РОМАН ЗЫКОВ

С DATA SCIENCE

КАК МОНЕТИЗИРОВАТЬ БОЛЬШИЕ ДАННЫЕ

Роман Зыков

Роман с Data Science. Как монетизировать большие данные

«Питер»
2021

УДК 004.62
ББК 32.973.233.02

Зыков Р.

Роман с Data Science. Как монетизировать большие данные / Р. Зыков — «Питер», 2021

ISBN 978-5-4461-1879-3

Как выжать все из своих данных? Как принимать решения на основе данных? Как организовать анализ данных (data science) внутри компании? Кого нанять аналитиком? Как довести проекты машинного обучения (machine learning) и искусственного интеллекта до топового уровня? На эти и многие другие вопросы Роман Зыков знает ответ, потому что занимается анализом данных почти двадцать лет. В послужном списке Романа — создание с нуля собственной компании с офисами в Европе и Южной Америке, ставшей лидером по применению искусственного интеллекта (AI) на российском рынке. Кроме того, автор книги создал с нуля аналитику в Ozon.ru. Эта книга предназначена для думающих читателей, которые хотят попробовать свои силы в области анализа данных и создавать сервисы на их основе. Она будет вам полезна, если вы менеджер, который хочет ставить задачи аналитике и управлять ею. Если вы инвестор, с ней вам будет легче понять потенциал стартапа. Те, кто «пилит» свой стартап, найдут здесь рекомендации, как выбрать подходящие технологии и набрать команду. А начинающим специалистам книга поможет расширить кругозор и начать применять практики, о которых они раньше не задумывались, и это выделит их среди профессионалов такой непростой и изменчивой области. Книга не содержит примеров программного кода, в ней почти нет математики. В формате PDF A4 сохранен издательский макет.

УДК 004.62
ББК 32.973.233.02

ISBN 978-5-4461-1879-3

Содержание

Роман Зыков
Роман с Data Science. Как монетизировать большие данные

Иллюстрации Владимира Вышванюка

которые вы можете найти в этой книге. На момент подготовки книги к изда-нию все ссылки на интернет-ресурсы были действующими.

Об авторе

Роман Владимирович Зыков, 1981 года рождения, в 2004 году получил степень бакалавра, а затем магистра прикладной физики и математики в МФТИ (Московском физико-техническом институте).

В 2002 году начал свой карьерный путь в аналитике данных (Data Science) в качестве технического консультанта в компании StatSoft Russia, российского офиса одноименной американской компании-разработчика пакета статистического анализа данных STATISTICA. В 2004 году был принят на должность руководителя аналитического отдела интернет-магазина Ozon.ru, где создавал аналитические системы с нуля, в том числе веб-аналитику, аналитику баз данных, управленческую отчетность, внес вклад в систему рекомендаций.

В 2009 году консультировал ряд проектов инвестиционного фонда Fast Lane Ventures и гейм-индустрии.

В 2010 году возглавил отдел аналитики в интернет-ритейлере Wikimart.ru.

В конце 2012 года стал сооснователем и совладельцем маркетинговой платформы для интернет-магазинов RetailRocket.ru. На текущий момент компания является безусловным лидером на рынке в России и успешно работает на рынках Чили, Голландии, Испании и других.

С 2007-го вел блог «Аналитика на практике» (KPIs.ru – ныне не существует), где евангелизировал анализ данных в применении к бизнес-задачам в электронной коммерции. Выступал на отраслевых конференциях, таких как РИФ, iMetrics, Gec 2014 вместе с Аркадием Воложем (Yandex), бизнес-конференциях в Дублине и Лондоне, в посольстве США (AMC Center), университете Сбербанка. Печатался в технологическом прогнозе PwC, ToWave, «Ведомостях», «Секрете фирмы».

В 2016 году прочитал мини-лекцию в концертном зале MIT в Бостоне о процессах тестирования гипотез.

В 2020 году был номинирован на премию CDO Award.

Благодарности

Я посвящаю эту книгу своей жене Екатерине и моим детям – Аделле и Альберту. Катя придала мне решимости написать книгу и приняла большое участие в редактировании текстов. За что я ей очень благодарен.

Также я благодарен своим родителям, которые вырастили и воспитали меня в очень непростое время. Отдельная благодарность моему отцу Владимиру Юрьевичу за то, что привил мне любовь к физике.

Я благодарен всем на моем долгом пути в аналитику данных. Илье Полежаеву, Большакову Павлу и Владимиру Боровикову за грамотное руководство, когда я только пришел в StatSoft. Бернару Люке, тогда генеральному директору Ozon.ru, а также коллегам в Ozon.ru: Александру Перчикову, Александру Алехину, Валерию Дьяченко – за совместное написание рекомендательной системы. Марине Туркиной и Ирине Коткиной – с вами было замечательно сотрудничать. Основателям проекта Wikimart.ru Камилю Курмакаеву и Максиму Фалдину – те знакомства в Калифорнии очень сильно повлияли на меня. Александру Аникину – ты очень крутой был тогда, а сейчас вообще звезда. Основателям проекта Ostrovok.ru – Кириллу Махаринскому и Сержу Фаге, а также Жене Курышеву, Роману Богатову, Феликсу Шпильману – с вами очень интересно было работать, я узнал много нового о разработке.

Я благодарен сооснователям Retail Rocket – Николаю Хлебинскому и Андрею Чижу. Отдельная благодарность венчурному фонду Impulse VC (Кириллу Белову, Григорию Фирсову, Евгению Пошибалову) – за то, что поверили в нас. Всем сотрудникам Retail Rocket, особенно моим ребятам Александру Анохину и Артему Носкову – вы лучшие.

Я благодарен психологу Елене Клюстер, с которой работаю уже несколько лет, за осознание своих собственных границ и своих истинных желаний. Благодарен Андрею Гузю, моему тренеру по плаванию, за аналитический подход к тренировкам. Оказывается, так можно, и не только профессионалам, но и любителям.

Выражаю благодарность всем моим виртуальным рецензентам, особенно Артему Аствацатурову, Александру Дмитриеву, Аркадию Итенбергу, Алексею Писарцову. Роману Нестеру – за рецензию на главу по этике данных.

Благодарен всем, кто способствовал изданию этой книги. Прежде всего Алексею Кузменко, который помог мне быстро найти издательство, минуя бюрократические препоны. Отдельная благодарность Владимиру Вышванюку за ироничные иллюстрации кота Вилли. Юлии Сергиенко и Наталье Римицан, которые делали все, чтобы эта книга вышла в свет.

