой слепоты могут быть катастрофическими и даже фатальными.

Цель этой книги — исследовать, как и почему возникают темные данные. Мы рассмотрим различные виды темных данных, проследим, что приводит к их появлению, и выясним, как не допустить этого. Мы разберемся с тем, какие меры имеет смысл предпринимать, когда становится ясно, что темные данные все же имеются. А еще мы посмотрим, как этими данными, несмотря на их отсутствие, можно воспользоваться.

Темные данные принимают различные формы, возникают по разным причинам, и эта книга среди прочего содержит классификацию типов темных данных, обозначаемых как DD-тип x.

Например, в медицине понятие «травма» означает повреждение с возможными долговременными последствиями. Доктор Евгений Миркес и его коллеги из Лестерского университета в Великобритании провели исследование базы данных TARN[1] и выяснили: из 165 559 зарегистрированных травм исход 19 289 случаев оказался неизвестным. Этот пример иллюстрирует распространенную форму темных данных — DD-тип 1: данные, о которых мы знаем, что они отсутствуют. Иначе говоря, нам известно, что травмы у этих пациентов чем-то закончились, — мы просто не знаем, чем именно.

Основная идея этой книги: хотя иметь много данных полезно, большие данные, то есть объем, — это еще далеко не все. И то, чего вы не знаете, те данные, которых у вас нет, могут быть важнее для понимания происходящего, чем те, которыми вы располагаете.

Исследователи обычно имеют некий идеальный список людей, от которых они хотели бы получить ответы, но, как правило, отвечают не все. Если все те, кто отвечает, каким-то образом отличаются от тех, кто этого не делает, то у исследователей появляется основание усомниться в достоверности статистической сводки для данной группы населения. В конце концов, если бы некий журнал затеял опрос своих подписчиков, задав им единственный вопрос: «Отвечаете ли вы на журнальные опросы?», тот факт, что 100% ответивших скажут «да», еще не говорил бы о том, что все подписчики отвечают на подобные опросы.

DD-тип 2: Данные, о которых мы не знаем, что они отсутствуют. Вот иллюстрация темных данных такого рода. В конце октября 2012 г. сильнейший ураган, получивший название «Сэнди», обрушился на восточное побережье Соединенных Штатов. В период с 27 октября по 1 ноября 2012 г. было опубликовано более 20 млн. твитов об урагане. Очевидно, что это идеальный материал, на основе которого можно получить непрерывную картину стихийного бедствия по мере его развития — вы видите, какие районы пострадали больше всего и куда направить экстренную помощь.

Однако спустя какое-то время анализ показал, что наибольшее количество твитов о «Сэнди» пришло с Манхэттена и лишь немногие поступали из таких районов, как Рокуэй и Кони-Айленд. Причина заключалась не в том, что ураган пощадил их, а в том, что на их территории оказалось меньше пользователей твиттера и меньшее число смартфонов, чтобы отправить твит. Представим крайний вариант. Если бы ураган «Сэнди» полностью уничтожил какой-нибудь населенный пункт, то оттуда вообще бы не поступало никаких твитов.

Как-то на информационном брифинге бывший министр обороны США Дональд Рамсфелд охарактеризовал темные данные второго типа, да так удачно, что его высказывание стало знаменитым: «Есть известные неизвестные; то есть мы знаем, что есть какие-то вещи, которых мы не знаем. Но есть также неизвестные неизвестные — те, о которых мы не знаем, что мы их не знаем».[2] Этот замысловатый пассаж стал объектом насмешек для разнообразных СМИ, но их критика была несправедливой. То, что сказал Рамсфелд, было сущей правдой и имело глубокий смысл.

Так вы думаете, у вас есть все данные?

Через кассу супермаркета в базу попадают данные по всем транзакциям, совершенным всеми покупателями, и по всем товарам. Такие данные еще называют исчерпывающими. Но мы бы хотели знать, кто, что, когда и сколько купит в будущем. Другими словами, нам нужны данные, которые не собираются. Это связано с самой природой времени, и здесь фигурируют темные данные DD-тип 7: Данные, меняющиеся со временем.

Помимо этого, интересно узнать, как вели бы себя люди, если бы мы, скажем, более плотно заставили товарами полки, или разместили их как-то иначе, или изменили часы работы супермаркета. Такие

---

[1] ChatGPT поясняет: TARN (Trauma Audit and Research Network) – британская медицинская база данных, специализирующаяся на сборе и анализе данных о тяжелых травмах. Она является одной из крупнейших в мире систем мониторинга травматических повреждений и их исходов.

[2] Оригинальная цитата Дональда Рамсфелда на брифинге Министерства обороны США 12 февраля 2002 года (D. Rumsfeld, Department of Defense News Briefing, February 12, 2002) звучит так: Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say, we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know.

данные называются контрфактуальными, поскольку они противоречат реальным фактам — они о том, что случилось бы, если бы произошло нечто, чего на самом деле не происходило. Контрфактуальные данные классифицируются как DD-тип 6: данные, которые могли бы существовать.

