Много цифр. Анализ больших данных при помощи Excel

Автор:
Джон Форман

Много цифр. Анализ больших данных при помощи Excel

Джон Форман

Казалось бы, термин «большие данные» понятен и доступен только специалистам. Но автор этой книги доказывает, что анализ данных можно организовать и в простом, понятном, очень эффективном и знакомом многим Excel. Причем не важно, сколько велик ваш массив данных. Техники, предложенные в этой книге, будут полезны и владельцу небольшого интернет-магазина, и аналитику крупной торговой компании. Вы перестанете бояться больших данных, научитесь видеть в них нужную вам информацию и сможете проанализировать предпочтения ваших клиентов и предложить им новые продукты, оптимизировать денежные потоки и складские запасы, другими словами, повысите эффективность работы вашей организации. Книга будет интересна маркетологам, бизнес-аналитикам и руководителям разных уровней, которым важно владеть статистикой для прогнозирования и планирования будущей деятельности компаний.

Джон Форман

Много цифр. Анализ больших данных при помощи Excel

Переводчик А. Соколова

Редактор Л. Мамедова

Руководитель проекта М. Шалунова

Корректор Е. Чудинова

Компьютерная верстка К. Свищёв

Дизайн обложки Ю. Буга

* * *

Editor’s choice – выбор главного редактора

Разговоры о Big Data идут уже давно, есть и книги на эту тему. Но в общем и целом все они были о том, что Big Data – «круто», этим занимаются ведущие компании мира, а вот и кейсы от этих компаний.

Теперь же у нас есть книга, которая показывает, как работать с Big Data практически, причем без сложных программ, на обычном Excel. Изучив ряд несложных приемов, руководители малого и среднего бизнеса смогут находить в массивах своих данных неочевидные зависимости, которые позволят получить серьезное конкурентное преимущество.

Знания – это сила, а знания, полученные из больших данных, – большая сила.

Сергей Турко, главный редактор издательства «Альпина Паблишер»

Моей жене Лидии.

То, что ты делаешь каждый день – круто! Если бы не ты, я бы лишился волос (и ума) миллиард лет назад

Введение

Что я здесь делаю?

Наверняка где-нибудь в прессе, финансовой литературе и журналах или на конференции вы слышали что-то об обработке данных, их представлении и анализе – том, что составляет «науку о данных». Эта наука может предсказать результаты выборов, рассказать о ваших покупательских привычках больше, чем вы осмелились бы поведать маме, и определить, на сколько лет сокращают вашу жизнь сырные буррито с чили.

В последнее время вокруг науки о данных наблюдается некоторый ажиотаж, который начинает оказывать давление на многие виды бизнеса. Не занимаясь анализом данных, вы рискуете потерпеть неудачу в конкурентной борьбе. Обязательно появится кто-нибудь, разработавший очередной новый продукт под названием «Что-то-про-графы-и-большие-данные», – и уничтожит ваш бизнес.

Сделайте глубокий вдох.

Не все так мрачно! Вас, несомненно, спасет то, что большинство тех, кто считает себя «доками» в науке о данных, делают все ровно наоборот. Они начинают с покупки программ и нанимают консультантов. Они тратят все свои деньги еще до того, как поймут, чего же они на самом деле хотят. Заказав программные инструменты, они считают, что сделали главное и можно расслабиться.

Прочитав эту книгу, вы будете на голову выше этих «специалистов». Вы будете иметь точное представление о том, что такое техники анализа данных и как они используются. И когда придет время планировать, нанимать и покупать, вы уже будете знать, как применить возможности науки о данных с пользой именно для вашей конкретной компании.

Цель этой книги – введение в практическую науку о данных в комфортном режиме беседы. Надеюсь, что по окончании чтения священный ужас перед этим таинственным «зверем» – данными – сменится энтузиазмом и мыслями о том, как с их помощью поднять свой бизнес на новый уровень.

Рабочее определение науки о данных

В некоторой степени наука о данных – синоним таких терминов, как бизнес-аналитика; исследование операций; бизнес-интеллект; промышленный шпионаж; анализ, моделирование и раскрытие данных (также называемое обнаружением знаний в базах данных, или ОЗБД). Иными словами, нынешняя наука о данных – просто новый виток того, чем люди занимаются уже довольно долго.

После расцвета вышеозначенных и других дисциплин произошел скачок в технологиях. Совершенствование аппаратной и программной платформ сделали легким и недорогим сбор и анализ больших объемов данных во всех областях – будь то продажи и маркетинг, запросы HTTP с вашего сайта или информация для поддержки клиентов. Малый бизнес и некоммерческие организации могут теперь привлекать аналитиков, содержание которых раньше могли себе позволить только большие корпорации.

