Я посвящаю эту книгу
моей любящей жене Сэнди –
за ее терпение, поддержку и понимание

Содержание

От издательства

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте www.dmkpress.com, зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу dmkpress@gmail.com; при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу http://dmkpress.com/authors/publish_book/ или напишите в издательство по адресу dmkpress@gmail.com.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг, мы будем очень благодарны, если вы сообщите о ней главному редактору по адресу dmkpress@gmail.com. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательство «ДМК Пресс» очень серьезно относится к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу электронной почты dmkpress@gmail.com.

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

Предисловие

В последние годы направления в высших учебных заведениях, посвященные анализу данных, приобрели большую популярность. По общим оценкам, на данный момент запущено уже около 250 программ магистратуры в этой области, и их список пополняется ежегодно. Книга, которую вы держите в руках, базируется на моем более чем 40-летнем опыте преподавания аналитических дисциплин на факультете экономики и предпринимательства Келли при Индианском университете в Блумингтоне. Она основана на моих лекциях, демонстрациях в Excel, раздаточных и иных материалах, которые я использую в процессе преподавания. За время своей деятельности я смог понять, с чем у студентов возникают наибольшие трудности, и разработать необходимые стратегии, которые включил в эту книгу.

Существует немало книг, посвященных анализу данных, и все они очень разные. Отличительной чертой этой книги является использование в качестве инструмента аналитики данных инструмента под названием KNIME. KNIME – это программное обеспечение с открытым исходным кодом, в котором вы практически все можете делать мышкой, что до предела облегчает процесс понимания предметной области. В отличие от языков R и Python, использующихся в большинстве книг по аналитике, KNIME не требует серьезного и досконального изучения, а позволяет многие операции выполнять на интуитивном уровне, практически без программирования. *Рабочие процессы* (workflow), создаваемые в KNIME, предстают перед пользователем в наглядном виде, что позволяет легко определять, какие действия выполняются и в каком порядке.

Графический интерфейс KNIME напоминает дорогие коммерческие продукты, такие как SAS Enterprise Miner, IBM SPSS Modeler и Microsoft Azure Machine Learning. Несмотря на открытость исходного кода, KNIME располагает тысячами простых в использовании *узлов* (node), предназначенных для анализа данных, в том числе с применением техник глубокого обучения, интеллектуального анализа текста и ансамблевых методов машинного обучения.

Открытый исходный код позволяет использовать KNIME как в личных целях, так и в продуктовой среде. В отличие от многих схожих продуктов, KNIME не требует покупки дорогостоящих ежегодных лицензий. Раньше студенты часто жаловались мне на то, что по завершении обучающих курсов по аналитике они не могли воспользоваться полученными знаниями ни самостоятельно, ни в компаниях из-за дороговизны лицензий на программное обеспечение, применявшееся во время обучения.

Еще одним преимуществом KNIME является возможность использовать в рабочих процессах скрипты на языках R и Python. Лично я иногда прибегаю

к помощи скриптов на R, если не могу найти подходящего узла KNIME или когда мне требуется более богатый функционал. Но знание языка R совсем не обязательно для использования KNIME, и в этой книге при необходимости я буду давать исходные коды скриптов.

Кроме того, в книге я буду приводить как фундаментальные концепции используемых методов, так и их более простое описание. К примеру, анализ главных компонент обычно представляется в виде последовательности матричных вычислений с использованием разложения по собственным значениям, и математика не проясняет, как это работает. Помимо уравнений, я включил в книгу описание техники получения собственных значений в Excel путем максимизации взвешенных средних для набора переменных. Это один из примеров, показывающих, что книга написана языком, доступным для читателей, незнакомых с техническими подробностями некоторых алгоритмов и методов, но в то же время может оказаться полезной и людям, свободно владеющим языком статистики и математики.

Все демонстрации рабочих процессов KNIME я буду сопровождать соответствующими рисунками и подробным описанием используемых узлов. Кроме того, все приведенные в книге процессы, а также наборы данных для них доступны для загрузки. Таким образом, вы можете скачать их, запустить у себя на компьютере, внести нужные вам изменения и использовать в своей работе. Загрузить рабочие процессы можно на странице книги на сайте издательства «ДМК Пресс», а также по адресу https://tinyurl.com/KNIMEWorkflows.