Поскольку число возможных причин возникновения темных данных, по сути, не ограничено, знание того, на что следует обращать внимание, является чрезвычайно важным для предотвращения ошибок и просчетов. Именно с этой целью в нашей книге и представлено описание DD-типов:

- DD-тип 1: данные, о которых мы знаем, что они отсутствуют;
- DD-тип 2: данные, о которых мы не знаем, что они отсутствуют;
- DD-тип 3: выборочные факты;
- DD-тип 4: самоотбор;
- DD-тип 5: неизвестный определяющий фактор;
- DD-тип 6: данные, которые могли бы существовать;
- DD-тип 7: данные, меняющиеся со временем;
- DD-тип 8: неверно определяемые данные;
- DD-тип 9: обобщение данных;
- DD-тип 10: ошибки измерения и неопределенность;
- DD-тип 11: искажения обратной связи и уловки;
- DD-тип 12: информационная асимметрия;
- DD-тип 13: намеренно затемненные данные;
- DD-тип 14: фальшивые и синтетические данные;
- DD-тип 15: экстраполяция за пределы ваших данных.

Глава 2. Обнаружение темных данных. Что мы собираем, а что нет

В этой главе мы рассмотрим три основных метода создания наборов данных, а также пути возникновения темных данных, связанные с каждым из них. Вот три основные стратегии создания наборов данных.

- Сбор данных обо всех интересующих нас объектах.
- Сбор данных о некоторых элементах совокупности.
- Изменение условий. Первые две стратегии помогают собрать данные наблюдения. Если же вы меняете условия сбора данных, иначе говоря, вмешиваетесь, то такие данные называются экспериментальными. Экспериментальные данные особенно важны, потому что они могут дать информацию о контрфактуальности.

Извлечение, отбор и самоотбор данных

Данные все чаще сохраняют, отправляют в базы данных и там аккумулируют. То же самое происходит и с побочными или, как их еще называют, выхлопными данными (по аналогии с выхлопными газами), которые в дальнейшем помогают добиться лучшего понимания, усовершенствовать системы или восстановить картину событий, если что-то пошло не так.

Выхлопные данные, описывающие людей, называются административными. Особая сила административных данных заключается в том, что они сообщают не то, что люди говорят о своих действиях (как, например, в случае опросов), а то, что они делают на самом деле. Такие данные показывают, что люди купили, где они это купили, что они ели, какие поисковые запросы делали и т.д. Считается, что административные данные намного точнее демонстрируют реалии общества, чем ответы людей на вопросы об их действиях и поведении.

Мое первое настоящее знакомство с темными данными состоялось в сфере банковских услуг для потребительского сектора: кредитные и дебетовые карты, персональные займы, автокредиты, ипотека и прочие подобные вещи. Мне сделали заказ на создание «системы показателей» — статистической модели для прогнозирования вероятности неплатежей, которая могла бы использоваться при принятии решений о предоставлении кредитов. Мне был открыт доступ к большому набору данных, содержащему информацию из заявок предыдущих клиентов, а также их кредитные истории, показывающие действительную картину того, платили они или нет по своим обязательствам.

Проблема заключалась в том, что банк хотел получить модель, позволяющую делать прогнозы в отношении всех будущих заявителей. Предоставленные мне данные, безусловно, не были генеральной совокупностью, отражавшей всех заявителей — они касались лишь тех, кто уже прошел

процесс отбора. Надо полагать, состоявшиеся клиенты получили кредиты, потому что им был присвоен статус приемлемого риска в соответствии с каким-то более ранним механизмом отбора.

DD-тип 2: данные, о которых мы не знаем, что они отсутствуют, особенно обманчив, потому что у нас, как правило, нет оснований подозревать существование таких данных. Допустим, вы читаете лондонскую The Times от 29 декабря 2017 г. и, так же как и я, узнаете, что, «по данным полиции, число сексуальных домогательств, предположительно совершенных водителями такси по отношению к пассажирам, возросло на одну пятую за три года». Объяснение, лежащее на поверхности, состоит в том, что совершается все больше подобных правонарушений. Но есть и другое объяснение, вытекающее из темных данных: число совершенных преступлений не меняется, зато растет число сообщений о них. Если мы видим внезапное изменение шага во временном ряду значений, это может быть связано не только с тем, что поменялись параметры наблюдаемой реальности, но и с тем, что изменилась сама процедура сбора данных. Это также проявление темных данных DD-тип 7: Данные, меняющиеся со временем.

От нескольких ко многим

Исследуя выборки мы входим в область темных данных DD-тип 4: самоотбор. Один из ярких примеров такого рода проблем дают президентские выборы 1936 г. в США. На основе опросов популярный журнал The Literary Digest предсказал, что победит кандидат от республиканцев Альфред Лэндон. Тем не менее Франклин Рузвельт, кандидат от демократов, одержал уверенную победу. Результаты этих выборов и ошибочный прогноз The Literary Digest часто связывают с темными данными, полученными в ходе опроса. Не стоило в качестве рамки выборки использовать телефонные справочники. В то время телефоны были чем-то вроде предметов роскоши и в основном принадлежали состоятельным людям, среди которых было больше сторонников республиканцев. Таким образом, в выборке была завышена доля людей, которые планировали голосовать против Рузвельта.