От издательства

Ваши замечания, предложения, вопросы отправляйте по адресу comp@piter.com (издательство «Питер», компьютерная редакция).

Мы будем рады узнать ваше мнение!

На веб-сайте издательства www.piter.com вы найдете подробную информацию о наших книгах.

Введение

Дайте мне точку опоры, и я переверну Землю.
Архимед

Дайте мне данные, и я переверну всю вашу жизнь.
Data Scientist Архимед

Данные повсюду – начиная от алгоритмов «Тиндера», который «матчит» вас с далеко не случайными людьми, и заканчивая информационными войнами, которые ведут политики. Никого уже не удивляет, что за каждым нашим шагом пристально следят: будь то история запросов в браузере телефона или ваши действия в офлайне. Задержитесь на секунду у витрины спортивного магазина – и ждите его таргетированную рекламу в соцсетях с минуты на минуту. Расскажите коллеге, что натворил ваш кот, – и вот сухие корма и наполнители уже тут как тут в вашей ленте.

Особо впечатлительные могут впасть в паранойю – но данные в этом не виноваты. Все зависит от того, в чьи руки они попадут. С анализом данных связано очень много мифов, а data scientist – одна из самых перспективных и «сексуальных» профессий будущего. В своей книге я намерен развенчать мифы и рассказать, как все обстоит на самом деле. Надеюсь, читатель, ты, как и я, окажешься на «светлой» стороне силы.

Я окончил МФТИ в начале нулевых и тогда же возглавил аналитический отдел интернет-магазина Ozon.ru, где создавал аналитические системы с нуля. Я консультировал инвестиционные фонды, гигантов ритейла и гейм-индустрии, а восемь лет назад стал сооснователем и совладельцем маркетинговой платформы для интернет-магазинов RetailRocket.ru. Сейчас компания не просто является безусловным лидером на рынке в России, но и успешно работает на рынках Чили, Голландии, Испании и Германии. В 2016 году я прочитал лекцию в концертном зале MIT в Бостоне про процессы тестирования гипотез. В 2020 году номинировался на премию CDO Award.

Считается, что нужно потратить 10 000 часов для того, чтобы стать очень хорошим специалистом в своей области. Анализом данных я занимаюсь с 2002 года, когда это не было так популярно и хайпово. Так вот, чтобы получить эти заветные 10 000 часов, нужно проработать 10 000 часов / 4 часа в день / 200 дней в году = 12.5 лет. Я в полтора раза превысил эту цифру, поэтому, надеюсь, получилось написать книгу, которая будет очень полезна для вас, дорогие читатели.

Эта книга о том, как превращать данные в продукты и решения. Она основывается не на академических знаниях, а на моем личном опыте анализа данных длиной почти в двадцать лет. Сейчас существует очень много курсов по анализу данных (data science) и машинному обучению (machine learning, ML). Как правило, они узкоспециализированы. Отличие этой книги в том, что она, не утомляя частностями, дает цельную картину и рассказывает о том:

- как принимать решения на основе данных;
- как должна функционировать система;
- как тестировать ваш сервис;
- как соединить все в единое целое, чтобы на выходе получить «конвейер» для ваших данных.

Для кого эта книга

Эта книга предназначена для думающих читателей, которые хотят попробовать свои силы в области анализа данных и создавать сервисы на их основе.

Она будет вам полезна, если вы менеджер, который хочет ставить задачи аналитике и управлять ею. Если вы инвестор, с ней вам будет легче понять потенциал стартапа. Те, кто «пилит» свой стартап, найдут здесь рекомендации, как выбрать подходящие технологии и набрать команду. А начинающим специалистам она поможет расширить свой кругозор и начать применять практики, о которых вы раньше не задумывались – и это выделит вас среди профессионалов такой непростой и изменчивой области.

Как читать эту книгу

Я писал эту книгу так, чтобы ее можно было читать непоследовательно. Краткое содержание каждой главы:

Глава 1 «Как мы принимаем решения» описывает общие принципы принятия решения, как данные влияют на них.

Глава 2 «Делаем анализ данных» вводит общие понятия – с какими артефактами мы имеем дело, когда анализируем данные. Кроме того, с этой главы я начинаю поднимать организационные вопросы анализа данных.

Глава 3 «Строим аналитику с нуля» рассказывает об организации процесса построения аналитики: от первых задач и выбора технологии, заканчивая наймом.

Глава 4 «Делаем аналитические задачи» – полностью о задачах. Что такое хорошая аналитическая задача, как ее проверить. Технические атрибуты таких задач – датасеты, описательные статистики, графики, парный анализ, технический долг.

Глава 5 «Данные» о том, что говорят о данных – объемы, доступы, качество и форматы.

Глава 6 «Хранилища данных» рассказывает, зачем нужны хранилища, какие они бывают, также затрагиваются популярные системы для Big Data – Hadoop и Spark.

Глава 7 «Инструменты анализа данных», полностью посвящена наиболее популярным способам анализа от электронных таблиц в Excel до облачных систем.

Глава 8 «Алгоритмы машинного обучения» является базовым введением в машинное обучение.

Глава 9 «Машинное обучение на практике» является продолжением предыдущей главы: даются лайфхаки, как изучать машинное обучение, как работать с машинным обучением, чтобы оно приносило пользу.

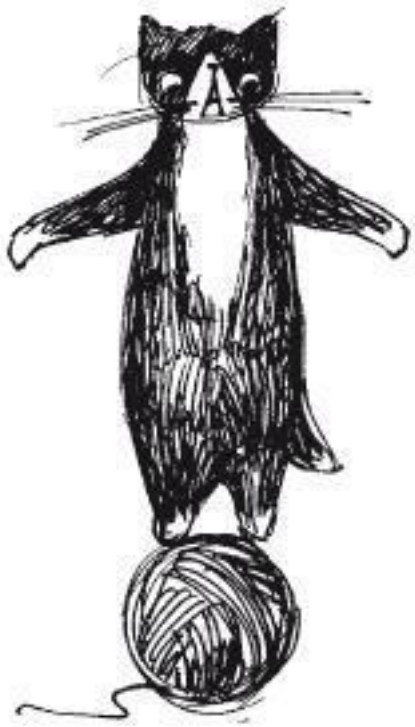
Глава 10 «Внедрение ML в жизнь: гипотезы и эксперименты» рассказывает о трех видах статистического анализа экспериментов (статистика Фишера, байесовская статистика и бутстрэп) и об использовании А/Б-тестов на практике.

