Data Science

для карьериста

Жаклин Нолис
Эмили Робинсон

MANNING

Библиотека программиста (Питер)

Жаклин Нолис

Data Science для карьериста

«Питер»
2020

УДК 004.62
ББК 32.973.233.02

Нолис Ж.

Data Science для карьериста / Ж. Нолис — «Питер», 2020 — (Библиотека программиста (Питер))

ISBN 978-5-4461-1734-5

Все мы хотим построить успешную карьеру. Как найти ключ к долгосрочному успеху в Data Science? Для этого понадобятся не только технические ноу-хау, но и правильные «мягкие навыки». Лишь объединив оба этих компонента, можно стать востребованным специалистом.Узнайте, как получить первую работу в Data Science и превратиться в ценного сотрудника высокого уровня! Четкие и простые инструкции научат вас составлять потрясающие резюме и легко проходить самые сложные интервью. Data Science стремительно меняется, поэтому поддерживать стабильную работу проектов, адаптировать их к потребностям компании и работать со сложными стейкхолдерами не так уж и легко. Опытные дата-сайентисты делятся идеями, которые помогут реализовать ваши ожидания, справиться с неудачами и спланировать карьерный путь. В формате PDF A4 сохранен издательский макет.

УДК 004.62
ББК 32.973.233.02

ISBN 978-5-4461-1734-5

Содержание

Жаклин Нолис, Эмили Робинсон
Data Science для карьериста

От Эмили для Майкла
и
от Жаклин для Хизер, Эмбер и Лауры
за любовь и поддержку, которую вы давали нам на всем этом пути

Переводчик *А. Попова*

Предисловие

«Как мне устроиться на такую же работу, как у вас?»

Нам как опытным дата-сайентистам постоянно задают этот вопрос. Порой он звучит прямо, а в других случаях нас спрашивают о том, какие решения мы принимали в течение карьерного пути, чтобы оказаться на этом месте. На самом деле люди, задающие подобные вопросы, постоянно испытывают трудности, так как ресурсов, объясняющих, как встать на путь Data Science или расти профессионально в этом направлении, очень мало. Многие дата-сайентисты ищут помощь по вопросам карьеры, но зачастую не находят внятных ответов.

Хотя в блогах мы постили тактические советы о том, что делать в определенные моменты работы в Data Science (DS), мы также решили разобраться с отсутствием адекватного текста, описывающего весь карьерный путь в этой области от начала до конца. Эта книга призвана помочь тысячам людей, которые слышат о Data Science и о машинном обучении, но не знают, с чего начать, а также тем, кто уже занят в этой области и хочет понять, как продвинуться по карьерной лестнице.

Мы были рады возможности поучаствовать в создании этой книги. Нам обеим казалось, что наш опыт и точки зрения дополняли друг друга и помогли в написании лучшей книги для вас. Мы – это:

• *Жаклин Нолис* (Jacqueline Nolis). Я получила степень бакалавра и магистра математических наук, а также кандидатскую степень в области исследования операций. Когда я начинала работать, такого понятия, как *Data Science* (*DS*), еще не было, и мне пришлось выстраивать свой карьерный путь одновременно с попытками определения этой области. Теперь я работаю консультантом и помогаю компаниям растить команды, занимающиеся DS.

• *Эмили Робинсон* (Emily Robinson). Я получила степень бакалавра в области теории принятия решений и степень магистра менеджмента. Окончив трехмесячный курс по Data Science в 2016 году, я начала работать в этой сфере, специализируясь на A/B-тестировании. Сейчас я работаю старшим дата-сайтентистом в компании Warby Parker и занимаюсь некоторыми проектами компании.

На своем карьерном пути мы создавали портфолио проектов и испытывали стресс от адаптации на новой работе. Когда нас не брали на желаемую должность, нам было обидно. Когда наш анализ положительно влиял на бизнес, мы торжествовали. Мы сталкивались с проблемами, работая со сложными деловыми партнерами, и нам помогали наставники, оказывающие поддержку. Хотя этот опыт многому нас научил, истинная ценность заключается в том, чтобы делиться этим опытом с другими.

Цель этой книги – стать руководством по вопросам карьеры в области Data Science. Она описывает путь, который человек пройдет, работая в этом направлении. Мы начнем с азов: расскажем, как получить базовые навыки и понять, что на самом деле представляют собой направления работы в DS. Затем мы объясним, как эту работу получить и освоиться на новом месте. Расскажем, как вырасти в должности и в конечном итоге стать руководителем или уйти в другую компанию. Мы намерены сделать эту книгу ресурсом, к которому дата-сайентисты будут возвращаться на новых этапах своей карьеры.

Поскольку основное внимание в этой книге уделено карьере, мы решили не заострять внимание на технических аспектах Data Science. Мы не будем обсуждать выбор гиперпараметров модели или нюансы пакетов Python. Здесь не будет ни одного уравнения или строчки кода – мы знаем, что об этом уже написано множество замечательных книг. Мы же, напротив, хотели обсудить часто упускаемые из виду, но не менее важные нетехнические знания, которые нужны для достижения успеха.

Мы включили в эту книгу много подробностей из личного опыта уважаемых дата-сайентистов. В конце каждой главы вы найдете интервью с реальными специалистами. Они расскажут, как справлялись с трудностями, рассматриваемыми в главе. Мы были очень рады получить удивительные, подробные и откровенные ответы этих людей и считаем, что их примеры из жизни могут научить гораздо большему, чем любое заявление, которое мы могли бы написать.

При написании этой книги мы намеренно решили сосредоточиться на уроках, которые извлекли, будучи профессионалами в области Data Science, а также общаясь с другими членами сообщества. Иногда мы заявляем о чем-нибудь, с чем не все могут согласиться, например предлагаем всегда писать сопроводительное письмо при поиске работы. Мы решили, что поделиться мнениями, которые, на наш взгляд, будут полезными для дата-сайентистов, важнее, чем пытаться написать что-либо содержащее только объективные истины.

Мы надеемся, что эта книга станет для вас полезным руководством в построении карьеры в области Data Science. Когда мы сами были начинающими специалистами, нам не хватало такой книги. Зато теперь она есть у вас.

Благодарности

Прежде всего хотели бы поблагодарить наших супругов Майкла Берковица (Michael Berkowitz) и Хизер Нолис (Heather Nolis). Без них эта книга не появилась бы (не только потому, что Майкл писал первые черновики некоторых разделов, несмотря на то что он профессиональный игрок в бридж, а вовсе не дата-сайентист, и не потому, что Хизер стремилась заполнить половину книги контентом о машинном обучении).

Хотим поблагодарить сотрудников компании Manning, которые помогли нам пройти этот путь, улучшили книгу и вообще сделали ее выход возможным. Особая благодарность нашему редактору Карен Миллер (Karen Miller), которая помогала нам придерживаться графика и координировала работу.

Спасибо всем редакторам, которые читали рукопись на разных этапах и давали неоценимые подробные отзывы. Вот их имена: Бринджар Смари Бьярнасон (Brynjar Smári Bjarnason), Кристиан Таудал (Christian Thoudahl), Даниэль Берец (Daniel Berecz), Доменико Наппо (Domenico Nappo), Джефф Барто (Geoff Barto), Густаво Гомес (Gustavo Gomes), Хагай Люгер (Hagai Luger), Джеймс Риттер (James Ritter), Джефф Ньюман (Jeff Neumann), Джонатан Твадделл (Jonathan Twaddell), Кшиштоф Енджеевский (Krzysztof Jedrzejewski), Малгожата Родацка (Malgorzata Rodacka), Марио Гизель (Mario Giesel), Нараяна Лалитананд Сурампуди (Narayana Lalitanand Surampudi), Пин Чжао (Ping Zhao), Риккардо Маротти (Riccardo Marotti), Ричард Тобиас (Richard Tobias), Себастьян Пальма Мардонес (Sebastian Palma Mardones), Стив Сассман (Steve Sussman), Тони М. Дубицкий (Tony M. Dubitsky) и Юл Вильямс (Yul Williams). Спасибо также нашим друзьям и членам семьи, которые прочитали книгу и внесли свои предложения: Элин Фарнелл (Elin Farnell), Аманда Листон (Amanda Liston), Кристиан Рой (Christian Roy), Джонатан Гудман (Jonathan Goodman) и Эрик Робинсон (Eric Robinson). Ваш вклад помог оформить эту книгу и сделать ее максимально полезной для наших читателей.

Наконец, хотим поблагодарить всех, кто согласился дать нам интервью: Роберт Чанг (Robert Chang), Рэнди Ау (Randy Au), Джулия Силдж (Julia Silge), Дэвид Робинсон (David Robinson), Джесси Мостипак (Jesse Mostipak), Кристен Керер (Kristen Kehrer), Райан Уильямс (Ryan Williams), Брук Уотсон Мадубуонву (Brooke Watson Madubuonwu), Джарвис Миллер (Jarvis Miller), Хилари Паркер (Hilary Parker), Хизер Нолис (Heather Nolis), Сейд Сноуден-Акинтунде (Sade Snowden-Akintunde), Мишель Кейм (Michelle Keim), Рене Теате (Renee Teate), Аманда Касари (Amanda Casari) и Анджела Басса (Angela Bassa). Кроме того, мы благодарны тем, кто участвовал в создании примечаний на протяжении всей книги и предлагал вопросы для интервью в приложении: Вики Бойкис (Vicki Boykis), Родриго Фуэнтеальба Картес (Rodrigo Fuentealba Cartes), Густаво Коэльо (Gustavo Coelho), Эмили Барта (Emily Bartha), Трей Кози (Trey Causey), Элин Фарнелл (Elin Farnell), Джефф Аллен (Jeff Allen), Элизабет Хантер (Elizabeth Hunter), Сэм Бэрроуз (Sam Barrows), Решама Шейх (Reshama Shaikh), Габриэлла де Кьерос (Gabriela de Queiroz), Роб Штамм (Rob Stamm), Алекс Хейз (Alex Hayes), Людамила Джанда (Ludamila Janda), Аянти Дж. (Ayanthi G.), Аллан Батлер (Allan Butler), Хизер Нолис (Heather Nolis), Йерун Янссенс (Jeroen Janssens), Эмили Спан (Emily Spahn), Тереза Иофчиу (Tereza Iofciu), Бертил Хатт (Bertil Hatt), Райан Уильямс (Ryan Williams), Питер Болдридж (Peter Baldridge) и Хлинур Хадльгримссон (Hlynur Hallgrímsson). Все эти люди предоставили ценную информацию, и вместе они знают гораздо больше, чем мы.