Я считаю, что популярная «телефонная» теория неверна. Дело в том, что, хотя было разослано 10 млн анкет, лишь около четверти тех, кто их получил, то есть около 2,3 млн человек, потрудились ответить. Более трех четвертей опрошенных просто проигнорировали анкеты — их политические взгляды стали темными данными. Если республиканские избиратели были более заинтересованы в выборах, чем сторонники Рузвельта, то они с большей вероятностью откликались на опрос. Это и создало впечатление большинства в пользу Лэндона. Искажение самоотбора было устранено, когда состоялись настоящие выборы.

Расхождение прогноза и результатов выборов было следствием темных данных, но не из-за ошибок при составлении выборки, а по причине разной вероятности того, что республиканские и демократические избиратели вообще откликнутся на опрос.

Экспериментальные данные

Перейдем к третьему способу сбора данных – рандомизированным контролируемым исследованиям (РКИ). Все было бы замечательно, если бы пациенты придерживались протоколов лечения, принимая лекарства строго по графику вплоть до самого конца исследования. Но, к сожалению, темные данные часто встречаются и в такого рода исследованиях, проникая в них в форме выбывших. Выбывшие — это те, кто выбыл из исследования по тем или иным причинам.

Например, в исследовании, в котором реальное лечение сравнивается с плацебо, побочные эффекты более вероятны при реальном лечении. Это может означать, что выбывшие будут чаще появляться в группе лечения. Это еще один пример ошибки выжившего — те, кто «выживают» или продолжают лечение до конца исследования, не являются репрезентативной выборкой для всей группы.

Общим термином направления статистики, которая занимается оптимальным распределением испытуемых в процессе лечения, чтобы наилучшим образом проанализировать эффективность этого лечения, является *экспериментальный дизайн*. Рандомизированное исследование с двумя группами — самый простой из возможных и широко используемых дизайнов изучения эффективности лечения, политики или иных видов вмешательства.

Принципы экспериментального дизайна были разработаны ведущим британским статистиком сэром Рональдом Фишером, который ставил сельскохозяйственные эксперименты.

Остерегайтесь человеческих слабостей

В этой главе мы рассматриваем три основных способа сбора данных. Полученные данные рассеивают мрак вокруг и освещают нам новые миры. Но стратегии их сбора разрабатываем и воплощаем мы сами. А наши решения о том, какие данные собирать основываются на предыдущем опыте, который может не отражать того, что ждет нас в будущем. Более того, наши решения обусловлены нашим эволюционным развитием. Мы уязвимы для всевозможных подсознательных когнитивных искажений. *Эвристика доступности* — одно из таких искажений. В ее основе лежит склонность судить о вероятности события, исходя из того, насколько легко мы можем вспомнить соответствующий пример.

Другое когнитивное искажение — ошибка базового процента.

Предвзятость подтверждения — еще один риск, которому подвержено наше восприятие. Мы склонны выискивать информацию, которая поддерживает нашу точку зрения, и игнорировать данные, противоречащие ей. Психологические эксперименты отчетливо демонстрируют, что то, во что мы верим, может влиять на то, что мы помним.

Глава 3. Определения и темные данные. Что именно вы хотите узнать?

Одна из фундаментальных причин возникновения темных данных — использование неподходящих определений. Измерение инфляции основано на изучении цен на определенный набор товаров и услуг, называемых потребительской корзиной (которой в природе, конечно, не существует), и отслеживании изменений средней цены на эту корзину. Но, как рассчитать среднее значение? Недавно Великобритания перешла от индекса инфляции, основанном на среднем арифметическом, к индексу на основе среднего геометрического, что привело его в соответствие с индексами большинства других стран. Когда меняется определение, вы начинаете смотреть на вещи с другой точки зрения, поэтому меняются и аспекты данных, которые вы видите.

Бывают и другие причины, по которым темные данные возникают в индексах инфляции: прежде чем производить расчет, необходимо решить, какие товары и услуги включать в потребительскую корзину и как именно получать информацию о ценах.

Причинность

Когда демографические исследования показали связь между раком легких и курением, выдающийся статистик Рональд Фишер отметил, что это необязательно означает, что курение вызывает рак. Он привел несколько возможных причин возникновения такой связи явлений, в том числе вероятность того, что и рак легких, и предрасположенность к курению могут быть вызваны каким-то третьим общим фактором, например на генетическом уровне. Здесь мы имеем дело с классическим примером темных данных DD-тип 5: неизвестный определяющий фактор — некоторой неизмеренной переменной, которая служит причиной и того и другого и даже приводит к корреляции между ними, при этом сами по себе изучаемые явления непосредственно не влияют друг на друга.

Парадокс!

Иногда последствия влияния темных данных DD-тип 5: неизвестный определяющий фактор могут ставить в тупик.