Фрэнк Асито, Блумингтон, штат Индиана, США

Об авторе

Фрэнк Асито (Frank Acito) – заслуженный профессор Университета Индианы, доктор философии (Ph.D.) Государственного университета Буффало. Прошел постдокторантуру в области эконометрики и вариационного анализа. С 2011 года занимает пост главы факультета экономики и предпринимательства Келли при Индианском университете в Блумингтоне. Обладает более чем 40-летним стажем преподавания аналитических дисциплин на факультете экономики. Автор множества научных статей и обладатель многих научных премий. Женат, имеет троих детей.

О переводчике

Александр Гинько, обладающий богатым опытом работы в сфере ИТ и более десяти лет посвятивший переводам книг и статей на самые разные темы, в последние годы специализируется на переводе книг в области бизнес-аналитики и программирования для издательства «ДМК Пресс» по направлениям Python, SQL, R, Power BI, DAX, статистика, машинное обучение, нейронные сети, Excel, Power Query, Tableau, ... На данный момент в активе Александра уже порядка 25 книг, включая одну авторскую, и он продолжает плодотворно работать над переводом новых книг.

Возможно, вам также будут интересны книги *Введение в статистическое обучение с примерами на Python* (https://dmkpress.com/catalog/computer/statistics/978-5-93700-217-4) и *Машинное обучение сквозь призму Excel. Примеры и упражнения* (https://dmkpress.com/catalog/computer/data/978-5-93700-238-9) в переводе Александра.

Помимо перевода книг, Александр ведет свой канал в Telegram (https://t.me/alexanderginko_books), на котором вы можете из первых уст получить ответы на все интересующие вас вопросы об уже переведенных книгах, находящихся в работе и запланированных на будущее. Также на канале можно найти промокоды на все книги Александра для покупки книг на сайте издательства «ДМК Пресс» с большими скидками. Купить книги Александра и следить за переводом новых книг в режиме реального времени можно и с помощью его бота в Telegram по адресу https://t.me/alexanderginko_books_bot.

Глава 1

Введение в аналитику

Аналитика предполагает применение моделей, построенных на основе данных, с целью улучшения результатов и снижения затрат и рисков как в коммерческих, так и в некоммерческих организациях. Эрик Сигель (Eric Siegel)[1] ввел термин *эффект предсказания* (prediction effect) для описания того, как модели делают предсказания относительно людей, документов, действий или неких сущностей (Siegel, 2013). Он также предположил, что применяемые прогностические модели не должны быть исключительно точными в своих предсказаниях. Все, что нам нужно, – это чтобы предсказания с использованием анализа данных были точнее методов, которые мы использовали до этого.

Аналитика (analytics) стала неким объединяющим термином для несвязанных областей применения разнообразных приложений в бизнесе. Иногда под аналитикой понимают статистический и математический анализ данных, применяющийся в таких областях, как продажи, услуги, системы поставок и логистика, здравоохранение и защита данных. В этих областях аналитика обычно подразделяется на три основных типа на основе поставленной цели: описательная, предсказательная и предписывающая – в порядке увеличения сложности.

Описательная аналитика (descriptive analytics) направлена на исследование и интерпретацию данных с целью ответа на вопросы «Что произошло?» или «Что происходит в данный момент?». Обычно для этого используются столбчатые диаграммы, диаграммы рассеяния, линейные графики, диаграммы размаха и гистограммы. Кроме того, описательная аналитика активно использует сводную статистику, включающую в себя *меры центральной тенденции* (measures of central tendency) и информацию о корреляции в наборе. Это помогает лучше понять природу и распределение исходных данных.

Предсказательная аналитика (predictive analytics) концентрируется на прогнозировании непрерывных переменных и классификации категори-

---

[1] Доктор философии, известный лектор и просветитель, основатель конференции Predictive Analytics World. – *Прим. перев.*

альных откликов. Перед построением предсказательных моделей важно воспользоваться техниками описательной аналитики для лучшего понимания, очистки и подготовки данных. Предсказательная аналитика включает в себя такие методы, как множественная регрессия, деревья решений, нейронные сети и многие другие.