Глава 11 «Этика данных». Я не смог пройти мимо этой темы, наша область начинает все больше и больше регулироваться со стороны государства. Здесь поговорим о причинах этих ограничений.

Глава 12 «Задачи и стартапы» рассказывает об основных задачах, которые я решал в e-commerce, а также о моем опыте сооснователя проекта Retail Rocket.

Глава 13 «Строим карьеру» больше предназначена для начинающих специалистов – как искать работу, развиваться и даже когда уходить дальше.

Глава 1
Как мы принимаем решения



«Итак, главный принцип – не дурачить самого себя. А себя как раз легче всего одурачить. Здесь надо быть очень внимательным. А если вы не дурачите сами себя, вам легко будет не дурачить других ученых. Тут нужна просто обычная честность.

Я хочу пожелать вам одной удачи – попасть в такое место, где вы сможете свободно исповедовать ту честность, о которой я говорил, и где ни необходимость упрочить свое положение в организации, ни соображения финансовой поддержки – ничто не заставит вас поступиться этой честностью. Да будет у вас эта свобода».

Нобелевский лауреат Ричард Фейнман, из выступления перед выпускниками Калтеха в 1974 году

Монетизация данных возможна лишь тогда, когда мы принимаем на основе этих данных правильные решения. Однако делать выбор, руководствуясь только статистикой, – плохая идея: как минимум нужно уметь читать их между строк и слушать свою интуицию (gut feeling). Поэтому в первой главе я расскажу про принципы, которыми я пользуюсь, принимая решения на основе данных. Я проверял на своем опыте – они работают.

Решения принимать непросто, ученые даже придумали новый термин «усталость от решений» (decision fatigue) [7]. Мы накапливаем стресс, совершая выбор каждый день сотни раз: и в какой-то момент, когда уже полностью вымотаны необходимостью принимать решения, можем махнуть рукой и начать действовать наугад. Я не зря привел в начале этой книги цитату

выдающегося физика, нобелевского лауреата Ричарда Фейнмана. Она напрямую касается как аналитики данных, так и вообще нашей жизни.

Как принимать верные решения, оставаясь честным с собой?

В книге «Биология добра и зла. Как наука объясняет наши поступки» профессор Стэнфордского университета, нейробиолог Роберт Сапольски [1] пишет, что на наши поступки, а значит и решения, влияет множество факторов: среда, в которой мы выросли, детские травмы, травмы головы, гормональный фон, чувства и эмоции. На нас всегда влияет множество факторов, которые мы даже не осознаем. Мы необъективны!

Лично я принял как данность, что гораздо легче принять необъективное и срезать углы, чем объективное, потому что для второго нужны серьезные усилия.

Вспомните об этом, когда будете предоставлять цифры кому-либо для принятия решения. И даже мои сотрудники указывали мне на то, что я сам нарушаю принципы объективности при утверждении результатов некоторых А/Б-тестов. Тогда я возвращался к реальности и соглашался с ними – объективность важнее моих априорных решений до проведения эксперимента.

В современном мире решения мы вынуждены принимать быстро и в условиях неопределенности. Но это не катастрофа. В квантовой физике, в отличие от классической, мы не знаем точно, где находится электрон, но знаем вероятность его нахождения. И вся квантовая физика базируется на этих вероятностях. С решениями точно так же – никто не знает истины, мы просто пытаемся угадать «правильное» с определенной долей успеха. И именно для этого нужны данные – увеличить вероятность успеха ваших решений!

Четыреста сравнительно честных способов

Остап Бендер знал четыреста сравнительно честных способов отъема денег у населения. Профессиональный аналитик знает примерно столько же способов «повернуть» цифры в сторону «нужного» решения. К сожалению, это очень распространено в политике: вспомните, как государства рапортовали о количестве зараженных во время пандемии вируса COVID-19. Показатели смертности в России были занижены [6]. Оказалось, что если человек болел коронавирусом и умер от сопутствующего заболевания, то в соответствующую статистику не попадет. В большинстве же западных стран одной положительной пробы на коронавирус было достаточно, чтобы попасть в статистику. Если копнуть глубже, мы видим, что у всех разная методика и разные цели. Существуют объективные и субъективные причины неточности таких цифр.

Первая причина – объективная: много бессимптомных носителей вируса, они не обращаются к врачам. Здесь требуется «ковровое» тестирование населения, которое подразумевает случайную выборку из всей популяции определенной местности. Тестирование добровольное, значит, кто-то не придет. Некоторые – потому, что у них есть симптомы коронавируса, и если это будет обнаружено в процессе тестирования, то их запрут дома на двухнедельный карантин. А это может привести к потере заработка. В итоге мы получим выборку, смещенную в сторону здоровых людей, а значит, и заниженную оценку количества заболевших.

Вторая причина тоже объективная – нет денег на массовое тестирование населения.

А вот третья причина – субъективная: власти хотят уменьшить официальную статистику заболевших, чтобы снизить панику среди населения и успокоить международное сообщество. Умение понимать эти причины и читать данные между строк – важное качество аналитика, которое позволяет ему делать более объективные выводы.

В работе я постоянно с этим сталкивался. Сейчас все живут на KPI, поэтому руководитель будет не очень-то рад плохим цифрам – премия висит на волоске. Возникает искушение найти показатели, которые улучшились. Нужно быть очень сильным руководителем, чтобы принять отрицательные результаты и внести коррекцию в работу. Аналитик данных как исследователь несет личную ответственность за результат своих цифр.

Чему можно научиться у Amazon?

Мне всегда нравились письма Джеффа Безоса (основателя Amazon.com) акционерам. Например, еще в 1999 году он писал про важность систем персональных рекомендаций на сайте, которые сейчас стали стандартом в современной электронной коммерции. Меня заинтересовали два его письма: 2015 [2] и 2016 [3] годов.

В первом из них Безос писал про «Фабрику изобретений» (Invention Machine). Он точно знает, о чем говорит, – само провидение вело Amazon через тернии электронной коммерции. Попутно в компании изобретали много вещей, абсолютно новых для рынка: система рекомендаций, А/Б-тесты (да-да, именно они были пионерами тестирования гипотез для веба), AWS (Amazon Web Services), роботизация склада, кнопки на холодильник для мгновенного заказа порошка и многое другое.