Изучение показателей выживаемости среди пассажиров и членов экипажа Титаника выявило нечто любопытное. На судне было 908 членов экипажа, из которых выжило только 212 человек, то есть 23,3%, а из 627 пассажиров третьего класса — тех, чьи каюты находились на нижних палубах корабля и кому было труднее выбраться, — выжил только 151 человек, то есть 24,1%. Хотя показатели выживаемости в этих двух группах не сильно отличаются, тем не менее мы видим, что вероятность выживания пассажиров была несколько выше. Но изучите показатели выживаемости мужчин и женщин по отдельности:

а — в целом; б — мужчины и женщины отдельно

		Члены экипажа,%	Пассажиры третьего класса,%
а		212/908 = **23,3**	151/627 = **24,1**
б	Мужчины	192/885 = **21,7**	75/462 = **16,2**
	Женщины	20/23 = **87,0**	76/165 = **46,1**

Рис. 1. Доля членов экипажа и пассажиров третьего класса «Титаника», переживших крушение

Скрининг

Подбор персонала — одна из сфер применения скрининга. Заявки подает множество кандидатов, которые отсеиваются после первичного изучения биографических данных и заполненных анкет. Кандидаты из короткого списка приглашаются на собеседование. Первичное изучение выполняет ту же роль, что и инструмент скрининга. Кандидатов, попавших на собеседование, но не прошедших его, можно рассматривать как ложноположительные результаты — они казались подходящими на основании резюме, а более глубокий анализ показал, что это не так. Но нельзя забывать и о тех кандидатах среди отсеянных до собеседования, которые подошли бы идеально. В медицине такие результаты называются ложноотрицательными, и, конечно, все это тоже темные данные.

Выбор на основе прошлого

Вполне логично прогнозировать будущие результаты на основе прошлых. Но, к сожалению, прошлое может быть очень ненадежным путеводителем по будущему. Организации приходят в упадок, производители автомобилей обновляют модельный ряд, а рестораны меняют своих владельцев.

Странный феномен, заставляющий нас ожидать того, что хорошие показатели должны ухудшиться, а плохие улучшиться, даже если процесс остался неизменным, называется возвратом к среднему. Это проявление темных данных DD-тип 3: выборочные факты.

Термин «возврат к среднему» был введен блестящим эрудитом викторианской эпохи Фрэнсисом Гальтоном. Он заметил, что (в среднем) дети высоких людей хотя и оказывались выше среднего роста, но при этом не были настолько же высокими, как их родители, а дети, рожденные невысокими людьми, хотя и были ниже среднего роста, но все-таки превосходили своих родителей в росте (подробнее см. Фрэнсис Гальтон. Наследственность таланта, её законы и последствия).

Глава 4. Непреднамеренные темные данные. Видим одно, регистрируем другое

Как правило, пытаться понять смысл большой таблицы данных, просто разглядывая ее, занятие малопродуктивное. Чтобы облегчить задачу, нужно сначала *обобщить* значения. Другими словами, мы анализируем данные, сжатые до формата сводок, который нам удобнее воспринимать. Например, мы вычисляем средние значения и диапазоны значений, а также более сложные статистические обобщения, такие как коэффициенты корреляции, коэффициенты регрессии и факторные нагрузки. Однако, по определению, любое обобщение означает жертвование деталями или, что то же самое, затемнение данных (DD-тип 9: обобщение данных).

Также важно тщательно выбирать статистические сводки, чтобы они соответствовали вопросу, который мы хотим задать. Средний доход, рассчитанный как среднее арифметическое, в небольшой компании из десяти сотрудников, девять из которых зарабатывают в год по $10 000, а один — $10 млн составляет более $1 млн. Такая информация будет вводить в заблуждение, например, того, кто претендует на работу в этой компании. По этой причине доходы и благосостояние часто рассчитываются не как средняя, а как медианная величина, когда половина зарплат ниже, а половина выше этого значения.

Глава 5. Стратегические темные данные. Уловки, обратная связь и информационная асимметрия

В математике есть очень глубокая и мудрая теорема, названная в честь ее первооткрывателя Курта Геделя, которая, если упростить формулировку, гласит следующее: любая достаточно сложная система аксиом содержит утверждения, которые нельзя ни доказать, ни опровергнуть в рамках этой системы (см. Эрнест Нагель, Джеймс Рой Ньюмен. Теорема Гёделя). В жизни это означает, что даже тщательно разработанные системы неизбежно содержат прорехи. Одной из сфер, где мы регулярно наблюдаем такие прорехи, является налоговое право. Легальные схемы минимизации налогов возникают как следствие неопределенностей или упущений в налоговом законодательстве.

Агентская проблема — еще одна иллюстрация темных данных, тесно связанных с уловками. Она возникает, когда одно лицо (агент) может принимать решения от имени другого лица (принципала). Например, сотрудники принимают решения от имени своего работодателя, а политики действуют от лица своих избирателей. Сотрудники начинают использовать знания и информированность для своей выгоды в ущерб работодателю; политики отворачиваются от избравшего их народа и начинают действовать в собственных интересах, тем самым вставая на скользкий путь, ведущий к диктатуре (см. Законы Хаммурапи и проблема принципал – агент).