Предписывающая аналитика (prescriptive analytics) применяется для помощи в принятии решений и отвечает на вопрос «Что нужно сделать?». Для разработки предписывающих моделей используются имитационное моделирование, рекомендательные системы и оптимизационные методы. При этом для успеха предписывающей аналитики необходима высокая точность предварительно построенных предсказательных моделей.

Хотя в некоторых примерах мы будем использовать техники описательной и предписывающей аналитики, главным образом мы сосредоточимся в этой книге на методах предсказательной аналитики, точность которых, как мы упомянули ранее, напрямую отражается на принятии дальнейших решений. Описательная аналитика очень важна, но отчеты и дашборды на основе нее могут дать лишь общее представление об исходных данных, а принимать решения придется на основе собственного опыта и знаний предметной области. В то же время предписывающие модели способны помочь в принятии реальных решений на практике.

1.1. Рост спроса на аналитику

По итогам исследования Digital Readiness Survey 2021 года 89 % респондентов в области информационных технологий сообщили о росте потребности в бизнес-аналитике за последние два года[1]. Годом позже опрос руководителей производства и логистики показал, что 22 % из них уже используют средства предсказательной и предписывающей аналитики, а четыре пятых респондентов собираются воспользоваться ими в ближайшие пять лет[2]. Кроме того, уже в 2019 году 60 % руководителей медицинских учреждений сказали, что активно используют в работе методы предсказательной аналитики. А не так давно отделы по управлению персоналом вышли на лидирующие позиции в области использования аналитических приложений.

Росту востребованности аналитики и систем принятия решений в организациях поспособствовали следующие факторы:

- бурный рост объемов, разнообразия и скорости изменения данных;
- повышение эффективности программного и аппаратного обеспечения на фоне снижения их стоимости;
- рост спроса на принятие решений на основе данных.

---

[1] https://www.manageengine.com/the-digital-readiness-survey-2021.

[2] 2022 «MHI Annual Industry Report» (https://www.mhi.org/publications/report).

Бурный рост объемов, разнообразия и скорости изменения данных

Сегодня данных собирается значительно больше в сравнении с прошлыми годами. При этом увеличился не только объем данных, но также их насыщенность и разнообразие. На протяжении долгих лет аналитические подходы главным образом применялись только к структурированным количественным данным с понятным форматом и четко определенными полями, а предметными областями были финансы, производство, продажи и управление персоналом. Сегодня, с появлением большого количества данных от коммерческих и некоммерческих организаций, из социальных сетей, мобильных устройств и т. д., аналитика серьезно расширила свое присутствие. К тому же появились техники эффективного извлечения информации из текстовых данных, изображений, видео и аудио.

В далеком 2001 году аналитик Gartner Дуг Лейни (Doug Laney) сформулировал постулат, получивший название 3V и использующийся для описания взрывного роста данных[1]. Лейни заметил, что рост данных происходит по трем следующим направлениям:

- **объем** (volume): количество, или размер, данных. Большие данные, что ясно из названия, характеризуются беспрецедентными объемами. Даже префиксы, используемые для количественного измерения таких данных, не поспевают за их объемами. Сегодня уже используются следующие единицы измерения: мега- (10^6), гига- (10^9), тера- (10^{12}), пета- (10^{15}), экса- (10^{18}), зетта (10^{21}) и йотта- (10^{24});
- **разнообразие** (variety): сюда относится полнота вариативности источников данных, включая структурированные базы данных, неструктурированные данные, автоматизированное чтение которых существенно затруднено, временные ряды, пространственные данные, языковые структуры, медиа и данные с подключенных устройств;
- **скорость изменения** (velocity): скорость, с которой данные извлекаются, сохраняются, изменяются и анализируются. Иногда это происходит в реальном времени.