О книге

Книга «Data Science для карьериста» поможет вам войти в сферу DS и стать профессионалом. В ней рассказывается том, кто такие дата-сайентисты, как получить необходимые навыки и какие шаги нужно предпринять, чтобы устроиться на работу. После трудоустройства эта книга поможет вам понять, как развиваться в своей должности и стать в итоге частью сообщества Data Science, а также дорасти до уровня старшего специалиста. Прочитав ее, вы станете уверенно смотреть на предстоящий карьерный путь.

Для кого эта книга

Эта книга предназначена для людей, которые еще не начали работать в Data Science, но в перспективе рассматривают такую возможность, а также для тех, кто только начал трудиться в этой сфере. Начинающие специалисты получат навыки, которые необходимы, чтобы стать дата-сайентистами, а джуниоры узнают, как повысить свою экспертность. Многие темы в книге вроде прохождения интервью и обсуждения оффера – это полезные ресурсы, к которым стоит возвращаться на любом этапе карьерного пути.

Структура книги

Эта книга разбита на четыре части, посвященные этапам, которые проходит начинающий дата-сайентист. В первой части книги, «Data Science. С чего начать», рассказывается о том, что такое DS и какие навыки нужны для работы в этой сфере:

• В главе 1 вы узнаете о функциях дата-сайентиста, а также о различных должностях с аналогичным названием.

• В главе 2 представлено пять примеров компаний, в которых трудятся дата-сайентисты, и показано, как культура и тип каждой из них влияют на работу.

• Глава 3 описывает различные пути, которые можно выбрать для получения важных для дата-сайентиста навыков.

• Из главы 4 вы узнаете, как создавать проекты и делиться ими для создания портфолио.

Во второй части книги, «Как попасть в Data Science», объясняется весь процесс поиска вакансий:

• В главе 5 рассказывается о поиске вакансий и о том, как понять, ради каких из них стоит стараться.

• В главе 6 мы расскажем, как написать сопроводительное письмо и составить резюме, а затем скорректировать их под каждую конкретную вакансию.

• В главе 7 подробно описывается, как проходит интервью и чего от него следует ожидать.

• Из главы 8 вы узнаете, что делать после того, как получен оффер, и как обсуждать его детали.

В третьей части, «Осваиваемся в Data Science», рассматриваются основные моменты первых месяцев работы:

• В главе 9 рассказывается о том, чего следует ожидать в первые несколько месяцев работы в Data Science, а также о том, как провести это время максимально продуктивно.

• В главе 10 рассматривается процесс проведения анализа, являющегося ключевым компонентом большинства должностей в Data Science.

• Глава 11 фокусируется на внедрении моделей машинного обучения, что является необходимым для специалистов, занимающих инженерные должности.

• В главе 12 объясняется, как общаться со стейкхолдерами, – дата-сайентисты занимаются этим чаще, чем большинство других технических специалистов.

В четвертой части, «Как подняться по карьерной лестнице в Data Science», рассматриваются темы для более опытных специалистов, которые ищут способ профессионально вырасти:

• Из главы 13 вы узнаете, что делать с неудавшимися проектами Data Science.

• В главе 14 показано, как стать частью более широкого сообщества дата-сайентистов с помощью участия в конференциях и разработки открытого исходного кода.

• Глава 15 представляет собой руководство по принятию сложного решения об уходе с должности специалиста Data Science.

• Глава 16 – заключительная; в ней рассказывается о должностях, которые могут получить дата-сайентисты по мере продвижения по карьерной лестнице.

Наконец, в приложении мы собрали для вас более 30 вопросов, которые можно услышать во время интервью, а также предложили примеры хороших ответов. Мы пояснили, какие навыки оцениваются при каждом вопросе и как на них лучше отвечать.

Если вы новичок в области Data Science, то начинайте читать с самого начала, а если вы уже работаете в этой сфере, то переходите сразу к той главе, которая предлагает решение вашей текущей задачи. Несмотря на то что последовательность глав соответствует развитию карьеры в этой сфере, их можно читать в произвольном порядке в соответствии с вашими потребностями.

В конце каждой главы – интервью со специалистами, занятыми в разных индустриях. Они рассказывают, как рассмотренные вопросы коснулись их в работе. Мы выбрали тех специалистов, которые внесли весомый вклад в развитие Data Science и которым пришлось пройти интересный путь прежде, чем стать профессионалами.

От издательства

Карьера в Data Science не зависит от страны, в которой вы живете и учитесь. Чтобы двигаться вперед, необходимо лучше понимать, чего от вас ждет работодатель или хедхантер.

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.

Об авторах

Эмили Робинсон

Написала Жаклин Нолис

Эмили Робинсон – блестящий старший дата-сайентист в компании Warby Parker; ранее она работала в DataCamp и Etsy.

Впервые я встретила Эмили на Data Day Texas 2018, когда она была одной из немногих слушательниц моего доклада о Data Science в индустрии. В конце моего выступления она подняла руку и задала прекрасный вопрос. К моему удивлению, через час мы поменялись местами – теперь уже я слушала, как она спокойно проводила восхитительную презентацию, и с нетерпением ждала возможности поднять руку и задать ей вопрос. В тот день я уже поняла, какой она трудолюбивый и умный специалист. Несколько месяцев спустя, когда пришло время искать соавтора для моей книги, Эмили Робинсон была первым кандидатом в списке на эту роль. Отправляя ей электронное письмо, я думала, что мне, скорее всего, откажут: она, пожалуй, была «не моего уровня».

Работа с Эмили над этой книгой была сплошным удовольствием. Она очень заботится о трудностях младших специалистов по работе с данными, а еще у нее есть способность четко выделять важное. Она всегда качественно выполняет свою работу и каким-то образом умудряется одновременно писать статьи в блогах. Наблюдая за ней на других конференциях и общественных мероприятиях, я видела, как она общалась со многими дата-сайентистами, каждый из которых чувствовал себя с ней комфортно. Она также является экспертом в области A/B-

тестирования и экспериментирования, хотя ясно, что для нее это просто временный этап. При желании она могла бы взять любую другую область DS и стать в ней экспертом.

Единственное, что меня расстраивает, так это то, что я пишу эти слова о ней на финальном этапе создания книги, и, как только мы закончим, возможность сотрудничать с Эмили появится уже у кого-то другого.

Жаклин Нолис

Написала Эмили Робинсон

Когда меня спрашивают о том, стоит ли писать книгу, я всегда отвечаю: «Только если у вас будет соавтор». Но это еще не все. Полный ответ должен быть таким: «Только если у вас будет такой же веселый, душевный, щедрый, умный, опытный и заботливый соавтор, как Жаклин». Я не знаю, каково писать книгу с «нормальным» соавтором, потому что Жаклин всегда была просто потрясающей, и мне невероятно повезло поработать с ней над этим проектом.

На фоне такого образованного человека, как Жаклин, вы запросто можете почувствовать себя неловко. У нее есть степень кандидата наук в промышленной инженерии и $100 000 за победу в третьем сезоне телевизионного реалити-шоу «Король ботанов». Жаклин работала директором по аналитике и основала собственное успешное консалтинговое агентство. Она выступает на конференциях по всей стране и регулярно получает приглашения от своей альма-матер приехать и провести карьерные консультации для студентов-математиков (ее специализация). Когда она выступает на онлайн-конференциях, ее забрасывают комплиментами вроде «это лучшее, что я когда-либо слышал», «превосходное выступление», «действительно полезно», «отличная живая презентация». Но Жаклин никогда не дает людям повода чувствовать себя недостойно или плохо из-за того, что они чего-то не знают; наоборот, она любит делать сложные понятия простыми, как, например, в ее презентации «Глубокое обучение – это нетрудно, даю слово».

Ее личная жизнь тоже впечатляет – у нее прекрасный яркий дом в Сиэттле, где она живет со своей подругой, сыном, двумя собаками и тремя кошками. Надеюсь, однажды она приютит соавтора, чтобы заполнить немного оставшегося места. Она со своей подругой Хизер даже провели презентацию перед аудиторией в тысячу человек об их опыте в использовании R для развертывания моделей машинного обучения в производство T-Mobile. А еще у них, пожалуй, самая милая история знакомства: они встретились на том самом шоу «Король ботанов», где Хизер также была участницей.

Я очень благодарна Жаклин за этот опыт, ведь она могла бы заработать гораздо больше, занимаясь чем-то гораздо менее утомительным, чем написание этой книги вместе со мной. Надеюсь, что наша работа подтолкнет начинающих дата-сайентистов стать частью сообщества людей, таких же прекрасных, как Жаклин.

Об обложке

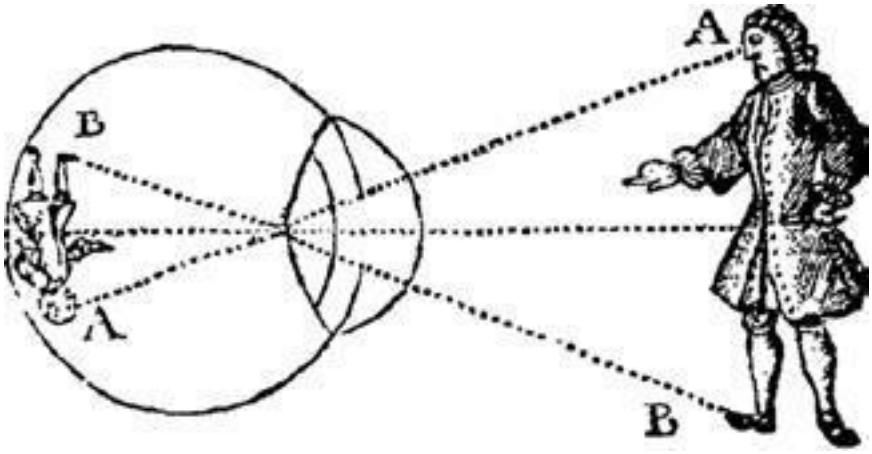
Сен-Совер

Рисунок на обложке книги называется «Femme de l'Aragon», или «Арагонская женщина». Иллюстрация позаимствована из книги Жака Грассе де Сен-Совера (1757–1810) «Костюмы разных стран» (фр. Costumes de Différents Pays), изданной во Франции в 1797 году. Каждая иллюстрация тщательно прорисована и раскрашена вручную. Богатое разнообразие коллекции Сен-Совера ярко отражает то, насколько далекими в культурном плане были города и регионы еще каких-то 200 лет назад. Будучи изолированными, люди говорили на разных языках и диалектах. На улицах городов и деревень по одежде можно было легко определить статус человека, его место жительства и род занятий.

С тех пор манера одеваться сильно изменилась, а разница между регионами, ранее такая заметная, практически исчезла. Сегодня различать жителей разных континентов стало гораздо труднее, не говоря уже о разных городах, регионах или странах. Возможно, мы отказались от культурного многообразия в пользу более разносторонней личной жизни – и уж точно в пользу более разнообразной и быстрой технологической жизни.