Закон Кэмпбелла доступно и кратко излагает, почему уловки так опасны в контексте государственной политики. Он гласит: «Чем шире какой-либо количественный показатель используется для принятия социальных решений, тем больше он подвержен злоупотреблениям и тем больше искажаются социальные процессы, которые контролируются с его помощью». Закон Гудхарта говорит нечто подобное, хотя и в более мягкой форме: «Когда показатель становится целью экономической политики, он перестает быть хорошим измерителем».

Обратная связь

Хорошие результаты тестирования вдохновляют на приложение дополнительных усилий, которые могут привести к увеличению масштабов успеха и возникновению желания еще больше нарастить усилия. Это пример механизма обратной связи, в котором измеряемые данные возвращаются назад, чтобы влиять и изменять свои значения. Эти механизмы встречаются и в психологии: знание того, что за вами наблюдают, может побудить вас старательнее выполнять задание (эффект Хоторна). Особенно ярко механизмы обратной связи проявляют себя при образовании финансовых пузырей.

Пузырем на финансовых рынках называют необоснованное значительное повышение цен на акции (или другие виды активов), за которым следует резкое падение. Хотя фундаментальная стоимость компании и является одним из факторов, влияющих на рыночную цену ее акций, определяет эту цену нечто совсем другое — готовы или не готовы ее платить участники рынка. На этот счет есть прекрасная аналогия с конкурсом красоты, приведенная выдающимся экономистом Джоном Мейнардом Кейнсом: «...Это не тот случай, когда выбирают самых хорошеньких, полагаясь на свой вкус, и даже не тот, когда полагаются на мнение большинства о красоте. Здесь мы встаем на третью ступень, которая заставляет нас предвидеть мнение большинства в отношении того, каким будет мнение большинства. А некоторые, как мне думается, применяют на практике четвертую, пятую и более высокие ступени» (см. Профессиональное инвестирование, как конкурс красоты).

Информационная асимметрия

Информационная асимметрия — это термин для ситуаций, в которых одна сторона располагает большей информацией, чем другая. Иными словами, для одной из сторон какие-то данные являются темными, и это ставит ее в невыгодное положение в переговорах или конфликтах.

В статье 1970 г. «Рынок “лимонов”: неопределенность качества и рыночный механизм» лауреат Нобелевской премии экономист Джордж Акерлоф иносказательно описал, какие тяжелые последствия может иметь информационная асимметрия. На сленге продавцов подержанных машин «лимонами» называют автомобили низкого качества или с дефектами. В противоположность «лимонам» качественные автомобили именуют «персиками» (см. Джордж Акерлоф. Рынок «лимонов»: неопределенность качества и рыночный механизм).

Покупатели подержанных машин не могут быть уверены в исправности выбранного автомобиля. При прочих равных условиях их покупка может с одинаковой вероятностью оказаться и «лимоном», и «персиком». Поэтому покупатели готовы платить только некую среднюю цену. Но у продавцов есть преимущество — они точно знают, где «лимоны», а где «персики», и, конечно, не хотят продавать последние за такую среднюю цену. Поэтому «персики» они придерживают, толкая покупателям одни «лимоны». Покупатели быстро обнаруживают подвох, и, соответственно, еще ниже опускают цены, по которым готовы покупать, что, в свою очередь, становится для продавцов еще большим аргументом против продажи «персиков». Возникает обратная связь, которая заставляет уйти с рынка владельцев «персиков» и в результате снижает как цены, так и качество продаваемых автомобилей.

Постоянно ищите информационную асимметрию и почаще задавайтесь вопросом: что он, она или они могут знать такого, чего не знаете вы?

Глава 6. Умышленно затемненные данные. Мошенничество и обман

Все виды мошенничества объединяет одно — сокрытие информации. Мошеннические действия варьируют от обычной тщательной проверки регистрационных записей до сложных статистических методов, от моделирования типичного поведения клиентов с помощью машинного обучения и интеллектуального анализа данных до специальных программных фильтров, которые обнаруживают определенные виды транзакций со сложной структурой. Что касается темных данных, мораль очевидна: если что-то выглядит слишком хорошо, чтобы быть правдой, вероятно, с этим «что-то» не все в порядке. Скорее всего, оно призвано скрыть какую-то правду.

Глава 7. Наука и темные данные. Природа познания

Наука становится таковой, если к ней применим критерий Поппера, или фальсифицируемость (см. Карл Поппер. Предположения и опровержения. Рост научного знания). Основная идея заключается в том, что вы выдвигаете некое потенциальное объяснение изучаемого явления (теорию, догадку или гипотезу), а затем проверяете его, наблюдая, насколько прогнозируемые вашим объяснением последствия соответствуют тому, что происходит на самом деле. Если перевести на язык этой книги, то для подтверждения или опровержения нашей теории, мы должны сопоставить данные, которые она прогнозирует, с полученными экспериментальными данными. Если прогнозные данные не соответствуют действительности, то теория заменяется на другую, модифицируется или расширяется, пока не станет не только успешно подтверждать прошлое, но и предсказывать будущее.

