Посвящается моим детям Элли, Уильяму и Эллен.
Элли было три года, когда она узнала, что ее папа — «доктор».
Озадаченно посмотрев на меня, она сказала: «Но ведь ты
не помогаешь людям…»
Памятуя об этом, я также посвящаю эту книгу вам, читатель.
Надеюсь, что она вам поможет.
— Алекс

Посвящается Стивену и Мелиссе.
— Джордан

Оглавление

Предисловие

Книга «Разберись в Data Science» вышла очень своевременно, учитывая текущую ситуацию с данными и аналитикой в организациях. Давайте кратко пробежимся по последним событиям. Начиная с 1970-х годов лишь немногие передовые компании эффективно использовали данные и аналитику для принятия решений и обоснования своих действий. Большинство игнорировало этот ценный ресурс или не придавало ему особого значения.

В 2000-х годах ситуация стала меняться, и компании начали понимать, как они могут изменить свою ситуацию с помощью данных и аналитики. К началу 2010-х годов интерес стал смещаться в сторону «больших данных», которые изначально появились в интернет-компаниях, а затем распространились по всей экономике. В связи с возросшим объемом и сложностью данных в компаниях возникла роль «дата-сайентиста», опять же, сначала в Силиконовой долине, а затем повсюду.

Однако как только фирмы начали приспосабливаться к большим данным, в период с 2015 по 2018 год акцент во многих фирмах снова сместился, на этот раз в сторону искусственного интеллекта. Сбор, хранение и анализ больших данных уступили место машинному обучению, обработке естественного языка и автоматизации.

В основе этих быстрых сдвигов фокуса лежал ряд допущений относительно данных и аналитики, распространенных внутри организаций. Я рад сообщить, что книга «Разберись в Data Science» разрушает многие из них и делает это весьма своевременно. Многие люди, внимательно наблюдающие за этими тенденциями, уже начинают признавать, что эти допущения направляют нас по непродуктивному пути. В оставшейся части этого предисловия я опишу пять взаимосвязанных допущений и то, как изложенные в этой книге идеи обоснованно опровергают их.

Допущение 1. Аналитика, большие данные и ИИ — совершенно разные явления.

Многие полагают, что «традиционная» аналитика, большие данные и ИИ — это отдельные явления. Однако авторы книги «Разберись в Data Science»

справедливо считают, что эти вещи тесно связаны друг с другом. Все они требуют статистического мышления, использования традиционных аналитических подходов, вроде регрессионного анализа, а также методов визуализации данных. Предиктивная аналитика — это, по сути, то же самое, что и контролируемое машинное обучение. Кроме того, большинство методов анализа данных работают с наборами данных любого размера. Короче говоря, главный по данным может эффективно работать во всех трех областях, так что заострять внимание на различиях между ними не очень продуктивно.

Допущение 2. В этой песочнице могут играть только дата-сайентисты. Мы часто прославляли дата-сайентистов, полагая, что только они способны эффективно работать с данными и аналитикой. Тем не менее в настоящее время зарождается важная тенденция к демократизации этих идей, и все больше организаций расширяют полномочия «гражданских специалистов по работе с данным». Автоматизированные инструменты машинного обучения упрощают создание моделей, которые отлично справляются с прогнозированием. Разумеется, нам все еще нужны профессиональные дата-сайентисты для разработки новых алгоритмов и проверки работы гражданских специалистов, занимающихся сложным анализом. Однако организации, которые демократизируют занятие аналитикой и наукой о данных, привлекая к этому «любителей», способны значительно расширить использование этих важных возможностей.

Допущение 3. Дата-сайентисты — это единороги, обладающими всеми необходимыми навыками.
Мы привыкли полагать, что дата-сайентисты, умеющие разрабатывать модели, также способны решать все остальные задачи, связанные с внедрением этих моделей. Другими словами, мы считаем их своеобразными «единорогами», которые могут все. Но таких «единорогов» нет вообще, или они существуют лишь в небольшом количестве. Главные по данным, которые понимают не только основы науки о данных, но и особенности бизнеса, а также способны эффективно управлять проектами и выстраивать деловые отношения, будут чрезвычайно ценны как участники проектов по работе с данными. Они могут стать продуктивными членами команд дата-сайентистов и повысить вероятность того, что проекты по работе с данными принесут бизнесу пользу.

Допущение 4. Чтобы преуспеть в работе с данными и аналитикой, вам необходимы выдающиеся математические способности и много тренировок.