В то время когда большинство книг о компьютерах так похожи, издательство Manning отмечает изобретательность и инициативность компьютерного бизнеса с помощью книжных обложек, основанных на богатом разнообразии жизни регионов двухсотлетней давности, оживающей благодаря иллюстрациям Грассе де Сен-Совера.

Часть 1
Data Science. С чего начать

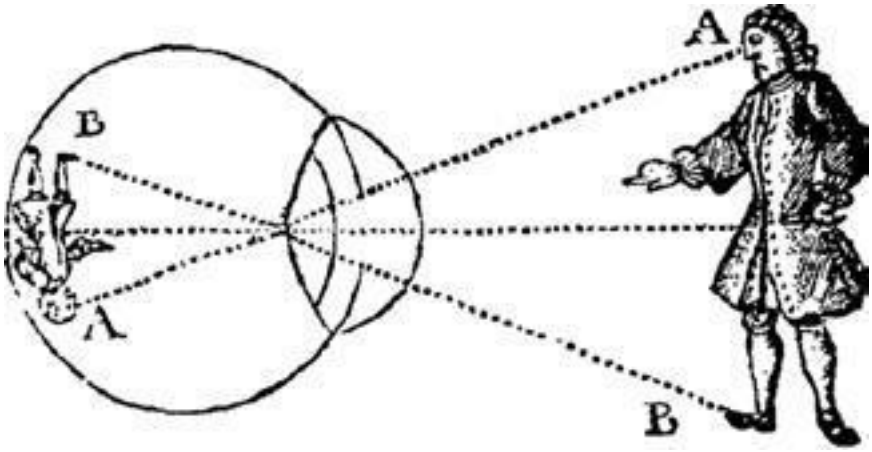


Если вы загуглите «как стать специалистом Data Science», перед вами, скорее всего, появится обширный список, содержащий навыки от статистического моделирования до программирования на Python, а также информация об эффективном общении и проведении презентаций. В одной вакансии может описываться роль, схожая с ролью специалиста по статистике, в то время как другой работодатель ищет кого-то с дипломом магистра информатики. Интернет вам предложит различные варианты приобретения нужных навыков – от возвращения в университет на магистерскую программу до прохождения учебного курса или практики анализа данных на текущем месте работы. В совокупности все эти способы могут показаться непреодолимыми, особенно для тех, кто еще до конца даже не определился с решением стать дата-сайентистом.

Для вас есть хорошая новость: не существует ни одного специалиста по Data Science, который обладал бы всеми этими навыками. У дата-сайентистов есть общий фундамент знаний, но каждый из них специализируется в конкретной области, причем настолько, что многие не смогут поменяться обязанностями. Первая часть этой книги призвана помочь вам разобраться во всех этих специализациях и в том, как принимать наилучшие решения для старта вашей карьеры. К концу у вас будет понимание того, как начать поиск работы.

В главе 1 раскрываются основы работы в Data Science, включая описание необходимых навыков и различных специализаций. В главе 2 подробно рассказывается о роли дата-сайентиста и о пяти типах компаний – это поможет вам лучше понять, на что будет похожа реальная работа. В главе 3 описываются различные пути приобретения навыков, а также преимущества и недостатки каждого из них. Из главы 4 вы узнаете, как создать портфолио как для практического опыта, так и для потенциальных работодателей.

1. Что такое Data Science?



В этой главе

- Три основных направления Data Science.
- Разные типы должностей в области Data Science.

«Самая сексуальная работа XXI века», «Лучшая работа в Америке»… Дата-сайентист – должность, названия которой даже не существовало до 2008 года, теперь является одной из самых востребованных среди соискателей, а работодатели не могут найти достаточное число подобных сотрудников. У такого ажиотажа есть веская причина: Data Science – это быстро развивающаяся область, медианная базовая зарплата специалистов которой в США в 2019 году составила более $100 000 (http://mng.bz/XpMp). В хорошей компании дата-сайентисты пользуются большой автономией и постоянно изучают что-то новое. Они используют свои знания для решения серьезных задач: например, работают с врачами во время испытаний лекарственных препаратов, помогают спортивной команде в подборе новобранцев или изменяют модель ценообразования для бизнеса по производству виджетов. Наконец, в главе 3 мы поговорим о том, что универсального способа стать дата-сайентистом нет. В эту сферу приходят люди с разным образованием, поэтому вы не ограничены своей бакалаврской специальностью.

Однако не вся работа в сфере DS идеальна. И у компаний, и у соискателей бывают нереалистичные ожидания. Например, компании, плохо знакомые с Data Science, могут считать, будто один человек может решить все их задачи с помощью данных. Когда дата-сайентист наконец принят на работу в такую компанию, он сталкивается с бесконечным списком дел. Ему могут поручить немедленно внедрить систему машинного обучения, при том что никакие работы по подготовке или очистке данных предварительно не проводились. Иногда случается так, что никто не может ему помочь, направить или хотя бы посочувствовать при возникновении проблем. Мы поговорим об этом подробнее в главах 5 и 7, где расскажем, как не оказаться в не подходящих для новичка компаниях, а в главе 9 посоветуем, что делать, если вы попали в неприятную ситуацию.

С другой стороны, соискатели могут подумать, что им никогда не придется скучать. Они могут рассчитывать на то, что стейкхолдеры будут просто следовать их советам, дата-инженеры смогут в мгновение ока исправить любые проблемы с качеством данных, а сами они получат самые быстрые вычислительные ресурсы из возможных для реализации своих моделей. На самом деле дата-сайентисты тратят много времени на очистку и подготовку данных, а также на организацию работы с учетом ожиданий и приоритетов других команд. Проекты не всегда оказываются удачными. Высшее руководство может давать клиентам нереалистичные обещания о работе ваших моделей. Основные обязанности могут заключаться в работе с архаичной систе-

мой данных, которую невозможно автоматизировать, – каждую неделю она будет требовать многочасового монотонного труда только на их очистку. Дата-сайентисты могут обнаружить множество статистических или технических ошибок с серьезными последствиями в предыдущих расчетах, но они не будут никого интересовать. При этом специалисты настолько перегружены работой, что им просто некогда что-либо исправлять. Дата-сайентиста могут попросить подготовить отчеты, подтверждающие решение руководства, поэтому он может беспокоиться о том, что его уволят в случае, если он предоставит независимое мнение.

Эта книга поможет вам пройти путь становления в качестве специалиста по Data Science и построить карьеру. Мы хотим, чтобы вы получили все преимущества работы в этой сфере и избежали большинства подводных камней. Возможно, вы работаете в смежной области вроде маркетинговой аналитики и подумываете сменить сферу деятельности. Или, может быть, вы уже работаете дата-сайентистом, но ищете новое место работы и полагаете, что подошли к предыдущему процессу поиска недостаточно хорошо. Возможно, вы хотите продолжить карьеру, выступая на конференциях, участвуя в разработке open source, или же стать независимым консультантом. Мы уверены, что, каким бы ни был ваш нынешний уровень, эта книга окажется вам полезной.

В первых четырех главах мы описали, как можно начать путь в Data Science и создать портфолио: так мы попытались решить парадокс, когда опыт можно получить только при изначальном владении практическими навыками. В части 2 мы покажем, как составить сопроводительное письмо и резюме, с которыми вас точно пригласят на собеседование, и расскажем, как создать сеть контактов для получения рекомендации. Мы также рассмотрим стратегии переговоров, которые, как показывают исследования, позволят вам получить наилучшие условия оффера.

Как дата-сайентисту вам необходимо будет разрабатывать методы анализа, взаимодействовать со стейкхолдерами и, возможно, даже участвовать в развертывании модели в производство. Часть 3 поможет понять, как устроены все эти процессы и как можно самому настроиться на успех. В части 4 вы найдете стратегии, которые помогут вам собраться с силами в тех неизбежных случаях, когда ваш проект терпит крах. А когда вы будете готовы, мы поможем вам решить, как продолжать свою карьеру – стать менеджером, остаться исполнителем или даже стать независимым консультантом.

Однако прежде, чем начать этот путь, вы должны разобраться в том, кто такие дата-сайентисты и какую работу они выполняют. Data Science – это очень широкое поле деятельности, которое включает в себя много направлений, и чем лучше вы понимаете разницу между ними, тем успешнее вы сможете в них развиваться.

1.1. Что такое Data Science?

Data Science (DS) – это практика использования данных, с помощью которой можно попытаться понять и решить реальные задачи. Эта концепция не нова; люди анализируют объемы и тенденции продаж с тех пор, как изобрели ноль. Однако за последнее десятилетие нам стало доступно экспоненциально большее количество данных, чем прежде. Появление компьютеров помогло генерировать их, и только путем машинных вычислений можно обрабатывать так много информации. С помощью компьютерного кода дата-сайентист может преобразовывать или накапливать данные, проводить статистический анализ или тренировать модели машинного обучения (МО). В результате могут быть созданы отчет, информационная панель или модель МО, которую можно будет запустить в непрерывную работу.

Например, если розничная компания не может определиться с местом для нового магазина, она может пригласить дата-сайентиста для проведения соответствующего анализа. Он соберет статистические данные об адресах доставки онлайн-заказов, чтобы понять, где нахо-

дится потребительский спрос. Специалист также может совмещать выводы о местонахождении клиентов с информацией о демографической ситуации и доходах в этих местах на основании данных переписи населения. С помощью этих датасетов можно найти оптимальное место для нового магазина и создать презентацию Microsoft PowerPoint, чтобы представить рекомендации вице-президенту компании по коммерческой деятельности.

В другой ситуации та же розничная компания захочет увеличить объем онлайн-заказов с помощью персональных рекомендаций во время шоппинга. Дата-сайентист может загрузить статистику прежних онлайн-заказов и создать модель машинного обучения, которая будет учитывать набор товаров в корзине покупателя и на его основании прогнозировать, что еще ему можно предложить. После этого он будет работать с командой инженеров компании, чтобы каждый раз, когда клиент совершает покупки, новая модель МО показывала рекомендуемые товары.

При попытке освоить сферу DS многие люди сталкиваются с одной проблемой: слишком уж много нужно изучить. Например, программирование (но какой язык?), статистику (но какие методы наиболее важны на практике, а какие в основном академические?), машинное обучение (но чем оно отличается от статистики или ИИ?) и предметную область в той отрасли, в которой они хотят работать (но что, если вы не знаете, где хотите работать?). Кроме того, им необходимо овладеть бизнес-навыками вроде эффективной презентации результатов всем, начиная с других дата-сайентистов и заканчивая генеральным директором. А от вакансий, в которых требуется степень кандидата наук, многолетний опыт работы в Data Science и знание обширного перечня статистических и программных методов, становится только хуже. Как можно приобрести все эти навыки? С чего лучше начать? Что входит в базу?