Конечно, из-за того, что наука о данных используется как всеобъемлющее ученое словечко для обозначения аналитики сегодня, она чаще всего ассоциируется с техниками добычи данных (data mining), такими как искусственный интеллект, кластерный анализ и определение выбросов. Благодаря подешевевшей аппаратной поддержке, обеспечившей резкий рост количества переменных бизнес-данных, эти вычислительные техники стали опорой бизнеса в последние годы, хотя раньше они были слишком громоздкими для использования на производстве.

В этой книге я собираюсь дать широкий обзор всех разделов науки о данных. Вот определение, которое я буду использовать:

Наука о данных – это трансформация данных методами математики и статистики в рабочие аналитические выводы, решения и продукты.

Я определяю это понятие с точки зрения бизнеса. В нем упоминается применимый и полноценный готовый продукт, получаемый из данных. Почему? Потому что я занимаюсь этим не в исследовательских целях и не из любви к искусству. Я изучаю данные для того, чтобы помочь моей компании работать лучше и постоянно повышать свою эффективность; поскольку вы держите в руках мою книгу, подозреваю, что наши намерения схожи.

Используя это определение, я собираюсь описать вам основные техники анализа данных, такие как оптимизация, прогнозирование и моделирование, а также затронуть наболевшие темы – искусственный интеллект, сетевые графы, кластерный анализ и определение выбросов.

Одни из этих техник довоенные в буквальном смысле слова. Другие внедрены в течение последних 5 лет. Но вы увидите, что возраст не имеет никакого отношения к сложности или полезности. Все эти техники – независимо от степени популярности – одинаково полезны для бизнеса при правильном выборе.

Вот почему вам нужно понимать, какая техника для решения какой проблемы подходит, как эти техники работают и как их моделировать. Довольно много людей имеют представление о сути одной или двух описанных мною техник – этим их знания и ограничиваются. Если бы у меня в ящике для инструментов был только молоток, наверное, я бы пытался решать все проблемы ударом посильнее. Совсем как мой двухлетний сын.

Но поскольку мне не два года, я предпочитаю иметь еще какие-то инструменты в своем распоряжении.

Но подождите, а как же большие данные?

Наверняка вы слышали термин «большие данные» даже чаще, чем «наука о данных». О них ли эта книга?

Ответ зависит от того, что понимать под большими данными. Если вы определяете большие данные как подсчет сводной статистики неструктурированного мусора, хранящегося в горизонтально масштабируемом NoSQL-массиве, то нет, это книга не о больших данных.

Если вы определяете большие данные как превращение переменных данных в решения и аналитические выводы с помощью ультрасовременных методов анализа (независимо от того, где хранится информация), тогда да, моя книга о больших данных.

В этой книге не рассматриваются системы управления базами данных, такие как MongoDB и Hbase. В ней не рассказывается о пакетах для разработчиков, таких как Mahout, Numpy, различных R-библиотеках и т. д. Для этого существуют другие книги.

Я сделал так намеренно. Эта книга игнорирует инструменты, хранилища и код. Вместо этого она, по возможности, фокусируется на методах. Многие думают, что если смешать хранение и извлечение данных с щепоткой очистки и агрегации, получится коктейль «Все, что нужно знать о больших данных».

Они ошибаются. Эта книга поможет вам беспрепятственно пробиться сквозь завесу многозначительной болтовни, которой нас окружают продавцы программного обеспечения для работы с большими данными и блогеры, и покажет вам, на что на самом деле способны ваши данные.
Что примечательно, для большинства этих техник объем ваших данных может быть любым – крошечным или огромным. Вы не обязаны иметь петабайт данных и энную сумму с пятью нулями на предсказание интересов вашей огромной клиентской базы. Иметь массив данных – это, конечно, замечательно, однако есть бизнесы, прекрасно обходящиеся и без этого «сокровища», более того – никому не хочется их генерировать. Например, мяснику, торгующему в моем родном квартале. Но это не значит, что его бизнесу помешало бы небольшое кластерное разделение «бекон/колбаса».

Если сравнивать книги с видами спорта, моя книга сравнима с гимнастикой. Никаких тренажеров и упражнений на выносливость. Поняв, как реализовывать техники с помощью базовых инструментов, вы обнаружите, что свободно можете применять их во многих технологиях, с легкостью моделировать их, правильно выбирать программные продукты у консультантов, формулировать задачи программистам и т. д.

Кто я?

Давайте прервемся ненадолго, и я расскажу вам о себе. Научный подход к изучению данных, который я проповедую, возник не вчера – к нему меня вел долгий путь. Много лет назад я был консультантом по менеджменту. Я работал над аналитическими проблемами таких организаций, как ФБР, министерство

обороны США, компания Coca-Cola, группы отелей Intercontinental и Royal Carribbean. Из всего этого опыта я вынес одно: наука о данных должна стать прерогативой не только ученых.