Еще одно похожее допущение сводится к тому, что для работы с данными человек должен быть очень хорошо подготовлен в этой области, а также хорошо разбираться в математике. Математические способности и подготовка, безусловно, очень важны, но авторы книги «Разберись в Data Science» утверждают (и я с ними согласен), что мотивированный ученик способен освоить необходимые навыки в достаточной степени для того, чтобы стать полезным участником проектов по работе с данными. Во-первых, общие принципы статистического анализа далеко не так сложны, как может показаться. Во-вторых, для того, чтобы «быть полезным» участником проектов по работе с данными, ваш уровень владения аналитикой не обязательно должен быть чрезвычайно высоким. Работа с профессиональными дата-сайентистами или автоматизированными ИИ-программами требует лишь любознательности и умения задавать хорошие вопросы, находить взаимосвязи между бизнес-проблемами и количественными результатами, а также обращать внимание на сомнительные предположения.

Допущение 5. Если в колледже или аспирантуре вы не занимались в основном количественными предметами, вам слишком поздно осваивать навыки, необходимые для работы с данными и аналитикой.

Это предположение подтверждается данными опросов. Согласно результатам опроса, проведенного компанией Splunk в 2019 году, в котором приняли участие около 1300 руководителей по всему миру, практически каждый респондент (98%) согласен с тем, что навыки работы с данными важны для специалистов будущего[1]. А 81% респондентов считает, что навыки работы с данными необходимы для того, чтобы стать старшим руководителем в их компаниях, а 85% согласны с тем, что ценность таких навыков в их фирмах будет расти. Тем не менее 67% респондентов заявили, что им неудобно получать доступ к данным или использовать их самостоятельно, 73% считают, что навыки работы с данными труднее освоить, чем другие бизнес-навыки, а 53% — что они слишком стары для освоения навыков работы с данными. Подобное пораженчество наносит ущерб как отдельным лицам,

[1] Splunk Inc., “The State of Dark Data,” 2019, **www.splunk.com/en_us/form/the-state-of-dark-data.html**.

так и организациям в целом, и ни авторы этой книги, ни я не считаем его оправданным. В ходе чтения этой книги вы увидите, что в этом нет ничего сложного!

Итак, отбросьте эти ложные допущения и станьте главным по данным. Это позволит вам повысить свою ценность как сотрудника и сделать свою организацию более успешной. Именно по этому пути движется мир, так что пришло время узнать больше о данных и аналитике. Я уверен, что процесс чтения книги «Разберись в Data Science» окажется гораздо более полезным и приятным, чем вы можете себе представить.

Томас Х. Дэвенпорт
Заслуженный профессор Бэбсон-колледжа, приглашенный профессор Бизнес-школы Саида при Оксфордском университете, научный сотрудник инициативы Массачусетского технологического института в сфере цифровой экономики, автор книг «Аналитика как конкурентное преимущество», «Внедрение искусственного интеллекта в бизнес-практику: Преимущества и сложности» и «Big Data @ Work»

Введение

Данные — это, пожалуй, важнейший аспект вашей работы, нравится вам это или нет. И, скорее всего, вы решили прочитать эту книгу, чтобы лучше в них разобраться.

Для начала стоит констатировать то, что уже почти превратилось в клише: в настоящее время мы создаем и потребляем больше информации, чем когда-либо прежде. Мы, без сомнения, живем в эпоху данных, которая породила массу обещаний, модных словечек и продуктов, многие из которых вы, ваши менеджеры, коллеги и подчиненные уже используете или будете использовать. Однако, несмотря на распространение этих обещаний и продуктов, проекты по работе с данными терпят неудачу с пугающей регулярностью[2].

Разумеется, мы не утверждаем, что все обещания пусты, а продукты — ужасны. Скорее, чтобы по-настоящему разобраться в этой области, вы должны принять фундаментальную истину: работа с данными очень сложна и сопряжена с нюансами и неопределенностью. Данные, безусловно, важны, но работать с ними совсем не просто. И все же существует целая индустрия, которая заставляет нас думать иначе, обещает определенность в мире неопределенности и играет на страхе компаний упустить выгоду. Мы называем это промышленным комплексом науки о данных.

ПРОМЫШЛЕННЫЙ КОМПЛЕКС НАУКИ О ДАННЫХ

Эта проблема касается всех. Компании бесконечно ищут продукты, которые думали бы за них. Менеджеры нанимают профессионалов в области аналитики, которые на самом деле таковыми не являются. Дата-сайентистов нанимают для работы в компаниях, которые к ним не готовы. Руководители вынуждены слушать техническую болтовню и делать вид, что понимают, о чем идет речь. Работа над проектами стопорится. Деньги тратятся впустую.

---

[2] Venture Beat. “87% of data science projects failing”: **venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production**