Если вы изучали различные области DS, возможно, вы знакомы с популярной диаграммой Венна, составленной Дрю Конвеем. По мнению Конвея (на момент создания диаграммы), Data Science находится на пересечении математики и статистики, знаний предметной области и навыков хакинга (то есть программирования). Это изображение часто берется за основу для определения того, кто такой специалист по работе с данными. На наш взгляд, компоненты науки о данных немного отличаются от того, что предложил Дрю Конвей (рис. 1.1).

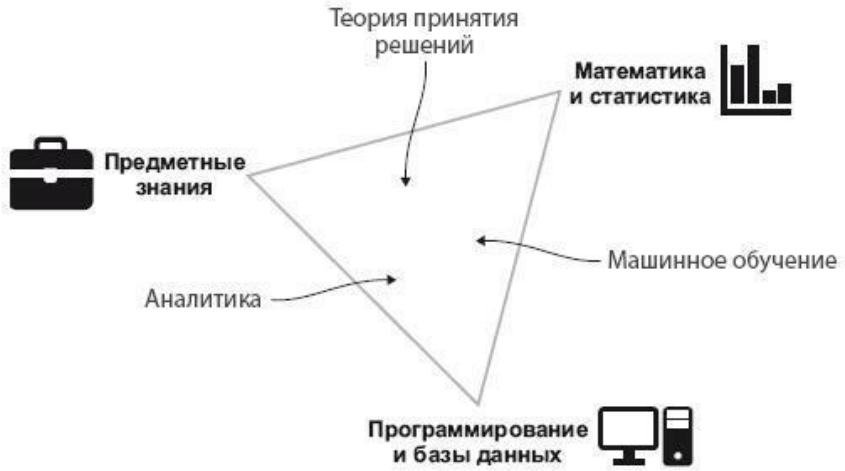


Рис. 1.1. Навыки, которые объединяются в DS, и то, как они сочетаются для выполнения разных функций

Мы изменили исходную диаграмму Венна, составленную Конвеем, на треугольник, потому что дело не в том, есть ли у вас навык или нет, а в том, что вы можете развить его лучше,

чем другие специалисты. Действительно, все три навыка являются фундаментальными и вам необходимо владеть каждым в определенной степени, но вам не обязательно быть экспертом во всех. Мы поместили в треугольник разные типы специальностей в сфере Data Science. Они не всегда однозначно соответствуют названиям должностей, а даже если и так, то в разных компаниях их названия могут отличаться. Итак, что означает каждый из этих компонентов?

1.1.1. Математика/статистика

На начальном уровне математика и статистика являются базой в работе с данными. Мы разделяем эту базу на три уровня знания:

• *Существование методов*. Если вы не знаете о какой-либо возможности, вы не можете ее использовать. Если дата-сайентисту нужно сгруппировать похожих клиентов, знание того, что это можно сделать статистическим методом (с помощью *кластерного анализа*), станет первым шагом.

• *Как применять методы*. Специалист по работе с данными должен не просто знать много методов – он должен различать нюансы их применения. Важно писать такой код, где они не только применяются, но и настраиваются. Если дата-сайентист хочет использовать кластеризацию методом *k*-средних, чтобы сгруппировать покупателей, он должен уметь делать это на языке программирования типа R или Python. Также он должен понимать, как настроить параметры метода, например как выбрать количество создаваемых групп.

• *Как выбрать подходящий метод*. В DS используется огромное количество методов, поэтому для дата-сайентиста важно быстро оценить, какой из них будет самым эффективным в каждом случае. В нашем примере с группировкой покупателей, даже если специалист сосредоточился на кластеризации, он может применять десятки различных методов и алгоритмов. Вместо того чтобы перебирать все доступные методы, он должен сразу отбросить бóльшую их часть и сосредоточиться всего на нескольких.

Эти типы навыков постоянно применяются в задачах по работе с данными. Приведем другой пример. Предположим, вы работаете в компании, занимающейся e-commerce. Ваш бизнес-партнер может поинтересоваться, в каких странах у вас самый большой средний чек. Это очень простой вопрос, если у вас есть готовые данные. Но вместо того, чтобы просто предоставить информацию и позволить партнеру делать выводы самостоятельно, вы можете копнуть глубже. Если у вас есть один заказ из страны А на $100 и тысяча заказов из страны Б средней стоимостью $75, то формально в стране А средний чек выше. Но можете ли вы с уверенностью сказать, что ваш бизнес-партнер должен вложиться в рекламу в стране А, чтобы увеличить количество заказов? Вряд ли. У вас есть только одна единица данных из этой страны, и она может оказаться статистически незначимой. А вот если бы у вас было 500 заказов из страны А, можно было бы протестировать разницу в стоимости заказов. Это значит, что, если бы эти показатели для стран А и Б действительно не различались, вы бы не получили прежний результат. В этом длинном примере дается оценка того, какие подходы были разумными, что следует учитывать и какие результаты были признаны несущественными.

1.1.2. Базы данных и программирование

Программирование и базы данных (БД) основываются на извлечении информации из БД компаний и написании чистого, эффективного, легко настраиваемого кода. Эти навыки во многом схожи с тем, что должен знать разработчик программного обеспечения. Вот только дата-сайентисты должны писать код, который выполняет анализ с неизвестным итогом, а не выдает заранее заданный результат. Стек данных каждой компании уникален, поэтому какой-то определенный набор технических знаний специалисту не нужен. В целом вам нужно уметь получать данные из базы, очищать их, обрабатывать, обобщать, визуализировать и обмениваться ими.

R и Python – основные языки программирования для большинства профессий DS. R берет свое начало в статистике и, как правило, лучше всего подходит для статистического анализа, моделирования, визуализации и составления отчетов. Python создавался как язык для разработки программного обеспечения и в дальнейшем приобрел огромную популярность в обработке данных. Python лучше R справляется с обработкой больших датасетов, проводит машинное обучение и поддерживает алгоритмы, работающие в реальном времени (например, модули рекомендаций в Amazon). Но благодаря вкладу многих участников возможности двух языков сейчас почти равны. Специалисты по работе с данными успешно используют R для создания моделей машинного обучения, запускаемых миллионы раз в неделю, а также делают чистый, презентабельный статистический анализ на Python.

R и Python наиболее популярны для обработки данных по нескольким причинам:

• Они бесплатны, и у них открытый исходный код. Это означает, что он создается многими участниками, а не одной определенной компанией или группой пользователей. В этих языках есть много пакетов, или *библиотек* (готовых блоков кода), которые можно использовать для сбора данных, их обработки, визуализации, статистического анализа и машинного обучения.

• Благодаря большому количеству пользователей каждого из этих языков дата-сайентистам легко найти помощь при возникновении проблем. И хотя в каких-то компаниях до сих пор используют SAS, SPSS, STATA, MATLAB или другие платные приложения, многие из них начинают переходить в своей работе на R или Python.

Хотя бóльшая часть анализа при обработке данных осуществляется на R или Python, часто приходится извлекать информацию из БД, и здесь на сцену выходит язык SQL. SQL – это язык программирования, который используется в большинстве БД для внутренней обработки данных и извлечения их из базы. Представим для примера дата-сайентиста, которому нужно проанализировать сотни миллионов записей о заказах клиентов компании и спрогнозировать, как со временем будет изменяться ежедневное количество заказов. Для начала он, скорее всего, напишет SQL-запрос для получения количества заказов за каждый день, после чего возьмет полученные данные и запустит статистический прогноз на R или Python. По этой причине SQL очень популярен в Data Science, и без знания этого языка вы далеко не продвинетесь.

Можно ли стать дата-сайентистом без программирования?

С данными можно успешно проделывать много вещей, используя только Excel, Tableau или другие BI-инструменты с графическими интерфейсами. Хотя код в них не пишется, часто заявляется, что этот софт так же функционален, как и программирование на R или Python. На самом деле многие дата-сайентисты действительно порой пользуются этими программами. Но могут ли они быть исчерпывающим набором инструментов? Мы говорим «нет». В реальности компаний, где DS-командам не приходится писать код, очень мало. Но даже если вам повезет оказаться в одной из них, у программирования все же есть ряд плюсов.

Первое преимущество программирования – воспроизводимость. Когда вы пишете код, а не пользуетесь программным обеспечением типа point-and-click, можно повторно запускать его при изменении данных хоть каждый день, хоть через полгода. Это преимущество также связано с контролем версий: вместо того чтобы переименовывать файл каждый раз при изменении кода, можно сохранить один файл и видеть всю его историю.

Второе преимущество – гибкость. Например, если в Tableau нет нужного вам типа графа, вы не сможете его создать. Но с помощью программирования можно написать собственный код, чтобы сделать то, о чем создатели и разработчики программных средств никогда даже не думали.

Третье и последнее преимущество языков с открытым исходным кодом, таких как Python и R, – это вклад в сообщество. Тысячи людей создают *пакеты* и публикуют их в открытом доступе на GitHub и/или CRAN (для R) и pip (для Python). Этот код можно скачать и использовать для решения своих задач. Так вы не зависите от числа функций, предлагаемых одной компанией или группой людей.

Другой ключевой навык – использование *контроля версий* для отслеживания изменений кода. Он позволяет организовать хранение файлов, выполнять откат до предыдущих версий и видеть, кто, когда и какие изменения вносил в файл. Этот навык чрезвычайно важен в Data Science и в разработке программного обеспечения. Например, если кто-то случайно изменил файл и испортил ваш код, вы можете восстановить его или посмотреть, что изменилось.

Безусловно, наиболее популярная система для контроля версий – это Git. Он часто используется вместе с GitHub, веб-службой хостинга для Git. Git позволяет сохранять (*фиксировать*) вносимые изменения, а также видеть всю историю проекта и то, как она менялась с каждой фиксацией. Если два человека по отдельности работают над одним и тем же файлом, Git гарантирует, что чья-либо работа не будет случайно удалена или перезаписана. Если вы захотите поделиться своим кодом или запустить что-то в производство, во многих компаниях вам обязательно потребуется Git, особенно если это компания с сильной командой проектировщиков.

1.1.3. Понимание бизнеса

Любая достаточно развитая технология неотличима от магии.
Артур Чарльз Кларк

У компаний, мягко говоря, разное понимание того, как работает Data Science. Часто руководство просто хочет решить определенную задачу и обращается к своим волшебникам DS. Основной навык, необходимый в Data Science, – это умение преобразовать бизнес-ситуацию в вопрос о данных, найти ответ на их основе и предоставить бизнес-решение. Бизнесмен может спросить: «Почему наши клиенты уходят?» Но у Python нет импортируемого пакета «почему уходят клиенты» – вы сами должны понять, как ответить на этот вопрос с помощью данных.