Так вот, в первом письме он рассуждает о том, как в больших компаниях принимаются решения об изобретении новых продуктов. Часто процесс утверждения выглядит так: все участники процесса (как правило, руководители департаментов компании) проставляют свои «визы». Если решение положительное, идея или гипотеза отправляются на реализацию. Здесь Безос предупреждает, что есть два типа решений и они не должны проходить один и тот же процесс утверждения.

Первый тип – решения, у которых нет или почти нет обратной дороги. Это как дверь, в которую можно войти, но нельзя выйти. Здесь нужно действовать очень внимательно и осторожно.

Второй тип – решения, у которых есть обратный ход. Дверь, в которую можно войти и выйти. Здесь он предлагает утверждать идею достаточно быстро, не мучая ее долго бюрократическими процедурами.

В письме 2016 года Безос противопоставляет компанию Дня 1 (Day 1), где сохраняется живая атмосфера создания компании и новых продуктов, компании Дня 2 (Day 2), которая статична и, как следствие, приходит к своей ненужности и смерти. Он выделяет 4 фактора, которые определяют компанию Дня 1:

- истинная одержимость покупателем (customer obsession);
- скепсис относительно моделей (a skeptical view of proxies);
- стремительное освоение внешних трендов (the eager adoption of external trends);
- стремительное принятие решений (the eager adoption of external trends).

Последний пункт мне кажется особенно важным в контексте этой книги. Для поддержания атмосферы компании Дня 1 требуется принимать быстрые и качественные решения. Мой шестилетний сын в таких случаях восклицает: «Но как?» Вот правила Безоса:

1. Никогда не использовать один-единственный процесс принятия решений (есть два типа решений, про которые я написал выше). Не дожидаться получения 90 % всей информации, нужной для принятия решения, – 70 % уже достаточно. Ошибаться не так страшно, если вы умеете быстро исправляться. А вот промедление, скорее всего, влетит вам в копеечку.

2. Не соглашайся, но позволяй. Когда руководителю предлагают идею талантливые и успешные сотрудники, а он не согласен с ней – ему стоит просто позволить им ее реализовать, а не тратить их усилия на то, чтобы убедить. Безос рассказал, как дали зеленый свет одному из сериалов Amazon Studios. Он считал, что запускать этот проект рискованно: Безосу эта история казалась сложной в производстве и не слишком интересной. Но команда с ним не соглашалась. Тогда он сказал – хорошо, давайте пробовать. Им не пришлось убеждать Безоса в своей правоте, и они сэкономили уйму времени. Сам он подумал так: эти ребята уже привезли домой одиннадцать премий «Эмми», шесть «Золотых Глобусов» и три «Оскара» – они знают, что делают, просто у нас разные мнения.

3. Быстро находите причины несогласия и эскалируйте их наверх вашим руководителям. Разные команды могут иметь разные взгляды на решение. Вместо того чтобы тратить время на изматывающих совещаниях в попытках договориться – лучше эскалировать проблему наверх.

Аналитический паралич

Поспешишь – людей насмешишь. Все самые страшные ошибки я совершил, когда торопился – например, когда 15 лет назад пришел в Ozon.ru, чтобы поднять аналитику с нуля и должен был каждую неделю делать огромную простыню метрик о деятельности всей компании без нормальных проверок. Из-за давления менеджмента и спешки в этом регулярном еженедельном отчете было множество ошибок, с последствиями которых мне еще долго пришлось разбираться.

Современный мир живет на бешеных скоростях, но расчет метрик нужно делать очень аккуратно, а значит, не быстро. Конечно, не стоит впадать в другую крайность – «аналитический паралич», когда на каждую цифру будет уходить очень много времени. Иногда попытки сделать правильный выбор приводят к тому, что я называю «аналитическим параличом» – когда уже пора принять решение, но не получается. Слишком высока неопределенность результата или рамки слишком жесткие. В аналитический паралич легко впасть, если пытаться принять решение чисто рационально, руководствуясь только логикой.

Яркий пример – книга «Проект Рози» Грэма Симсиона (кстати, одна из любимых книг Билла Гейтса и его жены). Молодой успешный ученый-генетик Дон ищет жену, но ни разу еще не продвинулся дальше первого свидания. Сочтя традиционный способ поиска второй половинки неэффективным, Дон решает применить научный подход. Его проект «Жена» начинается с подробнейшего 30-страничного вопросника, призванного отсеять всех неподходящих и выявить одну – идеальную. Понятно, что человека, который соответствовал бы такому списку требований, просто не существует. А потом он знакомится с девушкой, у которой нет ничего общего с его идеалом. Что из этого вышло – догадайтесь сами.

Второй пример – покупка машины. Когда я в последний раз делал это, то составил целую таблицу в Excel с техническими параметрами машин, вплоть до размера багажника в сантиметрах. Потом я целый год думал, ходил, смотрел, а в результате купил ту, которой и близко не было в моем списке, по велению сердца. Но на самом деле это было не веление сердца – просто за целый год поисков и анализа я понял, что в этом списке было по-настоящему важно для меня, а что нет.

Третий пример из моей профессиональной практики связан с гипотезами, точнее с тестами. Представьте себе, что вы вместо старого алгоритма рекомендаций разработали новый и хотите его протестировать. У вас есть 10 сайтов, где можно выполнить сравнение. В итоге вы получили: 4 выигрыша, 4 ничьи и 2 проигрыша. Стоит ли заменить старый алгоритм на новый? Все зависит от критериев решения, которые сформулировали перед тестом. Новый алгоритм должен победить на всех сайтах? Или вероятность выигрыша должна быть больше вероятности проигрыша? В первом случае очень высока вероятность того, что вы закопаетесь в бесконечных итерациях, «полируя» свой алгоритм до совершенства, особенно учитывая то, что тесты займут не одну неделю. Это типичная ситуация «аналитического паралича». Во втором – условие кажется легким. Хотя из практики скажу, что даже его выполнить бывает очень непросто.

Я считаю, что в решениях нужно идти на осознанный риск, даже если нет всей информации. В наше время, конечно, мир меняется слишком быстро, чтобы иметь роскошь долго делать выбор. Если решение не примете вы, это сделает кто-то за вас, например ваш конкурент.