До победы научной революции успехи познания сдерживались нежеланием собирать данные, которые могли бы опровергнуть теорию, — такова предвзятость подтверждения. Развитию науки препятствовало нежелание сделать темные данные видимыми. Если у вас имеется солидная теория, которой уже несколько веков, зачем искать данные, которые противоречат ей? Вспомнить хотя бы миазматическую теорию, господствовавшую в Европе, Индии и Китае с древнейших времен вплоть до XIX в., которая гласила, что эпидемии вызваны ядовитыми парами гниющей материи.

Наука дает объяснения, каждое из которых становится все более точным по мере углубления познания, но всегда остается возможность опровержения любого из этих объяснений новыми экспериментальными данными. Такая вероятностная природа теорий, допускающая их изменение по мере поступления новых данных, — это то, что отличает науку, например, от религии, которая никак не связана с доказательствами.

Проще говоря, наука — это процесс. В частности, она не является набором известных фактов, хотя для простоты, особенно при обучении, научный процесс зачастую подают именно так. Возможно, мы совершаем досадное упущение: научное образование в наших школах должно стать естественной колыбелью критического мышления, давая детям наряду с констатацией фактов инструмент, который позволит им в будущем лучше оценивать любую информацию.

Натыкаясь на темные данные

Обычно темные данные представляют собой проблему, требующую серьезного поиска: от нас скрыто нечто, что могло бы изменить наше понимание и, скорее всего, повлиять на наши действия. Но иногда мы совершенно случайно натыкаемся на темные данные, и перед нами внезапно открывается целый мир.

В «Структуре научных революций» философ Томас Кун пишет: «Именно это и происходит с новыми фундаментальными фактами и теориями. Они создаются непреднамеренно в ходе игры по одному набору правил, но их восприятие требует разработки другого набора правил. После того как они становятся элементами научного знания, наука... никогда не остается той же самой» (см. Томас Кун. Структура научных революций).

Стремление сделать эффектное открытие привело к практике препарирования данных бесконечным множеством методов и реконфигурирования наборов данных до тех пор, пока не будет найдено что-то существенное. Например, сравнивая две группы пациентов, мы можем измерить 100 характеристик каждого пациента, а затем сравнить средние значения двух групп по каждой из них. Было бы удивительно, если бы при этом не обнаружилось хотя бы несколько существенных отличий между группами — исключительно из-за случайных ошибок измерения. Такую манипуляцию иногда называют p-хакингом. Этот термин пришел из статистики и описывает явление, с которым стоит разобраться (см. Дэвид Шпигельхалтер. Искусство статистики. Как находить ответы в данных).

Смысл p-значения часто понимается неверно. Принято думать о нем как о показателе вероятности того, что гипотеза верна. Это не так. Гипотеза либо верна, либо ошибочна, а p-значение просто показывает вероятность получения определенных экстремальных результатов в первом случае, то есть когда гипотеза верна.

Термин p-хакинг появился благодаря пагубной практике проводить бесконечное множество проверок значимости без учета их количества. Почему это становится проблемой, понять несложно. Предположим, что мы проверяем 100 никак не связанных между собой гипотез, каждая из которых верна, но нам это неизвестно. Далее предположим, что мы рассматриваем p-значение на уровне 2%

для любой из этих 100 гипотез как достаточно низкое, чтобы отнестись к ней с сомнением. Для каждой взятой в отдельности проверки значимости это вполне разумно, поскольку означает, что вероятность ложных подозрений в отношении этой единственной гипотезы, если она верна, составляет всего 2%. Но в случае, если вы проводите для каждой из 100 гипотез 100 проверок с уровнем p-значения 2%, получается, что вероятность возникновения сомнений по крайней мере для одной из них составит 87%. Скорее всего, вы решите, что хотя бы одна из гипотез является ошибочной, даже если все они будут верны. Вспомните о достаточно долгих пытках данных! Если вы скрываете тот факт, что провели 100 проверок, по сути, превращая их в темные данные, то ваши выводы могут быть очень обманчивыми.

На эту тему есть анекдот. Экспериментатор А говорит экспериментатору Б, что у него большие проблемы с воспроизведением результатов, полученных Б. «Неудивительно, — отвечает тот, — ведь я тоже не смог получить их первые 100 раз, когда проводил эксперимент».

Часть II. Освещение и использование темных данных

Глава 8. Принцип работы с темными. данными

Связываем наблюдаемые и недостающие данные. Если набор данных оказался неполным, то ключевым фактором в борьбе с темными данными становится понимание того, почему эти данные отсутствуют. Я предлагаю классифицировать недостающие данные на три категории:

- Неигнорируемо потерянные наблюдения как зависимые от невидимых данных, или UDD (Unseen Data Dependent). Вероятность отсутствия наблюдений зависит от значений, которые неизвестны.
- Случайно потерянные наблюдения как зависимые от видимых данных, или SDD (Seen Data Dependent). Здесь вероятность невозможности наблюдения зависит от данных, которые ранее наблюдались.
- Абсолютно случайно потерянные наблюдения как независящие от данных, или NDD (Not Data Dependent). Вероятность отсутствия наблюдения никак не зависит от данных, имеющихся или нет.