Понимание бизнеса – это та грань, где ваши идеальные представления о Data Science встречаются с условиями реального мира. Недостаточно просто запросить информацию, не зная, как данные хранятся и обновляются в конкретной компании. Если компания предоставляет услуги по подписке, то где хранятся данные? Что произойдет, если кто-то изменит свою подписку? Обновляется ли строка этого пользователя или в таблицу добавляется еще одна? Нужно ли вам исправить какие-либо ошибки или несоответствия в данных? Если вы не знаете всего этого, вы не сможете дать точный ответ на такой простой вопрос, как: «Сколько у нас было подписчиков на 2 марта 2019 года?»

Понимание бизнеса также помогает задавать правильные вопросы. Когда стейкхолдер спрашивает вас, что делать дальше, вероятно, он имеет в виду: «Почему у нас нет больше денег?» Для ответа приходится задавать встречные вопросы. Если вы понимаете основной бизнес (а также вовлеченных лиц), то лучше разбираетесь в ситуации. Вы можете спросить в ответ, по какой линейке продуктов нужны рекомендации, или что-то вроде: «Хотели бы вы видеть большее участие определенного сектора нашей аудитории?»

Исчезнет ли Data Science?

В основе вопроса о том, что будет с Data Science через пару десятилетий, лежат две основные проблемы: автоматизация и перенасыщение рынка труда.

Некоторые этапы процесса обработки данных действительно можно автоматизировать. Автоматическое машинное обучение (AutoML) может сравнивать производительность различных моделей и выполнять определенные части подготовки данных (например, масштабирование переменных). Но эти задачи – лишь малая часть большого процесса. Например, данные часто нужно создавать самостоятельно, поскольку идеально чистыми они бывают очень редко. При этом нужно взаимодействовать с другими людьми, например с UX-специалистами или с инженерами, которые будут проводить опрос или регистрировать действия пользователей.

Что касается пузыря на рынке труда, то хорошим сравнением может послужить разработка программного обеспечения в 1980-х годах. По мере того как компьютеры становились дешевле, быстрее и популярнее, возникали опасения, что вскоре эти машины смогут выполнять все и программисты перестанут быть востребованными. Но все произошло ровно наоборот, и теперь в США работает более 1,2 миллиона разработчиков ПО (http://mng.bz/MOPo). Несмотря на исчезновение таких профессий, как веб-мастер, над разработкой, обслуживанием и улучшением веб-сайтов работает больше людей, чем когда-либо.

Мы полагаем, что в Data Science появится больше специализаций, что может привести к исчезновению самого понятия «дата-сайентист». Но многие компании все еще находятся на ранних стадиях изучения того, как использовать науку о данных, и им предстоит еще много работы в этом направлении.

Другая часть понимания бизнеса – это развитие общих бизнес-навыков вроде умения адаптировать презентации и отчеты для разных аудиторий. Иногда вы будете обсуждать лучшую методологию с кандидатами наук по статистике, а иногда вы будете выступать перед вице-президентом, который не занимался математикой уже 20 лет. Вам нужно донести информацию до слушателей, учитывая их особенности.

Наконец, по мере карьерного роста вы научитесь определять, в каких случаях Data Science может помочь бизнесу. Если вы хотели создать систему прогнозирования, а руководство не поддержало эту идею, можно самому стать частью руководства и решить этот вопрос. Старший дата-сайентист будет искать способы внедрения машинного обучения, так как знает его возможности и ограничения, а также то, какие виды задач выиграют от автоматизации.

1.2. Различные типы вакансий в Data Science

Комбинировать три основных навыка, необходимых в Data Science (и описанных в разделе 1.1), можно на разных по сути должностях. С нашей точки зрения, эти навыки объединяются тремя основными параметрами: аналитикой, машинным обучением и наукой о принятии решений. Каждая из этих областей служит разным целям компании и дает принципиально разные результаты.

При поиске вакансий в сфере Data Science следует меньше обращать внимание на названия должностей – лучше сконцентрируйтесь на описании обязанностей и на вопросах во время собеседования. Посмотрите на опыт работы людей, занимающихся наукой о данных, например какие должности они раньше занимали и на кого учились. Вы можете обнаружить, что должности людей, которые выполняют схожие функции, называются совершенно по-разному, и наоборот, под одним и тем же названием должности «дата-сайентист» может подразумеваться совершенно разная работа. В этой книге мы поговорим о различных типах вакансий, но помните, что названия в разных компаниях могут отличаться.

1.2.1. Аналитики

Аналитик берет данные и передает их нужным людям. После того как компания установит цели на год, их можно поместить на информационную панель, чтобы руководство могло отслеживать прогресс каждую неделю. Можно также встроить функции, которые позволят менеджерам легко разбивать значения по странам или типам продуктов. Эта работа включает в себя много очистки и подготовки данных и, как правило, меньше работы по их интерпретации. Специалист должен уметь находить и устранять проблемы с качеством данных, однако основные решение по ним принимает бизнес-партнер. Таким образом, задача аналитика – взять данные внутри компании, отформатировать, упорядочить и передать их другим специалистам.

Поскольку должность аналитика не связана со статистикой и машинным обучением, некоторые люди и компании считают, что она выходит за рамки Data Science. Однако для большей части работы вроде создания осмысленных визуализаций и принятия решений о конкретных преобразованиях требуются те же навыки, которые нужны и другим специалистам DS. Например, аналитика могут попросить создать автоматизированную информационную панель, которая показывает изменение количества подписчиков и позволяет фильтровать данные только по подписчикам определенных продуктов или в определенных географических регионах. Он должен будет найти соответствующие данные в компании, выяснить, как их преобразовать (например, изменив их с ежедневных на еженедельные новые подписки), а затем создать содержательный набор информационных панелей с удобным интерфейсом и ежедневным автоматическим обновлением без ошибок.

Короткое правило: аналитик создает *информационные панели и отчеты на основе данных*.

1.2.2. Машинное обучение

Инженер по машинному обучению разрабатывает модели МО и разворачивает их в производство для постоянной работы. Такой специалист может оптимизировать алгоритм ранжирования для результатов поиска на сайте интернет-торговли, создать систему рекомендаций или отслеживать модель в производстве, чтобы убедиться, что ее производительность не снизилась с момента запуска. Инженер по машинному обучению уделяет меньше времени таким вещам, как создание визуализаций для убеждения других людей в чем-то, и больше сосредоточен на программировании для анализа данных.

Существенное различие между этой ролью и другими заключается в том, что результаты работы в первую очередь предназначены для машин. Например, вы можете создавать модели МО, которые превращаются в интерфейсы прикладного программирования (API) для других устройств. Во многих отношениях вы будете ближе к разработчику программного обеспечения, чем к другим специалистам Data Science. Любому дата-сайентисту полезно следовать передовым методам программирования, а вы как инженер по машинному обучению просто обязаны это делать. Ваш код должен быть производительным, протестированным и написанным так, чтобы другие люди могли с ним работать. Поэтому многие инженеры по машинному обучению имеют опыт работы в области информатики.

Инженера по машинному обучению могут попросить создать модель МО, которая может в реальном времени прогнозировать вероятность оформления онлайн-заказа. Он должен будет найти архивные данные в компании, обучить на них модель МО, преобразовать ее в API, а затем развернуть API, чтобы веб-сайт мог запускать модель. Если по какой-либо причине эта модель перестанет работать, для решения проблемы пригласят инженера по машинному обучению.

Короткое правило: инженер по машинному обучению создает *модели, которые работают непрерывно*.

1.2.3. Теория принятия решений

Специалист по принятию решений превращает необработанные данные компании в информацию, которая помогает руководству определяться с дальнейшими действиями. Для этой работы нужно хорошо владеть различными математическими и статистическими методами и процессами принятия бизнес-решений. Кроме того, специалисты по принятию решений должны уметь создавать убедительные визуализации и таблицы, чтобы люди, не имеющие технических знаний, понимали их анализ. Хотя они много программируют, обычно их код одноразовый – он нужен только для конкретного анализа. Поэтому неэффективный или сложный в поддержке код просто сходит им с рук.

Специалист по принятию решений должен понимать потребности других людей в компании и находить способы выдавать нужную информацию. Например, директор по маркетингу может попросить его помочь определить, какие типы продуктов следует выделить в праздничном каталоге компании. Специалист по принятию решений может исследовать, какие продукты хорошо продавались и без каталога, договориться с командой по user research о проведении опроса и использовать принципы поведенческой психологии, чтобы провести анализ и предложить подходящие варианты. Результатом, скорее всего, будет презентация или отчет PowerPoint, который будет представлен продакт-менеджерам, вице-президентам и другим бизнесменам.

Специалист по принятию решений часто использует знания в области статистики, чтобы помочь компании делать выбор в условиях неопределенности. Например, он может отвечать за управление системой экспериментальной аналитики в компании. Многие компании проводят онлайн-эксперименты или A/B-тестирование, чтобы оценить эффективность изменений. Это изменение может быть простым, например добавление новой кнопки, или сложным, включающим изменение системы ранжирования результатов поиска или полное изменение дизайна страницы. Во время A/B-тестирования посетителям случайным образом предлагается одно из двух или нескольких условий, например *контрольная* группа использует старую версию домашней страницы, а *экспериментальная* – новую версию. По окончании эксперимента действия посетителей из двух групп сравнивают между собой.

Из-за случайности показатели в контрольной и экспериментальной группах редко совпадают. Предположим, вы подбрасываете две монеты и одна выпадает орлом 52 раза из 100, а другая – 49 раз из 100. Можете ли вы сделать вывод, что первая монета имеет склонность выпадать орлом? Конечно, нет! Но бизнес-партнер может посмотреть на эксперимент, увидеть, что коэффициент конверсии составляет 5,4 % в контрольной группе и 5,6 % в экспериментальной, и объявить последнюю успешной. Специалист по принятию решений помогает интерпретировать данные, применять передовые методы разработки экспериментов и так далее.

Короткое правило: специалист по принятию решений создает анализ, на основе которого дает *рекомендации*.

1.2.4. Смежные специальности

Хотя три специализации, о которых мы писали в предыдущих разделах, – это основа работы в Data Science, также бывает несколько других отдельных должностей, которые выходят за рамки этих категорий. Мы перечислим их здесь, потому что разбираться в существующих направлениях полезно и, возможно, вам предстоит сотрудничество с такими специалистами. Тем не менее если вы бы хотели заниматься чем-то из нижеописанного, эта книга может быть для вас менее актуальной.