Погрешности – правило штангенциркуля

Следующая вещь, с которой я столкнулся, – это точность цифр. Я много занимался анализом маркетинговой деятельности, в том числе маркетинговых акций. Моя задача заключалась в том, чтобы как можно более точно оценить их влияние на бизнес. Вообще реакция менеджеров на цифры разная – все радуются положительным результатам, не проверяя их; но когда видят отрицательные – сразу ищут ошибку. И скорее всего, «найдут». Видите ли, все метрики содержат ошибку. Вспомните лабораторные работы по физике в школе или институте, сколько мы мучились и считали погрешности. Системные, случайные… Сколько времени мы тогда тратили на то, чтобы подогнать результат под нужную закономерность?

В бизнесе и науке так делать нельзя, особенно если вы хотите быть хорошим аналитиком и не пользоваться вышеупомянутыми «сравнительно честными способами» повернуть цифры туда, куда нужно. Сейчас погрешность измерений веб-аналитики (системы измеряют посещаемость веб-сайтов) составляет около 5 %. Когда я еще работал в Ozon.ru, погрешность всей аналитической системы тоже была около 5 % (расхождение с данными бухгалтерии). У меня был серьезный случай – я обнаружил ошибку в коммерческой системе веб-аналитики Omniture Sitecatalyst (ныне Adobe Analytics): она не считала пользователей с браузером Opera. В результате погрешность измерений была очень большой – около 10 % всех совершенных заказов система, за которую мы платили более 100 тысяч долларов в год, безнадежно потеряла. С такой погрешностью ей тяжело было доверять – но, к счастью, когда я обнаружил ошибку системы и сообщил о ней в Omniture, их разработчики ее устранили.

При работе с погрешностями я вывел правило, которое называю Правилом штангенциркуля. Есть такой инструмент для измерения размеров деталей с точностью до десятых долей миллиметра. Но такая точность не нужна при измерении, например, размеров кирпича – это уже за пределами здравого смысла, достаточно линейки. Правило штангенциркуля я бы сформулировал так:

> Погрешность есть в любых измерениях, этот факт нужно принять, а саму погрешность – зафиксировать и не считать ее ошибкой (в одной из следующих глав я расскажу, как ее мониторить).

Задача аналитика – в разумной мере уменьшить погрешность цифр, объяснить ее и принять как данность. Как правило, в погоне за сверхточностью система усложняется, становится тяжелой с точки зрения вычислений, а значит, и более дорогой – ведь цена изменений становится выше.

Принцип Парето

Итальянский экономист и социолог Вильфредо Парето в 1897 году, исследуя структуру доходов итальянских домохозяйств, выяснил, что 80 % процентов всех их доходов приходится на 20 % из них.

Универсальный принцип, названный в его честь, был предложен в 1951 году, и сейчас принцип Парето звучит так: «20 % усилий дают 80 % результата».

Опираясь на свой опыт, я бы так сформулировал его на языке данных:

• 20 % данных дают 80 % информации (data science);

• 20 % фич или переменных дают 80 % точности модели (machine learning);

• 20 % из числа успешных гипотез дают 80 % совокупного положительного эффекта (тестирование гипотез).

Я почти 20 лет работаю с данными и каждый день убеждаюсь в том, что эта закономерность работает. Это правило лентяя? Только на первый взгляд. Ведь чтобы понять, какие именно 20 % позволят добиться результата, нужно потратить 100 % усилий. Стив Джобс в интервью Business Week в 98-м году сказал: «Простое сделать труднее, чем сложное: вам придется усердно поработать, чтобы внести ясность в ваши мысли, и тогда станет понятно, как сделать проще. Но это стоит того: как только вы достигнете этого, вы сможете свернуть горы».

Приведу пример того, как применяется правило Парето в машинном обучении. Для проекта обычно готовится ряд фич (входных параметров модели), на которых будет тренироваться модель. Фич может получиться очень много. Если выводить такую модель в бой, она будет тяжелой, требовать для своего поддержания много строк программного кода. Для такой ситуации есть лайфхак – посчитать вклад каждой фичи (feature importance) в результирующую модель и выбросить из модели фичи с минимальным вкладом. Это прямое использование правила Парето – 20 % фич дают 80 % результата модели. В большинстве случаев лучше модель упростить, пожертвовав небольшой долей ее точности, при этом проект будет в разы меньше исходного. На практике можно экономить время, подсмотрев фичи в решениях какой-нибудь схожей задачи на kaggle.com. Взять оттуда самые сильные из них и реализовать в первой версии собственного проекта.

Можно ли принимать решения только на основе данных?

Можно, но не всегда и везде. Области, где можно принимать решение только на основе данных, уже захвачены компьютерными алгоритмами. Они не устают и очень хорошо масштабируются. Тот же самый автопилот – уже относительно недалекое будущее: алгоритмы принимают решение на основе данных, поступающих к ним от датчиков, и управляют автомобилем.

Человек – универсальное существо, способное решать множество задач. Если задачу достаточно сузить, то можно сделать алгоритм, который будет работать быстрее тысячи человек. Но в отличие от человека, алгоритм не способен сделать ни шага в сторону от заданной схемы: его придется дорабатывать, внося каждое изменение. В этом и заключается вся суть автоматизации: сделать дешевле, быстрее и без участия человека. Поэтому все так одержимы идеей искусственного интеллекта.

На решения, принимаемые людьми, влияет много факторов. Один из них – так называемые когнитивные искажения, то есть систематические ошибки в восприятии и мышлении. Например, систематическая ошибка выжившего. Во время Второй мировой войны нью-йоркскому математику Абрахаму Вальду поручили исследовать пробоины на самолетах-бомбардировщиках, возвратившихся из боя, чтобы понять, в каких местах нужно усилить броню. Первое «логичное» решение – усилить броню в местах, поврежденных вражескими зенитками и пулеметами. Но Вальд понимал, что не может изучить все самолеты, включая те, что погибли. Проанализировав проблему как математик, он предложил бронировать те места, которые остались целыми, ведь самолеты с такими повреждениями не возвращались на базу, а значит, это самые уязвимые места.

Ошибку выжившего допустить очень легко. Чему нас учит пример Вальда? Тому, что нужно думать о всей генеральной совокупности. Ошибка выжившего является одной из форм когнитивных искажений.