Я работал с менеджерами, которые покупали симуляции, когда им были нужны модели оптимизации. Я работал с аналитиками, которые понимали только графики Ганта[1 - Популярный тип столбчатых диаграмм (гистограмм), который используется для иллюстрации плана, графика работ по какому-либо проекту. Является одним из методов планирования проектов. – Прим. ред.], так что абсолютно все приходилось представлять в виде этих графиков.
Как консультанту, мне было нетрудно расположить к себе покупателя, имея в арсенале любые старые бумаги и миленькую презентацию в PowerPoint, потому что они не могли отличить искусственный интеллект от бизнес-анализа, а бизнес-анализ – от BS.

Цель этой книги – расширение аудитории, способной понять и применить техники научного анализа данных. Я не пытаюсь обратить вас, уважаемые читатели, в специалистов по научной обработке данных против вашей воли. Я просто хочу, чтобы вы научились применять науку о данных настолько, насколько сможете, в той области, в которой вы уже хорошо разбираетесь.

Это заставляет задать вопрос: кто же вы?

Кто вы?

Не пугайтесь, я не использовал научный анализ данных, чтобы шпионить за вами. Я понятия не имею, кто вы, но заранее благодарен вам за то, что раскошелились на эту книгу.

Вот несколько архетипов (или личностей – для вас, маркетологи!), которые пришли мне на ум, когда я писал эту книгу. Возможно, вы:

• заместитель начальника по маркетингу и хотите использовать свои бизнес-переменные стратегическим образом, для оценки продукта и сегмента рынка, но не понимаете подходов, рекомендуемых разработчиками приложений

и переоцененными консультантами;

• аналитик, предсказывающий спрос, который знает, что история заказов фирмы содержит больше информации о клиентах, чем даже план на следующий квартал;

• руководитель розничного интернет-магазина, желающий угадать по данным о предыдущих заказах, когда клиент скорее всего «созреет» для очередной покупки;

• бизнес-аналитик, который в состоянии просчитать растущие денежные потоки и затраты на снабжение, но не знает, как перебросить мостик экономии на издержках;

• онлайн-маркетолог, который хочет чего-то большего для своей компании от бесплатных текстовых сервисов, таких как электронные письма или социальные сети. Пока же судьба разосланных сообщений незавидна – их открывают и тут же выбрасывают в корзину.

Иными словами, вы – читатель, который получает практическую пользу от дополнительной информации о научной обработке данных, но пока не нашел «свой конек» во всем многообразии техник. Цель этой книги – стряхнуть мишуру (код, инструменты и просто слухи) с науки о данных и обучить необходимым техникам на практических примерах, понятных любому, прошедшему курс линейной алгебры или вычислительной математики в институте. Если вы, конечно, их успешно сдали. Если нет – читайте медленно и не стесняйтесь пользоваться Википедией.

Никаких сожалений – только электронные таблицы

Эта книга не о программировании. Я даже готов гарантировать полное отсутствие (ну, по крайней мере, до главы 10) в ней кода. Почему?

Да потому что я не хочу тратить первые сто страниц на возню с Git, объявлением переменных среды и выступление Emacs против Vi.

Если вы пользуетесь исключительно Windows и Microsoft Office, работаете в государственной структуре и вам запрещено скачивать и устанавливать приложения из каких попало открытых источников и даже если MATLAB или ваш графический калькулятор наводили на вас леденящий ужас во времена студенчества, вам нечего бояться.

Нужно ли вам знать, как пишется код, чтобы перевести большую часть этих техник в автоматизированную, производственную форму? Непременно! Вы или кто-то из ваших коллег должен знать технологии хранения данных и уметь управляться с кодом.

Нужно ли вам знать, как пишется код, чтобы понимать, различать и моделировать эти техники? Совершенно ни к чему!

Именно поэтому я объясняю каждую методику с помощью электронных таблиц.

Ну, ладно, если по-хорошему, то я должен признаться, что все вышесказанное мною не совсем правда. Последняя глава этой книги – о переходе на язык программирования R, ориентированный на анализ данных. Она предназначена для тех из вас, кто захочет использовать эту книгу как трамплин к пониманию новых глубин аналитики.

Но электронные таблицы так устарели!

Электронные таблицы – не самый привлекательный инструмент из существующих.

Электронные таблицы стоят немного особняком. Они позволяют вам видеть данные и взаимодействовать с ними (или, по крайней мере, кликать на них). Они создают определенную свободу для маневра. Во время изучения анализа данных вам понадобится инструмент – привычный, понятный каждому, позволяющий двигаться быстро и легко в процессе. Это и есть электронные

таблицы.