Важность темных данных в экономике иллюстрируется тем фактом, что в 2000 г. американский экономист Джеймс Хекман был удостоен Нобелевской премии «за разработку теории и методов анализа селективных выборок», которой он занимался в 1970-х гг. Понятие «селективные выборки» — это еще один способ показать, что у вас не хватает данных, а есть только отдельные выборки, сделанные из имеющихся значений.

Классификация по категориям NDD, SDD, UDD очень полезна, поскольку для разных механизмов возникновения недостающих данных требуются разные типы решений. Если вы, скажем, исследуете сферу, где люди особенно чувствительны к сообщенным ими данным, то можете предположить, что недостающие значения принадлежат категории UDD. Например, в исследовании, касающемся употребления кокаина, недостающие данные с большей вероятностью будут из категории UDD, чем в исследовании на тему использования общественного транспорта.

Я выделяю три основных шага в работе с темными данными: предотвращение, обнаружение и исправление.

Предотвращение

Ошибки в данных предотвращаются, во-первых, благодаря пониманию того, какие именно ошибки бывают, и, во-вторых, путем создания систем, которые препятствуют их возникновению на этапе сбора данных. Например, если речь идет о дате рождения, то для машины не составит труда проверить, является ли она допустимой. Я слышал о случае, когда набор данных имел странный пик по датам рождения, приходившийся на 11 ноября 1911 г. Как выяснилось, дату рождения требовалось вводить шестью цифрами в формате день/месяц/год и программисты были в курсе, что люди иногда вводят 00/00/00, если не хотят указывать свой день рождения. Поэтому они запрограммировали форму сбора данных таким образом, что, если кто-то вводил шесть нулей, машина отклоняла дату и требовала повторить попытку. Но люди вбивали последовательность из шести единиц, что принималось базой данных и выглядело как 11 ноября 1911 г.

Обнаружение

Пример статистического обнаружения странностей встречается в распределении Бенфорда. Первое описание этого распределения (иногда его называют законом Бенфорда), по-видимому, было сделано

в 1881 г. американским астрономом Саймоном Ньюкомом. В своей работе он использовал логарифмические таблицы — до появления компьютеров с их помощью перемножали большие числа. Ньюком обратил внимание на тот факт, что первые страницы логарифмических таблиц всегда были замусолены больше, чем последующие. Закон был повторно открыт почти 60 лет спустя физиком Фрэнком Бенфордом, который провел обширное исследование, показавшее, что частое использование более ранних значений по сравнению с более поздними характерно для очень разных числовых таблиц (см. Закон Бенфорда или закон первой цифры).

Закон Бенфорда гласит: во многих полученных наборах чисел цифры от 1 до 9 встречаются в качестве начальных в разных пропорциях: 1 встречается примерно в 30% случаев, 2 — в 18% и т.д. по убывающей, вплоть до 9, которая служит наиболее значимой цифрой всего для 5% чисел в наборе. Закон Бенфорда посредством точной математической формулы как раз и описывает это распределение. Если данные отклоняются от распределения Бенфорда, то это повод проверить, не закралась ли какая-то ошибка или мошенничество (Miller S.J. Benford's Law: Theory and Applications).

Глава 9. Полезные темные данные. Переосмысление вопроса

Американским статистиком Брэдом Эфроном для определения точности оценок был разработан метод бутстреппинга (см. Стандартное отклонение и стандартная ошибка). Идея заключалась в том, чтобы принять единственную имеющуюся у нас выборку за всю совокупность. Затем, мы могли бы извлечь подвыборку из нашей выборки (каждая подвыборка должна иметь тот же размер, что и исходная выборка, благодаря многократному включению в нее каждого значения). Фактически точно так же, как мы могли бы извлечь много выборок из генеральной совокупности, мы можем извлечь много подвыборок из одной имеющейся у нас выборки.

Мнимые данные: байесовское априорное распределение

При рассмотрении симулирования мы предполагали, что правильно понимаем базовую структуру, процесс и механизм возникновения данных. Такая уверенность часто неоправданна. Мы можем иметь некоторое представление о процессе, но нам должно очень повезти, чтобы действительно знать все в деталях.

Учитывая фундаментальную субъективность и неопределенность причин наших представлений, вполне понятна неуверенность, которая может у нас возникнуть в отношении того, стоит ли что-то утверждать или принимать решения на основании собственного мнения по этому вопросу. Вместо этого следует собрать данные, чтобы добиться большей объективности. И это именно то, для чего нужен байесовский подход к статистике. В соответствии с ним следует взять наши первоначальные мнения о возможных значениях среднего, называемые априорными убеждениями, а затем корректировать их по мере поступления новых данных, что приводит к появлению апостериорных убеждений.