В анализе данных ошибка выжившего – это учет известного и пренебрежение неизвестным, но существующим. С этой ошибкой очень легко столкнуться, когда у нас есть какие-то данные, на основе которых нужно сделать вывод. Любые данные – это выборка, ограниченное число. Сама выборка сделана из генеральной совокупности. Если выборка сделана случайно и она достаточно большая, то все хорошо – большая часть закономерностей будет зафиксирована в выборке, и выводы будут объективными. Если же выборка была не случайной, как в нашем случае с самолетами, где в ней отсутствовали сбитые машины, – то, скорее всего, выводы будут ошибочными.

Например, в среднем только 1 из 100 посетителей сайта интернет-магазина совершает покупку. Если мы захотим улучшить свой сайт, чтобы больше покупателей покупали, то с какими посетителями нужно работать? Обычно дизайнеры и продуктологи обращают внимание на существующих покупателей из-за того, что с ними можно пообщаться, есть контактная информация из заказов, по ним есть хорошая статистика. Но эта выборка составляет всего лишь 1 % от всей генеральной совокупности посетителей; с остальными почти невозможно связаться – это «сбитые самолеты». В итоге будет смещение выводов в сторону «выживших», а значит, выводы анализа не будут работать для всех посетителей.

Еще одно когнитивное искажение – предвзятость результата (outcome bias). Представьте себе – вам предлагают два варианта на выбор:

- Сыграть в «Орла или решку» – если выпадет орел, получите 10 000 рублей.
- Сыграть игральной костью с шестью гранями – если выпадет 6, получите 10 000 рублей.

Какой вариант выберете? Естественно первый, в котором шанс выиграть 1 к 2, во втором варианте значительно хуже – 1 к 6. Монету подбросили – выпала решка, вы ничего не получили.

Тут же бросили кость, выпала шестерка. Будет ли обидно? Да, будет. Но было ли наше решение правильным?

Этот пример я взял из поста «Фокусируйтесь на решениях, а не на результате» [5] Кэсси Козырьков (Cassie Kozyrkov), которая работает директором по принятию решений [4] (Decision Intelligence) в Google. Она советует всегда оценивать верность решения, учитывая, какой именно информацией вы обладали в момент его принятия. Многие люди жалеют, что они не уволились с работы раньше и только потеряли время, откладывая это решение, – я и сам в свое время так думал. И это отличный пример предвзятости результата – мы понимаем, что нужно было уволиться раньше, только обладая той информацией, которая у нас есть на данный момент. Например, что с тех пор зарплата так и не выросла, а интересный проект, который мы предвкушали, так и не был запущен. Оценивая последствия своего решения (особенно неудачного), в приступе самокопания мы не должны забывать, что принимали решение в условиях неопределенности.

Глава 2
Делаем анализ данных

Когда я работал в компании Wikimart.ru, основатели этого проекта познакомили меня с Ди Джеем Патилом (DJ Patil). Ди Джей был тогда одним из ангелов-инвесторов проекта, он руководил аналитикой в LinkedIn, затем был ведущим аналитиком данных (Chief data scientist) Белого дома в Вашингтоне при администрации Барака Обамы, тогдашнего президента США. Встречался я с Ди Джеем несколько раз в Москве и в Кремниевой долине в Калифорнии. В Москву он приезжал для презентации своей мини-книги «Building Data Science Teams» («Построение команд аналитиков данных») [9], выпущенной издательством O'Reilly. В книге он обобщил опыт ведущих компаний Кремниевой долины. Очень рекомендую вам эту книгу, так как ее мысли мне близки, и их я проверил на практике. Вот как автор определяет организацию, управляемую данными:

> «A data-driven organization acquires, processes, and leverages data in a timely fashion to create efficiencies, iterate on and develop new products, and navigate the competitive landscape».

> «Организация, управляемая данными, своевременно получает, обрабатывает и использует данные для повышения эффективности, итераций и разработки новых продуктов, а также навигации в конкурентной среде».

Далее Ди Джей указывает на принцип «Если ты не можешь измерить, ты не можешь это исправить» («if you can't measure it, you can't fix it»), который объединяет самые сильные организации, эффективно использующие свои данные. Вот рекомендации Патила, которые следуют из этого принципа:

• Собирайте все данные, какие только возможно. Вне зависимости от того, строите ли вы просто отчетную систему или продукт.

• Продумывайте заранее и делайте вовремя измерение метрик проектов.

• Позвольте как можно большему количеству сотрудников знакомиться с данными. Множество глаз поможет быстрее выявить очевидную проблему.

• Стимулируйте интерес сотрудников задавать вопросы относительно данных и искать на них ответы.

Эти мысли я еще озвучу в главе про данные. А теперь самое время поговорить о том, что мы получаем на выходе анализа данных.

Артефакты анализа данных

Здесь и далее под артефактами я буду понимать осязаемый результат, физический или виртуальный объект.

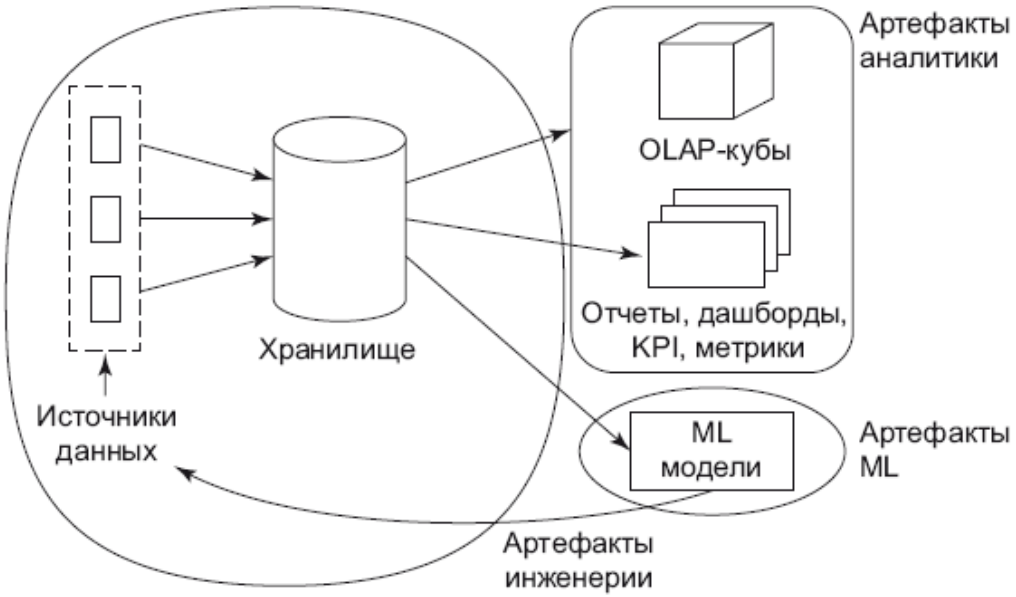


Рис. 2.1. Артефакты аналитики

Их можно разделить на три вида (рис. 2.1):
• артефакты бизнес-анализа данных (business intelligence);
• артефакты машинного обучения (machine learning);
• артефакты инженерии данных (data engineering).
Поговорим о них подробнее.