Бизнес-аналитик

Бизнес-аналитик занимается чем-то похожим на работу аналитика, но, как правило, использует меньше статистических знаний и навыков программирования. Его инструментом, вероятнее всего, будет Excel, а не Python, и он может вообще не создавать статистические модели. Хотя его функция аналогична функции аналитика, он выдает менее сложные результаты, поскольку используемые им программные средства и методы ограничены.

Если вы хотите заниматься машинным обучением, программированием или применением статистических методов, должность бизнес-аналитика может вас разочаровать, потому что не даст вам этих навыков. Кроме того, эта работа обычно оплачивается хуже, чем должности в Data Science, и считается менее престижной. Но она может стать хорошим стартом на пути к DS, особенно если у вас нет опыта работы с данными в бизнес-среде. Если вы хотите начать с роли бизнес-аналитика и вырасти до дата-сайентиста, ищите вакансии, где говорится о возможности получить необходимые для вас навыки, например в программировании на R или Python.

Инженер данных

Инженер данных занимается хранением данных в БД и обеспечением доступа к ним. Он не составляет отчеты, не проводит анализ и не разрабатывает модели; вместо этого он аккуратно хранит и форматирует данные в хорошо структурированных базах для других специалистов. Инженеру данных могут поручить хранение записей о клиентах в крупномасштабной облачной базе и добавление в нее новых таблиц по запросу.

Инженеры данных существенно отличаются от дата-сайентистов – они даже более редкие и востребованные специалисты. Такой сотрудник может помочь создать серверные компоненты данных внутренней экспериментальной системы компании и обновить поток обработки данных, когда задачи начинают занимать слишком много времени. Другие специалисты разрабатывают и отслеживают пакетные среды и потоковую передачу, управляя данными на всех этапах от сбора до обработки и хранения.

Если вас интересует инженерия данных, вам потребуются глубокие знания в области информатики; многие инженеры данных – это бывшие инженеры-программисты.

Вики Бойкис (Vicki Boykis): дано ли каждому стать дата-сайентистом?

Учитывая весь оптимизм (и большие потенциальные зарплаты, о которых пишут в новостях) в отношении Data Science, легко понять, почему эта сфера дает привлекательные возможности для карьерного роста, особенно если учесть, что диапазон и количество должностей в DS продолжают расти. Однако начинающему специалисту важно иметь реалистичное и детальное представление о том, как будет развиваться рынок Data Science в ближайшую пару лет, и в соответствии с этим корректировать свои решения.

Сегодня на сферу науки о данных влияет несколько основных тенденций. Во-первых, Data Science как область знаний существует уже десять лет и за это время прошла через ранние стадии цикла хайпа: ажиотаж в СМИ, быстрое внедрение и консолидация. Вокруг DS было много шума, ее обсуждали в медиапространстве, внедряли компании Кремниевой долины и не только, и сейчас мы находимся на этапе быстрого развития области в крупных компаниях и стандартизации таких программных средств обработки данных, как Spark и AutoML.

Во-вторых, в результате быстрого развития отрасли возник избыток новых специалистов, пришедших после изучения новых программ в

университетах, буткемпах или на онлайн-курсах. Число кандидатов на любую должность в области Data Science, особенно на начальном уровне, выросло с 20 человек на место до 100 или более. Теперь нередко можно увидеть даже 500 резюме на одну вакансию.

В-третьих, стандартизация наборов программных средств, обеспеченность рабочей силой и спрос на специалистов с опытом работы привели к изменениям в порядке распределения рабочих мест и к созданию иерархии должностей и функциональных обязанностей в Data Science. Например, в одной компании дата-сайентист может заниматься созданием моделей, а в другой – главным образом выполнением анализа SQL, что соответствует, скорее, должности аналитика.

Для тех, кто хочет прийти в Data Science с нуля, это означает несколько вещей. Во-первых, и это самое важное, они увидят, что рынок труда наполнен конкурентами. Особенно это касается тех, кто, в принципе, только начинает работать (например, выпускников колледжей), либо тех, кто пришел в отрасль из какой-либо другой сферы и конкурирует за место с тысячами таких же соискателей. Во-вторых, они могут претендовать на вакансии, которые не совсем соответствуют тому образу Data Science, который создается в СМИ, будто это исключительно написание и внедрение алгоритмов.

Учитывая эти тенденции, важно понимать, что изначально может быть непросто выделиться среди других кандидатов и попасть на финальный этап собеседования. И хотя стратегии, приведенные в этой книге, могут показаться сложными, они помогут вам привлечь внимание, а это необходимо в сложившихся условиях высокой конкуренции.

Инженер-исследователь

Ученый-исследователь разрабатывает и внедряет новые программные средства, алгоритмы и методологии, которые часто используются другими дата-сайентистами в компании. Такие должности почти всегда требуют наличия кандидатской степени, обычно в области информатики, статистики, количественных социальных наук или в смежных направлениях. Ученому-исследователю может потребоваться несколько недель, чтобы изучить и испытать методы повышения эффективности онлайн-экспериментов, повысить точность распознавания изображений в беспилотных автомобилях на 1 % или создать новый алгоритм глубокого обучения. Он даже может тратить время на написание исследовательских работ, которые будут редко использоваться в компании, но помогут поднять ее престиж и (в идеале) продвинуться в этой области. Поскольку эти должности требуют очень специфического опыта, мы не будем уделять им особого внимания в этой книге.

1.3. Выбор пути

В главе 3 мы рассмотрим несколько способов обучиться работе с данными, опишем преимущества и недостатки каждого из них, а также дадим несколько советов по выбору пути, подходящего именно вам. На этом этапе было бы неплохо задуматься, в каком направлении Data Science вы хотите специализироваться. Какой опыт у вас уже есть? Мы видели дата-сайентистов, которые в прошлом были инженерами, профессорами психологии, менеджерами по маркетингу, студентами программ статистики и социальными работниками. Часто знания, полученные в других профессиях и академических областях, могут помочь вам лучше справляться с работой в DS. Если вы уже работаете с данными, подумайте, в какой части треугольника вы находитесь. Довольны ли вы текущим положением? Хотите ли переключиться на другой тип работы в Data Science? Смена специализации зачастую вполне доступна.

1.4. Интервью с Робертом Чангом, дата-сайентистом из Airbnb

Роберт Чанг (Robert Chang) – дата-сайентист в Airbnb, который работает над продуктом Airbnb Plus. Ранее он занимался аналитикой продуктов, создавал конвейеры данных и модели, проводил эксперименты в «Команде роста» (Growth team) Twitter. Роберт ведет блог об инженерии данных, дает советы новичкам, а также рассказывает о работе в Airbnb и Twitter на странице https://medium.com/@rchang.

Расскажите о вашем первом опыте в Data Science.

Моей первой работой был анализ данных в The Washington Post. Еще в 2012 году я был готов оставить учебу и уйти в эту сферу, но не знал, чем именно хочу заниматься. Я надеялся стать специалистом по визуализации данных, так как был впечатлен работой в The New York Times. Когда я пошел на ярмарку вакансий в вузе и увидел, что в The Washington Post требуются сотрудники, я наивно предположил, что они, скорее всего, делают то же самое, что и The New York Times. Я подал заявку и получил работу, не особо вдаваясь в детали.

Если вам нужен пример того, как не следует начинать карьеру в Data Science, возьмите мой случай! Я получил работу в надежде заниматься либо визуализацией данных, либо моделированием, но очень быстро понял, что, скорее, выполняю обязанности инженера данных. Бˆольшая часть моих задач заключалась в создании конвейеров ETL (извлечение, преобразование, загрузка), повторном запуске скриптов SQL и попытках обеспечить запуск отчетов, чтобы можно было представлять ключевые показатели руководству. Тогда я пережил это очень болезненно; я понял, что то, чем мне хотелось заниматься, не соответствовало тому, что было нужно компании, и в конце концов уволился.

Но в последующие годы работы в Twitter и Airbnb я понял, что столкнулся с нормой, а не исключением. При работе с данными их нужно наращивать слой за слоем. Моника Рогати (Monica Rogati) опубликовала знаменитую статью об иерархии потребностей Data Science, попав в самую точку (http://mng.bz/ad0o). Но в то время мне не хватало опыта, чтобы оценить, как в действительности устроена работа в этой сфере.

На что следует обращать внимание при поиске работы в Data Science?

При поиске вакансий вам следует обращать внимание на состоянии инфраструктуры данных в компании. Если вы устроитесь в организацию, где куча сырых данных даже не размещена в хранилище, то уйдут месяцы или даже годы, прежде чем вы займетесь чем-то интересным вроде аналитики, экспериментов или машинного обучения. Если вы на такое не рассчитываете, то этап развития компании совершенно не будет соответствовать тому вкладу, который вы хотите внести в организацию.

Чтобы оценить ситуацию, можно задать вопросы вроде: «Есть ли у вас команда по созданию инфраструктуры данных?», «Как давно она создана?», «На что похож стек данных?», «Есть ли у вас команда дата-инженеров?», «Как они взаимодействуют с дата-сайентистами?», «Есть ли у вас процесс инструментального анализа логов, построения таблиц данных и помещения их в хранилище при создании нового продукта?» Если всего этого нет, вы станете частью команды, создающей все с нуля; приготовьтесь потратить на это немало времени.

Второе, на что нужно обращать внимание, – это люди. Особенно присмотритесь к трем типам сотрудников. Полагаю, вы не хотите быть первым дата-сайентистом в компании. Тогда вам следует искать команду с опытным руководителем. Он знает, как создать и поддерживать хорошую инфраструктуру и процессы, чтобы работа специалистов была эффективной. Также ищите менеджера, который поддерживает постоянное обучение. Наконец, очень важно, особенно для новичков, работать с техническим руководителем проекта или старшим специали-

стом по данным, у которого много практического опыта. Именно этот человек помогает вам лучше всего справиться с ежедневными задачами.

Какие навыки нужны data-сайентисту?

Я думаю, это зависит от того, на какую должность вы претендуете и чего от вас ожидает работодатель. Престижные компании, как правило, задают высокую планку – иногда необоснованно высокую, ведь к ним выстраивается очередь из желающих. Обычно они ищут «единорогов» – тех, кто работает с R или Python, а также отлично разбирается в инженерии данных, проектировании экспериментов, создании конвейеров ETL и моделей с последующим внедрением в производство. Очень уж много требований к кандидатам! Хотя со временем вы можете освоить все эти полезные навыки, не думаю, что они так уж нужны для начала работы в отрасли.