Этот процесс обновления или корректировки выполняется с использованием фундаментальной теоремы Байеса (см. Идеи Байеса для менеджеров). С нашей точки зрения, теорема Байеса объединяет ненаблюдаемые темные данные с реально наблюдаемыми, чтобы получить новое распределение мнений о среднем. Байесовская статистика играет чрезвычайно важную роль — это одна из двух (или по другой версии трех) основных школ статистического анализа.[3]

---

[3] Вот что думает ChatGPT о трех упомянутых школах:
Частотная (классическая) статистика – основана на вероятностях как пределе относительных частот при бесконечном числе повторений эксперимента. Основные методы включают оценивание параметров, проверку гипотез, доверительные интервалы. Это традиционный подход, лежащий в основе многих статистических тестов (например, критерий Стьюдента, ANOVA).
Байесовская статистика – трактует вероятность как субъективную меру уверенности, которая обновляется по мере поступления новых данных. Основана на теореме Байеса и включает понятие априорного и апостериорного распределений. Этот подход широко применяется в машинном обучении, медицинской диагностике и принятии решений в условиях неопределенности.
Фидуциарная статистика (или обобщенная теория вероятностей) – менее популярная школа, связанная с работами Рональда Фишера. В ней вероятность рассматривается как свойство параметров, а не данных, что отличает ее от частотного и байесовского подходов. Иногда третьей школой называют *интуиционистскую* (или импрессионную) статистику, которая связана с эвристическими методами принятия решений и анализом данных без строгого формального обоснования.

Сбор данных в темноте

Данные, идентифицирующие людей, можно анонимизировать, но есть возможность пойти еще дальше. Можно делать данные темными по мере их сбора и использования в расчетах, чтобы их вообще никто никогда не видел, но они по-прежнему были бы доступными для анализа.

Прежде всего это рандомизированный ответ — хорошо известная стратегия сбора конфиденциальной личной информации, такой как данные, касающиеся сексуального или нечестного поведения. Для примера предположим, что мы хотим знать, какая часть населения хотя бы раз в жизни совершала кражу. Прямой вопрос на эту тему в лучшем случае приведет к искаженным ответам, поскольку очевидно, что люди склонны лгать и отрицать. Вместо этого мы просим каждого человека подбросить монету, которую видит только он.

Люди проинструктированы, что, если выпадает орел, они должны правдиво ответить «да» или «нет» на вопрос «Совершали ли вы когда-нибудь кражу?», а если выпадает решка, то они должны просто ответить «да». Теперь для любого человека положительный ответ означает, что мы не будем знать, украл ли он что-то на самом деле или это монета упала решкой вверх. Но мы узнаем нечто большее. Поскольку вероятность того, что выпадет орел, равна 1/2, мы будем знать, что общее число ответивших «нет» — только половина тех, кто действительно ничего не крал. Так что удвоение этого числа скажет нам о том, сколько человек действительно не совершали краж. Вычитая это значение из общего числа, мы узнаем число тех, кому доводилось красть.

Стратегия рандомизированного ответа — способ скрывать данные по мере их сбора. Есть также способы скрывать данные во время расчетов. Защищенное многостороннее вычисление — это способ сбора информации в группе, при котором никто из ее участников не имеет доступа к чужим данным. Вот простейший пример. Предположим, мы хотим узнать среднюю зарплату в группе проживающих рядом людей, но все они очень чувствительны к раскрытию информации о своем заработке. В этом случае я прошу каждого из них разбить его зарплату на два числа, a и b, так, чтобы их сумма равнялась зарплате. Таким образом, тот, кто зарабатывает £20 000, может разделить их на £19 000 и £1000, или на £10 351 и £9649, или на £2 и £19 998, или даже на £30 000 и -£10 000.

Совершенно не важно, как именно люди разделят свою зарплату. Они могут использовать и положительные, и отрицательные числа, главное, чтобы выполнялось условие — эти числа должны складываться в зарплату. Затем все части a отправляются кому-то, кто складывает их и получает общее значение A. Все части b отправляются кому-то другому (важно, чтобы это был другой человек), который также складывает их, чтобы получить значение B. Последний шаг — просто сложить A и B и разделить на число человек, чтобы получить среднее значение. Обратите внимание, что на протяжении этого процесса никто не знает значений чужих зарплат. Даже те люди, которые складывают одни части, понятия не имеют, что представляют собой другие части.

Глава 10. Классификация темных данных. Путь в лабиринте

Самое важное послание этой книги: относитесь к данным с подозрением — по крайней мере пока не будет доказано, что они адекватны и точны. Предложенная систематизация DD-типов дает своего рода контрольный список опасностей и общих проблем, на которые следует обращать внимание, работая с любым набором данных. И, конечно, всегда необходимо помнить, что обнаружение одного DD-типа не исключает присутствия других.

Классическим примером того, какие идеи можно извлечь из данных нового типа, служит проект «Миллиард цен». Альберто Кавалло и Роберто Ригобон из Школы менеджмента Слоуна извлекли огромное количество онлайн-цен из интернета и использовали их для построения индексов инфляции. На основе этого открытого источника данных они показали, что уровень цен и динамика инфляции в Бразилии, Чили, Колумбии и Венесуэле примерно совпадают. Кроме того, они обнаружили «в Аргентине большое необъяснимое расхождение между уровнями инфляции по онлайн-ценам и по официальной статистике»2. Простого объяснения этого несоответствия не было. Кавалло заключил: «Результаты для Аргентины подтверждают подозрение, что правительство манипулирует официальной статистикой по инфляции. Это единственная страна, где онлайн-инфляция с течением времени значительно отклоняется от официальных оценок».