Бизнес-анализ данных

Бизнес-анализ данных (Business Intelligence, BI) – термин уже устоявшийся. Вот какое определение дает Википедия:

> «Business Intelligence – это обозначение компьютерных методов и инструментов для организаций, обеспечивающих перевод транзакционной деловой информации в человекочитаемую форму, пригодную для бизнес-анализа, а также средства для работы с такой обработанной информацией».

Под бизнес-анализом я подразумеваю объединение контекста бизнеса и данных, когда становится возможным бизнесу задавать вопросы к данным и искать ответы Первыми артефактами являются так называемые инсайты и гипотезы, вторыми – отчеты или дашборды, метрики и ключевые показатели (Key Performance Indicator). Поговорим подробнее об инсайтах и гипотезах.

Гипотезы и инсайты

Инсайт (insight) в переводе с английского – понимание причин. Именно за этим обращаются к аналитикам. В поиске инсайтов помогают аналитика и статистика:

• Цель аналитики заключается [10] в помощи формулирования гипотезы.

• Цель статистики [10] в том, чтобы эту гипотезу проверить и подтвердить.

Это требует пояснений. В бизнесе, да и в жизни тоже, мы ищем причину проблемы, задавая вопрос «почему?». Не зная причины, мы не можем принять решение. В игру вступает аналитика – мы формулируем список возможных причин: это и есть гипотезы. Чтобы это сделать, нужно задать несколько вопросов:

• Не происходило ли что-нибудь подобное раньше? Если да, то какие тому были причины? Тогда у нас будет самая первая и самая вероятная гипотеза.

• Обращаемся к бизнес-контексту: не происходило ли каких-либо неординарных событий? Часто как раз параллельные события влияют на возникновение проблемы. Еще плюс пара гипотез.

• Описательный анализ данных (exploratory data analysis): смотрим данные в аналитической системе (например, кубах OLAP), не видно ли каких-либо аномалий на глаз? Например, какие-либо распределения изменились во времени (типы клиентов, структура продаж и т. д.). Если что-то показалось подозрительным – дополняем список гипотез.

• Использование более сложных методов поиска аномалий или изменений, например, как описано здесь [11].

Наша цель – накидать как можно больше гипотез, не ограничивая фантазию, затем отсортировать их по списку в порядке убывания вероятности, чтобы найти верную гипотезу как можно быстрее. Или даже воспользоваться бритвой Оккама, выстроив гипотезы по возрастанию сложности проверки. Иначе можно столкнуться с аналитическим параличом: превратить задачу в научную работу, когда проверяются все гипотезы без исключения. Такого в реальной жизни не бывает, у нас всегда есть ограничения в ресурсах – как минимум во времени. Как только гипотезы готовы, приходит очередь статистики, с помощью методов которой они проверяются. Как это сделать – расскажу в главе про эксперименты в ML.

Когда я был директором по аналитике Retail Rocket (сервис рекомендаций для интернет-магазинов), мне и аналитикам часто приходилось заниматься расследованиями, ведь бизнес довольно большой – больше 1000 клиентов, и странности, с которыми приходится разбираться, случаются часто. Много приходится работать с так называемыми А/Б-тестами: это тесты, где аудитория сайта делится на две части случайным образом – первой части пользователей показывается одна версия сайта, второй – другая. Такие тесты обычно используют, чтобы оценить влияние изменений на бизнес-метрики сайта, когда первая версия – это старая версия или контрольная группа, а вторая – новая версия. Если это интернет-магазин – это, скорее всего, будут продажи. Далее к результатам теста применяются статистические критерии, которые подскажут достоверность изменений.

Такие тесты хорошо выявляют проблемы: например, версия сайта с обновленными рекомендациями Retail Rocket проиграла старой версии рекомендаций. Как только это становится известным, начинается расследование. Проверка начинается с интеграции, и это первая гипотеза: правильно ли передаются нам данные от интернет-магазина. Обычно на этом этапе решается 60–70 % проблем. Далее мы пытаемся найти отличие этого магазина от остальных в такой же тематике, например магазины одежды. Это вторая гипотеза. Третья гипотеза – возможно, мы изменили дизайн сайта таким образом, что полезная информация опустилась ниже на странице сайта. Четвертая гипотеза – тест мог отрицательно повлиять на определенные категории товаров. Собрав набор таких гипотез, мы начинаем их проверять примерно в такой последова-

тельности, как я описал. Довольно часто мы находим причину проблем, но иногда это не удается, его величество случай играет с нами в кошки-мышки, и эту мышку очень сложно найти.

Однажды клиент – магазин «Дочки-Сыночки» – тестировал наш сервис и сервис одного из наших российских конкурентов, чтобы выбрать лучший, и это превратилось в настоящий детектив [12]. Чтобы точно не проиграть в тесте, конкурент перемещал некоторое число пользователей, которые были близки к покупке, (например, добавили товар в корзину) из конкурентных (наших) сегментов в свой – причем делалось это не на постоянной основе, а в отдельные дни и часы. Основной метрикой сравнения была конверсия: процент пользователей, совершивших покупку. Ясно, что в той «мошеннической схеме» такой процент будет выше там, куда перетянули пользователей. Здесь компания Retail Rocket пошла на принцип! Мы стали копать. Через два месяца были обнаружены и опубликованы [12] факты подтасовки результатов. В итоге прошел ряд судебных процессов, и справедливость восторжествовала.

Отчеты, дашборды и метрики

Понятие самого отчета очень широкое, здесь я подразумеваю под ним табличное или иное графическое представление данных. Отчеты могут быть разными:

• Просто таблица с «сырыми» данными или так называемые «выгрузки», например, таблица с заказами клиентов.