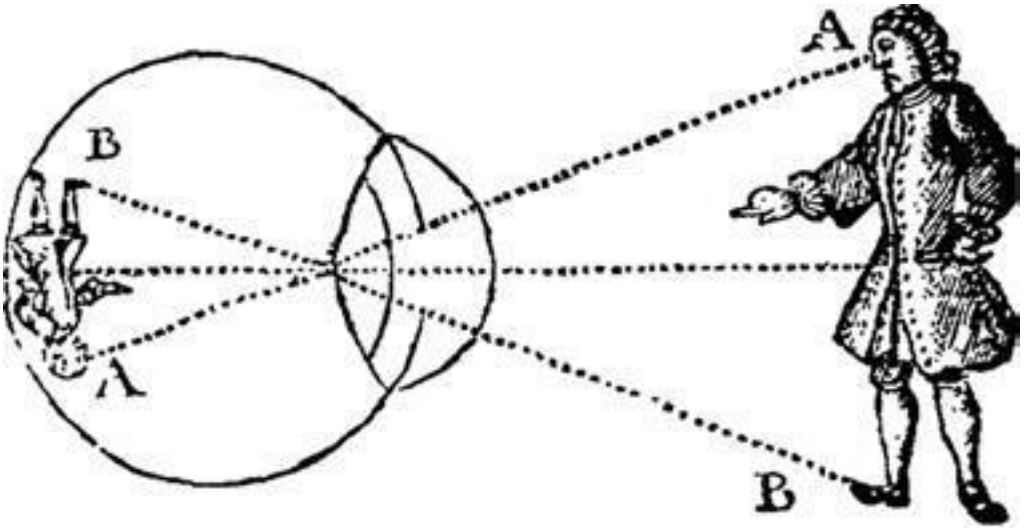
Если вы знаете R или Python и немножко SQL, это уже довольно неплохо для старта. Здорово, если вы можете выучить что-то наперед в целях карьеры, но мне кажется, что это необязательно. Гораздо важнее в принципе любить учиться. У ведущих технологических компаний могут быть более высокие требования, но они нужны скорее не для работы, а для того, чтобы выделить вас среди остальных. Следует различать основные навыки, необходимые для начала карьеры в Data Science, и те, которые неплохо бы иметь сотрудникам топовых компаний.

Итоги

• Набор навыков в Data Science зависит от людей и должностей. Хотя некоторые знания являются фундаментальными, специалисты по работе с данными не обязательно должны быть экспертами во всех смежных областях.

• У работы в Data Science разные направления: предоставление правильных, очищенных данных стейкхолдерам (аналитика), развертывание моделей МО в производство (машинное обучение) и использование данных для принятия решений (теория принятия решений).

2. Типы компаний в Data Science



В этой главе

- Типы компаний, нанимающие дата-сайентистов.
- Плюсы и минусы каждого типа компании.
- Комплекты технологий, которые можно увидеть на разных должностях.

Как уже было сказано в главе 1, в Data Science есть много разных специализаций: инженер-исследователь, инженер по машинному обучению, бизнес-аналитик и другие. Ваши рабочие обязанности будут зависеть от должности, а также от компании, в которую вы устроились. Ее размер, возраст, отрасль – все это влияет на типы проектов, сопутствующие технологии и командную культуру. Умение разбираться в архетипах компаний лучше подготовит вас к поиску работы, будь то ваша первая или очередная должность в Data Science.

Цель этой главы – сформировать у вас представление о повседневной работе некоторых стандартных видов компаний. Мы расскажем о пяти вымышленных фирмах, которым нужны дата-сайентисты. Все эти образы основаны на исследованиях и на нашем собственном опыте. Кроме того, они иллюстрируют основные принципы, которые можно широко применять при поиске работы. Хотя абсолютно одинаковых компаний не существует, знания об этих пяти архетипах поможет лучше понять потенциального работодателя.

Описанные нами стереотипы – не истина в последней инстанции, хоть они и основаны на тенденциях, которые мы наблюдаем в этих отраслях. Есть компании, которые вообще не соответствуют этим стереотипам, а еще бывает так, что отдельные команды отличаются по своей культуре и организации от остальной фирмы.

Хотя организации в этой главе вымышленные, все остальное написано настоящими дата-сайентистами, работающими в реальных компаниях!

2.1. КИТк: крупная информационно-технологическая компания

• Похожа на: Google, Facebook и Microsoft.
• Возраст компании: 20 лет.
• Количество сотрудников: 80 000.

КИТк – влиятельная технологическая компания, продающая облачные сервисы и специализированное ПО для повышения производительности – текстовые редакторы, серверное оборудование и бесчисленное количество разовых бизнес-решений. Свое огромное состояние компания использует для финансирования необычных проектов в области исследований и разработок (НИОКР), таких как беспилотные скутеры и технологии виртуальной реальности (VR). Об их исследованиях говорят в новостях, а большинство технических сотрудников – это инженеры, которые постепенно совершенствуют уже имеющиеся продукты, добавляют дополнительные функции, улучшают пользовательский интерфейс и запускают новые версии.

2.1.1. Команда: одна из многих в КИТк

В КИТк около тысячи дата-сайентистов. Они собраны в команды, каждая из которых поддерживает свой продукт или подразделение. Кроме того, специалиста могут направить в отдел другого профиля для всесторонней поддержки. Например, у команд проектировщиков VR-шлемов, маркетологов, специалистов по продвижению VR-шлемов и менеджеров цепочек поставок есть свой дата-сайентист.

Если бы вы стали членом одной из этих команд по анализу данных, то быстро бы адаптировались. В крупных организациях новых сотрудников нанимают ежедневно, поэтому в компании должны быть стандартные процессы выдачи ноутбука и обеспечения доступа к данным. Также сотрудников обучают работать со специализированным ПО. В команде вам поручат заниматься анализом данных для конкретной области. Это может быть создание отчетов и диаграмм, которые помогут менеджерам обосновать бюджеты проектов. Вам также могут поручить построение моделей МО – они передаются разработчикам для запуска ПО в продакшен.

Скорее всего, в вашей большой команде будет полно опытных специалистов. Поскольку КИТк – компания крупная и успешная, она может привлекать множество профессионалов. Вы будете работать в большой команде, члены которой нередко работают над практически несвязанными задачами, например один сотрудник может выполнять исследовательский анализ на R для директора, а другой – строить модель МО на Python для соседнего отдела. Размер команды – это и благословение, и проклятие в одном флаконе: вы можете обсудить свои идеи со многими экспертами, но большинство из них, скорее всего, не знакомы с вашими конкретными задачами. Кроме того, в команде есть устоявшаяся иерархия. К специалистам на более высо-

ких должностях, как правило, прислушиваются чаще, потому что они опытнее и в своей профессиональной сфере, и в работе с различными отделами КИТк.

Работа вашей команды – это здоровый баланс между поддержанием деятельности компании (например, составление ежемесячных отчетов и ежеквартальное обновление модели МО) и реализацией новых проектов (например, создание новых прогнозов). Руководитель команды должен искать золотую середину между потоком запросов от других команд, которым результаты нужны в ближайшее время, и желанием взяться за что-то инновационное – не востребованное сейчас, но полезное в долгосрочной перспективе. Крупные финансовые возможности КИТк позволяют компании заниматься инновациями и НИОКР гораздо больше, чем другим организациям. Благодаря этому, в свою очередь, команды охотно работают над новыми интересными проектами в Data Science.

2.1.2. Технология: продвинутая, но неупорядоченная

КИТк – крупная организация. При таких масштабах не избежать использования различных типов технологий между подразделениями. Один отдел может хранить данные о заказах и клиентах в базе Microsoft SQL Server, другой – записывать все в Apache Hive. Мало того, неупорядоченными являются не только технологии хранения данных, но и сами данные. Неупорядоченные технологии хранения – еще полбеды, ведь сами данные тоже ведутся по разным принципам. Одно подразделение индексирует записи о клиентах по номеру телефона, другое – по адресу электронной почты.

У большинства организаций такого же масштаба есть собственный арсенал технологий. Поэтому вам как сотруднику КИТк придется освоить способы работы с данными, характерные именно для этой компании. Изучение специализированного софта здорово поможет на текущей должности, но не в других фирмах.

Вам как специалисту по данным наверняка понадобится несколько видов инструментов. Поскольку КИТк – компания весьма крупная, она хорошо поддерживает распространенные языки, такие как R и Python. Некоторые команды порой работают с платными языками вроде SAS или SPSS, но это бывает реже. Если вы хотите использовать необычный язык, который нравится вам, но мало кем используется (скажем, Haskell), нужно будет получить согласие руководителя.

Комплекс технологий МО сильно различается в зависимости от отдела. Некоторые группы используют микросервисы и контейнеры для эффективного развертывания моделей, тогда как другие работают с устаревшими производственными системами. Разнообразие стека для развертывания ПО затрудняет подключение к API других команд; единой базы знаний или хотя бы понимания того, что происходит, попросту нет.

2.1.3. Плюсы и минусы КИТк

Быть дата-сайентистом в КИТк означает иметь потрясающую работу в потрясающей компании. А поскольку эта компания технологическая, сотрудники знают, кто такой специалист по данным и что полезного он может сделать. Когда все понимают вашу роль одинаково, это значительно облегчает работу. Если в компании много дата-сайентистов, значит, у вас будет широкий круг поддержки, а также возможность плавно влиться в команду и получить доступ к необходимым ресурсам. Оказаться в затруднении один на один – редкость.

В то же время у наличия толпы специалистов по работе с данными есть свои недостатки. Стек технологий сложен, в нем непросто ориентироваться, потому что создавался он разными людьми и разными способами. Может так случиться, что анализ, который вас попросили воссоздать, написал человек, который уже уволился, да еще и на незнакомом вам языке. Вам будет

сложнее выделиться среди множества других специалистов. Кроме того, может быть непросто найти интересный проект, потому что над многими из них уже работают другие люди.

Как устоявшаяся компания КИТк дает больше гарантий занятости. Риск увольнений есть всегда, но работа здесь не похожа на работу в стартапе, где финансирование может прекратиться в любой момент. Кроме того, в крупных компаниях руководители больше склонны искать новых сотрудников, чем увольнять старых, потому что увольнение сложно юридически.

У сотрудников КИТк много специализаций – это одновременно и хорошо, и плохо. Дата-инженеры, архитекторы данных, дата-сайентисты, маркетологи и другие выполняют разные задачи, связанные с Data Science, а значит, вокруг вас будет много людей, которым можно передать работу. Например, создавать собственную базу данных вас вряд ли заставят. С одной стороны, хорошо иметь возможность делегировать задачи, для которых у вас нет опыта, а с другой – так вы не получите новые навыки.

Еще один минус КИТк – бюрократия. В крупной компании введение новых технологий, поездки на конференции и запуск проектов придется согласовывать с начальством. Хуже того, от проекта, над которым вы работали годами, могут отказаться из-за конфликта между двумя руководителями, а ваш проект может «пострадать от шальной пули». Или, что еще хуже, ваш проект может пасть случайной жертвой конфликта двух руководителей – его могут просто закрыть.