Одними из важнейших и быстрорастущих источников данных сегодня являются встроенные датчики, собирающие и передающие информацию из подключенных устройств, таких как производственные машины, автомобили и бытовые приборы. Этот феномен, получивший название *интернет вещей* (Internet of Things), позволил сделать интеллектуальными системы охраны, автоматизировать сельскохозяйственную технику, лишить автомобиль водителя и многое другое. Данные сегодня генерируются буквально всеми устройствами, начиная от камер и светофоров и заканчивая датчиками частоты пульса. Это позволяет аналитикам внедряться в любые сферы

[1] Позже другие авторы расширили этот постулат до 5V, добавив термины *достоверность* (veracity) и *ценность* (value): https://www.techtarget.com/searchdatamanagement/definition/5-Vs-of-big-data.

жизнедеятельности человека. Розничные магазины фиксируют любые действия потенциальных покупателей, а медицинские учреждения отслеживают состояние здоровья своих пациентов удаленно. В традиционных индустриях производится постоянный мониторинг множества метрик и показателей логистики. Эти данные можно использовать для определения и предотвращения узких мест в деятельности компаний.

Некоторые из генерируемых данных представляют собой побочный продукт использования интернета, смартфонов и прочих технологий и устройств. Такие данные получили название *выхлоп данных* (data exhaust) по аналогии с выхлопом автомобиля.

Безусловно, рост объема доступных данных, особенно это касается персональных данных, использовать нужно с осторожностью. Законы не поспевают за техническим прогрессом, так что ответственность за этические вопросы в приложениях ложится на плечи аналитиков.

Повышение эффективности программного и аппаратного обеспечения на фоне снижения их стоимости

Эра вычислительных машин началась порядка 70 лет назад. В середине прошлого века компьютеры были величиной с комнату и требовали очень хорошего охлаждения. Кроме того, один или несколько операторов постоянно должны были присутствовать в помещении и контролировать все процессы вручную. С тех пор темпы процесса миниатюризации вычислительных устройств постоянно повышались и в итоге достигли невероятного уровня.

Первым на рост вычислительных мощностей еще в 1960-х годах внимание обратил профессор Калифорнийского технологического института и один из основателей компании Intel Corporation Гордон Мур (Gordon Moore). Закон Мура, согласно которому количество транзисторов, размещаемых на кристалле интегральной схемы, удваивается каждые 18–24 месяца, описал экспоненциальный рост мощности микропроцессоров начиная с 1965 года. Вас может заинтересовать любопытное сравнение между одной из первых вычислительных машин, выполненных полностью на транзисторах, IBM 7090, произведенной в 1961 году, и современным ноутбуком. Процессор последнего превышает в быстродействии IBM 7090 в 30 000 раз. Если бы такая производительность была у первого компьютера, он стоил бы на наши деньги порядка $20 млн вместо $500–2000, которые стоит современный ноутбук. Таким образом, сегодня в распоряжении аналитиков есть вычислительные мощности, способные обрабатывать огромные массивы данных с помощью самых эффективных техник буквально за секунды.

Кроме того, положительным образом на доступность программного обеспечения для датамайнинга и предсказательной аналитики повлияли еще несколько факторов. В частности, получили широкое распространение облачные вычислительные ресурсы, позволяющие по требованию выполнять объемные расчеты с оплатой только задействованных при этом мощностей. Лидерами в этой области являются компании Microsoft, Amazon Web Services, IBM и Google.

В то же время программное обеспечение, предназначенное для аналитических расчетов, постепенно и постоянно прибавляло в эффективности и наглядности и теряло в цене. Это существенно снизило порог входа в область построения аналитических моделей. Сегодня языки программирования с открытым исходным кодом, такие как Python и R, предоставляют бесплатный доступ к высококлассным аналитическим инструментам всем, кто пожелает. Раньше использование эффективных аналитических пакетов требовало довольно продолжительного обучения, а для их применения на практике нужно было вводить мудреные команды со сложным синтаксисом. Современные аналитические инструменты позволяют управляться едва ли не одной мышью, что делает их доступными для большего количества людей. Более того, многие программные комплексы с открытым исходным кодом, такие как KNIME, H2O, Orange и Weka, являются бесплатными.