• Отчет с «агрегированными» данными. Под агрегацией я подразумеваю суммы, количество и иные статистики. Например, таблица с именами клиентов и количеством заказов, который каждый из них совершил.

• Дашборды (dashboards) содержат ключевые показатели и метрики.

Первые два относительно просты и делаются через специальные системы, которые могут генерировать отчеты по запросу. Я стараюсь максимально оставить эту задачу на откуп пользователям. Почему? Потому, что тратить на это время высококвалифицированных сотрудников – значит стрелять из пушки по воробьям. Кстати, этим могут заняться стажеры-аналитики – отличный способ наработать опыт и понять бизнес-контекст. Как мотивировать пользователей стараться самостоятельно? Во-первых, они сэкономят время, которое обычно тратят на постановку задачи и ожидание результата. Во-вторых, получат возможность самим вносить правки и изменения – а значит творить. По моему опыту, обычно этим занимаются очень перспективные сотрудники, которые не бояться освоить новый инструмент, чтобы делать свою работу лучше. Остальным придется пройти через стандартный цикл планирования задач: а это время (дни, а иногда недели) и очень четкая формулировка технического задания. И кстати, все генеральные директора (Ozon.ru, Wikimart.ru, Ostrovok), с которыми я работал, пользовались OLAP-кубами со своих компьютеров. С их помощью они всегда могли ответить на простые вопросы, а если не получалось – обращались к аналитикам.

Теперь взглянем на дашборды и начнем с определения из Википедии:

> «Дашборд – это тип графического интерфейса, который делает возможным быструю оценку ключевых показателей для конкретной цели или бизнес-процесса. Часто подразумевается, что это просто другое название отчета о прогрессе или просто отчета».

Как правило, дашборд состоит из ключевых показателей и метрик, выраженных с помощью графических инструментов (графики, диаграммы и т. д.):

• Ключевой показатель (key performance indicator, KPI) – это индикатор, который показывает, насколько далеко мы находимся от цели, например отставание/опережение плана.

• Метрика – это цифра, которая характеризует процесс, обычно используется как справочная информация.

Главное различие между метрикой и ключевым показателем – наличие цели. В ключевых показателях она есть, в метриках обычно нет. Как правило, ключевые показатели используются в дашбордах компаний, где уже и продукт, и бизнес-процесс «устоялись». Уже накоплено некоторое множество данных, что делает возможным прогнозирование целей. Метрики я обычно вношу туда, где нет достаточно устойчивых процессов. Их обычно ставят, когда цель пока не прогнозируема. Зрелость дашборда можно определить по соотношению количества ключевых показателей и метрик: чем метрик больше, тем более незрелый дашборд мы получаем на выходе.

Обычно дашборды характеризуют какой-то бизнес-процесс (далее слово «процесс» без «бизнес»), например эффективность рекламы, складские остатки, продажи и т. д. Есть еще важные характеристики дашбордов:

• не является «простыней цифр»;

• показывает, где возникла проблема, но не дает ответа на вопрос почему.

Часто велико искушение сделать огромную простыню цифр, закрывающую все аспекты бизнеса. И я понимаю владельцев/менеджеров компаний – на старте проекта по построению внутренней аналитической системы всегда хочется большего. Я наблюдал это и в Ozon.ru, и в Ostrovok.ru. К слову, эти строки написаны по мотивам письма, которое я писал восемь лет назад операционному директору Ostrovok.ru, – он хотел получить от аналитиков ту самую «простыню». А я считаю такое цифровым «микроменеджментом», в нем легко запутаться, самые важные показатели похоронены среди второстепенных. С первого взгляда будет сложно понять, где возникла проблема, а это основная функция дашбордов. Бороться с этим можно, например, через внедрение OKR – цели и ключевые результаты (Objectives and Key Results) [13] – или системы сбалансированных показателей (Balanced Scorecard). В этой книге я не буду подробно останавливаться на этих методиках, но рекомендую вам с ними ознакомиться. Также можно чаще пользоваться графическими элементами, например, добавив на график линию тренда (с помощью семиточечного скользящего среднего, чтобы убрать недельную сезонность), будет легче заметить восходящий или нисходящий тренд.

Дашборд отвечает на вопрос, где есть проблема, а не почему она возникла. Может возникнуть искушение сделать огромный детальный отчет, чтобы быстро найти причину, – но тогда ваш дашборд превратится в простыню цифр, о которой я писал выше. В нем не будет интерактивности, и нужно будет «провалиться» внутрь этих цифр, чтобы проанализировать их, а для этого понадобятся совсем другие инструменты. Когда вам в следующий раз захочется это сделать, вспомните, удавалось ли вам хоть раз найти причину проблемы с помощью дашборда.

Никакой дашборд не заменит интерактивный анализ, для которого нужны соответствующая аналитическая система (SQL, OLAP, Google Data Studio, Tableau) и знание контекста. Мы никогда не сможем придумать ограниченный набор отчетов, которые будут отвечать на вопрос «почему». Максимум, что мы можем сделать, – наращивать (но не слишком) объем правильных метрик, исходя из инцидентов, за которыми будем следить.

Поэтому я всегда за лаконичные автоматические отчеты, которые будут отвечать на два вопроса: есть ли проблема и где она возникла. Если проблема есть, нужно лезть в интерактивные системы анализа данных.

Разработка дашбордов – это одна из самых нелюбимых работ у тех, кто занимается анализом данных. Когда я обсуждал этот вопрос с Ди Джеем Патилом, отметив, что 50 % времени аналитического отдела занимает работа над отчетностью, он сказал, что у них в LinkedIn тоже периодически накапливался пул таких задач и приходилось их закрывать. И взгрустнул. Но дашборды очень нужны – они помогают контролировать общее здоровье вашей системы – вверенных вам серверов и сетей, если вы системный администратор, или всей компании, если вы генеральный директор.

Артефакты машинного обучения

Раньше компьютером можно было управлять только с помощью прямых команд или инструкций: поверни сюда, дай назад, сложи и т. д. Это обычное, так называемое детерминированное программирование – для нас понятен алгоритм в виде инструкций, мы его описали, и компьютер подчиняется ему. Машинное обучение предполагает совершенно другой подход к программированию – обучение на примерах. Здесь мы показываем системе что-то с помощью примеров, тем самым избавляем себя от самостоятельного написания инструкций, что бывает совсем не просто. Это становится работой по обучению алгоритма ML.

Конец ознакомительного фрагмента.