КИТк – отличная компания для дата-сайентистов, которые хотят решать сложные задачи с помощью передовых методов. Это касается и специалистов по принятию решений, планирующих заниматься анализом, и инженеров МО, мечтающих создавать и развертывать модели. У крупных компаний есть масса задач и денег, чтобы пробовать новые вещи. Возможно, вы не сможете самостоятельно принимать важные решения, но будете знать, что внесли в них свой вклад.

Работа в КИТк не подойдет специалистам, которые хотят самостоятельно руководить и принимать решения. В большой компании есть установленные методы, протоколы и модели, которым придется следовать.

2.2. HandbagLOVE: устоявшийся ритейлер

- Похожа на: Payless, Bed Bath & Beyond и Best Buy[1].
- Возраст компании: 45 лет.
- Количество сотрудников: 15 000 (10 000 в розничных магазинах, 5000 в офисах).

HandbagLOVE – это розничная сеть с 250 точками по всей территории США, которая занимается продажей кошельков и клатчей. Здесь трудятся оформители магазинов и специалисты по повышению качества обслуживания клиентов. Компания на рынке уже давно, но новые технологии осваивать не спешит: прошло довольно много времени, прежде чем у нее появились первый веб-сайт и приложение.

[1] Американские сети магазинов одежды и товаров для дома с низкими ценами. – *Примеч. ред.*

В последнее время продажи HandbagLOVE упали, поскольку Amazon и другие интернет-магазины потеснили компанию на рынке. Руководство осознало очевидное и решило улучшить ситуацию с помощью технологий, инвестируя в онлайн-приложение и Amazon Alexa, а также пытаясь использовать накопленные данные. Финансовые аналитики HandbagLOVE уже много лет прекрасно рассчитывают совокупную статистику по заказам и клиентам, но лишь недавно компания подумала о том, чтобы нанять дата-сайентистов для лучшего понимания клиентов.

Новая группа специалистов по анализу данных была создана на базе службы финансовых аналитиков, которые ранее составляли отчеты по показателям эффективности компании в Excel. После дополнительного привлечения дата-сайентистов команда начала создавать более сложные продукты: ежемесячные статистические прогнозы роста клиентов в R, интерактивные информационные панели для лучшего понимания продаж, а также сегментацию, объединяющую клиентов в удобные группы для целей маркетинга.

Даже после создания моделей МО для новых отчетов и анализа HandbagLOVE далека от внедрения их в непрерывный рабочий процесс. Все рекомендации по продуктам на ее веб-сайте и в приложении основаны на продуктах МО от сторонних производителей. В команде по анализу данных надеются изменить ситуацию, но никому не известно, когда это все же произойдет.

2.2.1. Команда: небольшая группа, стремящаяся к росту

Команда полагается на специалистов по созданию отчетов, а не по машинному обучению, потому что оно для них в новинку. Никто не владел современными методами статистики и МО, так что сотрудникам приходилось вникать во все самостоятельно. Прекрасно, когда люди могут в одиночку изучать новые интересующие их техники. Обратная сторона медали – неэффективные или даже неправильные методы: в компании нет экспертов, которые могли бы проверить работу.

HandbagLOVE наметила общие пути продвижения специалистов по работе с данными на руководящие должности. К сожалению, они не подходят для сферы Data Science: это глобальные цели, скопированные из других областей вроде разработки ПО, потому что никто на самом деле не понимает, какие показатели использовать. Планируя повышение, вы должны убедить своего руководителя, что готовы перейти на следующий уровень, и, если повезет, он сможет получить одобрение для вашей кандидатуры. С другой стороны, если команда будет расти, вы быстро станете в ней старшим.

Сотрудников группы Data Science знают хорошо, потому что они делают отчеты и модели для других отделов компании (маркетинг, цепочка поставок, обслуживание клиентов). Команда пользуется уважением в компании и дружит с другими подразделениями. У дата-сайентистов HandbagLOVE гораздо больше полномочий, чем в других компаниях, из-за размера команды и ее влияния внутри организации. Их встречи с руководителями высшего звена на важных переговорах – обычное дело.

2.2.2. Технология: устаревшие методы, которые начинают меняться

В разговорах о технологиях в HandbagLOVE вы часто слышите фразу: «Ну, мы всегда так делали». Данные о заказах и клиентах хранятся в базе данных Oracle, которая напрямую связана с кассовым аппаратом и за 20 лет ни разу не менялась. Система вышла за пределы своих возможностей и претерпела множество изменений. Тем не менее она все еще работает. Другие данные также собираются и хранятся в центральной базе: информация с веб-сайта, центра обслуживания клиентов, рекламных акций и маркетинговых рассылок. Все эти серверы, которые обслуживает ИТ-команда, располагаются локально (*on-prem*), а не в облаке.

Когда все данные хранятся на одном большом сервере, можно свободно подключаться и объединять их как угодно. И хотя иногда запрос занимает вечность или перегружает систему, обходными путями обычно получается найти рабочий способ. Большинство аналитических операций выполняется на ноутбуке. Более мощный компьютер для обучения моделей получить непросто. У компании нет стека технологий для машинного обучения, потому что нет МО как такового.

2.2.3. Плюсы и минусы HandbagLOVE

Как сотрудник HandbagLOVE вы очень влиятельны и можете делать все, что считаете нужным. Можно предложить создать модель пожизненной ценности клиента, построить ее и использовать в компании и при этом не просить разрешения у кучи людей. Такую свободу дает сочетание размера компании и новизны сферы Data Science. И она того стоит: перед вами открываются невероятные возможности для принятия лучших, на ваш взгляд, решений. С другой стороны, вокруг не так много людей, к кому можно обратиться за помощью. Вы сами несете ответственность за то, чтобы все работало, а также за последствия в случае неудачи.

Стек технологий устарел, и вам придется потратить много времени на поиск обходных решений, что, безусловно, не очень практично. Возможно, вы захотите использовать более новый способ хранения данных или запуска моделей, но не получите технической поддержки. Если вы не можете создать какую-либо новую технологию самостоятельно, вам придется обходиться без нее.

Заработная плата будет ниже, чем в более крупных компаниях, особенно в технологических. У HandbagLOVE просто нет денег, чтобы платить за анализ данных. Кроме того, компания в любом случае не ищет лучших из лучших – ей просто нужны люди, которые умеют делать базовые вещи. При этом зарплата не будет совсем уж низкой: безусловно, она будет намного выше, чем у большинства сотрудников с тем же сроком работы.

HandbagLOVE подходит для дата-сайентистов, которым нравится принимать собственные решения, но при этом не нужны передовые технологии. Если вы не против использовать стандартные статистические методы и составлять рутинные отчеты, HandbagLOVE станет хорошим местом для развития карьеры. Если же вы хотите связаться с новейшими технологиями МО, то таких проектов будет крайне мало; кроме того, в компании практически не будет людей, которые поймут хоть что-то из того, о чем вы говорите.

2.3. Seg-Metra: стартап на ранней стадии

• Похожа на: тысячи неудачных стартапов, о которых вы даже не слышали.
• Возраст компании: 3 года.
• Количество сотрудников: 50.

Seg-Metra – молодая компания, чей продукт помогает клиентам оптимизировать вебсайты с помощью кастомизации уникальных сегментов плкупателей. В начале своей короткой истории Seg-Metra привлекла нескольких известных клиентов к использованию своих технологий и благодаря этому смогла получить больше финансирования от венчурных капиталистов. Теперь, имея миллионы долларов, компания хочет быстро увеличить размеры и улучшить продукт.

Самое крупное усовершенствование, которое основатели компании предлагали инвесторам, – добавление в продукт базовых методов машинного обучения, что было представлено как «передовой ИИ». Получив новое финансирование, основатели компании ищут инженеров МО для реализации задуманного. Им также нужны специалисты по принятию решений для составления отчетности об использовании продукта, чтобы лучше понять, как его оптимизировать.

2.3.1. Команда (какая еще команда?)

Новый дата-сайентист вполне может оказаться первым в компании. Или же стать одним из первых и подчиняться, скорее всего, тому, кого взяли раньше всех. Поскольку команда новая, протоколов практически не будет – никаких устоявшихся языков программирования, практик, способов хранения кода или официальных совещаний.

Именно тот дата-сайентист, которого взяли первым, будет отдавать все распоряжения. Скорее всего, культура команды будет зависеть от его личностных качеств. Если этот человек открыт для обсуждения и доверяет другим членам команды, то они смогут принимать решения вместе, например обсуждать, какой язык использовать. Если этот человек привык все контролировать и не готов прислушиваться к мнению других, он будет принимать такие решения самостоятельно.

В такой неструктурированной среде может вырасти очень сплоченный коллектив. Команда Data Science всеми силами пытается заставить работать новые технологии, методы и

программные средства, и в результате формируются глубокие связи и дружба. С другой стороны, те, у кого нет власти, могут испытывать огромное эмоциональное насилие со стороны руководства, а поскольку компания небольшая, никто не понесет за это ответственности. Независимо от того, как именно будет развиваться компания Seg-Metra, специалистов по работе с данными здесь ждет непростое время.

Работа команды может захватывать или раздражать – каждый день по-разному. Часто дата-сайентисты проводят анализ впервые, например делают первую попытку использовать данные о покупках для сегментации клиентов или развертывают первую нейронную сеть. Аналитические и инженерные задачи, которые решаются впервые, захватывают дух, ведь это неизведанная территория внутри компании, а специалисты по работе с данными становятся первопроходцами. Иногда работа может быть изнурительной, например когда уже пора предоставить инвестору готовую демоверсию, а модель все еще не сходится. Даже если у компании есть данные, сама инфраструктура может быть настолько запутана, что их просто невозможно использовать. Несмотря на хаотичность работы, выполнение всех этих задач в Seg-Metra помогает дата-сайентистам очень быстро освоить множество навыков.

2.3.2. Технология: передовые методы, собранные воедино

Поскольку Seg-Metra – молодая компания, ей не приходится поддерживать устаревшие технологии. Кроме того, хочется произвести впечатление на инвесторов, а сделать это гораздо проще, когда располагаешь эффектным стеком технологий. Поэтому Seg-Metra использует самые современные и лучшие методы разработки ПО, хранения и сбора данных, а также анализа и отчетности. Информация хранится в современных облачных сервисах: локально ничего не делается. Дата-сайентисты подключаются напрямую к этим базам и создают модели нейронных сетей МО на крупных экземплярах виртуальных машин Amazon Web Services (AWS) с обработкой графическим процессором. Эти модели развертываются с помощью современных методов программной инженерии.

Конец ознакомительного фрагмента.