Многие инструменты на рынке позволяют быстро и без труда строить базовые визуальные представления данных в виде столбчатых и круговых диаграмм, линейных графиков и диаграмм рассеяния. Но это лишь малая часть полного джентльменского набора аналитика, который включает в себя отображение содержимого в динамике с анимацией, мониторинг в реальном времени и создание богатых интерактивных дашбордов на основе множества визуальных элементов. Полноценные аналитические программные комплексы, такие как Power BI Desktop, Tableau и Qlik, позволяют создавать визуализации, граничащие с шедеврами мирового искусства, по которым можно делать глубокие выводы.

Рост спроса на принятие решений на основе данных

Сегодня продвинутая аналитика используется во всех типах организаций, как в коммерческих, так и в некоммерческих. В то же время в полной мере пользу от углубленного анализа данных осознают лишь те, кто принимают на его основе решения. В основной же своей массе компании до сих пор не раскрыли для себя потенциал применения аналитики на практике.

Самый ранний пример использования предсказательной аналитики на практике восходит к далекому 1689 году, когда в компании Lloyd Company догадались воспользоваться историческими данными для оценки рисков путешествий по морю и тем самым сделали себе состояние в области страхования таких поездок[1]. Также можно вспомнить не такой давний пример использования аналитики компанией Fair Isaacs Corporation, которая в 1950-х годах разработала систему оценки рейтинга кредитоспособности потенциальных заемщиков в США (FICO score)[2]. С тех пор аналитику стали использовать на практике гораздо чаще и во всех без исключения областях деятельности. К примеру, в недавнем прошлом появилась тенденция к использованию аналитических инструментов в отделах по работе с персоналом в компаниях.

---

[1] «A Brief History of Predictive Analytics». 2019 (https://medium.com/@predictivesuccess/a-brief-history-of-predictive-analytics-f05a9e55145f).

[2] «FICO History» (https://www.fico.com/en/history).

Это позволило автоматизировать процесс отбора резюме, лучше справляться с текучкой кадров, оптимизировать премиальные выплаты, осуществлять планирование кадрового резерва и курсы повышения квалификации сотрудников.

Таким образом, преимущества от использования данных, полученных в результате углубленного анализа и принятых на их основе тактических и стратегических решений, дали новый толчок к развитию области практической аналитики. В результате в руководящих отделах компаний стали все реже принимать решения, основываясь на интуиции, и все чаще опираться на аналитические выводы. Знаменитый отчет под названием *Большие данные: следующий рубеж инноваций, конкуренции и производительности*, опубликованный международной консалтинговой компанией McKinsey, предложил глубокий взгляд на область больших данных и предвосхитил огромную пользу от добычи данных в таких областях, как здравоохранение, обслуживание населения, управление, розничные продажи, производство и логистика. В отчете также подчеркивался острый дефицит опыта интеллектуального анализа данных и недостаток управленцев, способных грамотно работать с данными.

1.2. Применение аналитики

Аналитика используется в компаниях очень по-разному. Иногда креативный подход к аналитике не знает границ. К примеру, в 2008 году компания Google запустила проект под названием Project Oxygen, призванный проанализировать работу своего управляющего персонала. Обработав более 10 тыс. отзывов о работе сотрудников, в компании смогли сформулировать принципы руководства, способствующие снижению текучки кадров и сохранению важных и ценных работников. В результате это позволило существенно повысить уровень удовлетворенности сотрудников компании и снизить напряженность.

Аналитика активно используется в профессиональном и любительском спорте для повышения результатов команд, поиска новых игроков и увеличения прибыли.

В индустрии, связанной с недвижимостью, аналитика применяется в самых разных аспектах. Безусловный лидер по использованию аналитических инструментов в этой отрасли – сеть отелей Marriott – активно использует накопленные данные для поиска новых источников дохода, привлечения новых посетителей, повышения уровня доверия у людей, время от времени пользующихся услугами сети, и развития бизнеса в других областях.

Не секрет, что компания Amazon располагает сразу несколькими рекомендательными системами по книгам, товарам широкого потребления, музыке и видео, позволяющими предсказать покупательское поведение посетителей сайта. Многие магазины используют рекомендательные системы, но у Ama-