Давайте, наконец, скажем себе: «Я человек и обладаю чувством собственного достоинства. Я не должен делать вручную работу программного фреймворка, чтобы научиться анализировать данные».

А еще электронные таблицы отлично подходят для прототипирования! Конечно, вы не запустите с их помощью производственную модель ИИ[2 - Искусственный интеллект. – Прим. пер.] для вашего интернет-магазина из программы Excel, но зато сможете понять характер заказов, спрогнозировать, какие продукты в будущем вызовут интерес потребителей, и разработать прототип модели для определения целевой аудитории.

Используйте Excel или LibreOffice

Все примеры, с которыми вам придется работать, отображаются в таблицах Excel.

На сайте этой книги (www.wiley.com/go/datasmart (http://www.wiley.com/go/datasmart)) размещены электронные таблицы с открытым доступом для каждой главы, так что вы сможете следить за ходом повествования. Если вы по натуре склонны к риску, можете стереть оттуда все данные, кроме исходных, и сделать всю работу самостоятельно.

Эта книга совместима с Excel версий 2007, 2010, 2011 для MacOS и 2013. В первой главе достаточно подробно рассматриваются различия между версиями.

У большинства из вас есть доступ к Excel и вы наверняка уже применяете его в вашей работе для отчетности или хранения информации. Но если по какой-то причине этой программы у вас нет, вы можете ее либо купить, либо воспользоваться бесплатным аналогом от LibreOffice (www.libreoffice.org (http://www.libreoffice.org/)).

А как же Google Drive?

Кто-то из вас наверняка спросит, можно ли при решении задач, которые нам предстоят, использовать Google Drive – облачный сервис, доступный с любого устройства, как почтовый ящик. Что и говорить, вариант заманчивый…
К сожалению, он не будет работать.

Google Drive отлично справляется с небольшими таблицами, но того, чем собираетесь заниматься вы, он просто не выдержит. Процесс добавления строк и колонок уже раздражает, реализация поиска решения просто ужасна, а у графиков даже нет линий тренда!

Хотелось бы мне, чтобы было иначе, но увы…

LibreOffice – открытый бесплатный ресурс, имеющий практически всю функциональность Excel. Я даже думаю, что его собственный поиск решений предпочтительнее, чем у Excel. Так что если вы не раздумали читать эту книгу – вперед!

Условные обозначения

Чтобы помочь вам извлечь из текста максимальную пользу, я ввел в эту книгу несколько условных обозначений.

Вставки

Вставки типа той, в которой вы только что прочитали про Google Drive, раскрывают «побочные» темы, упомянутые в тексте.

Внимание!

Эти разделы содержат важную информацию, напрямую связанную с текстом, которую я рекомендую все время держать в уме.

Заметки

Здесь вы найдете советы, подсказки, приемы и все в этом духе, что пришлось к слову в текущем обсуждении.

Частенько я буду вставлять в текст небольшие кусочки кода Excel вроде этого:

=CONCATENATE(“THIS IS A FORMULA”, “IN EXCEL!”)/

=СЦЕПИТЬ(“ЭТО ФОРМУЛА”, “В EXCEL!”)

Мы выделяем курсивом новые термины и важные слова при первом упоминании. Названия файлов, веб-страниц и формул в тексте выглядят так:

http://www.john-foreman.com.

Итак, начнем

В первой главе я намерен заполнить некоторые пробелы в ваших познаниях об Excel, после чего вы сможете погрузиться непосредственно в практику. К концу книги вы не только будете иметь представление о нижеперечисленных техниках, но и приобретете опыт их применения:

• оптимизация с использованием линейного и интегрального программирования;

• работа с временными рядами данных, определение трендов и изменений сезонного характера, а также прогнозирование методом экспоненциального

сглаживания;

• моделирование методом Монте-Карло в оптимизации и прогнозировании сценариев для количественного выражения и адресации рисков;

• искусственный интеллект с использованием общей линейной модели, функции логистических звеньев, ансамблевых методов и наивного байесовского классификатора;

• измерение расстояния между клиентами с помощью близости косинусов угла, создание K-ближайших граф, расчет модулярности и кластеризация клиентов;

• определение выбросов в одном измерении по методу Тьюки или в нескольких измерениях с помощью локальных факторов выброса;

• применение пакетов R для использования результатов работы других программистов при выполнении этих задач.

Если хотя бы что-то из вышесказанного звучит для вас воодушевляюще – продолжайте чтение! Если пугающе – то тоже продолжайте! Я торжественно обещаю разжевывать все как можно тщательнее.

Итак, без лишней суеты приступим!

1. Все, что вы жаждали знать об электронных таблицах, но боялись спросить

В этой книге я исхожу из того, что вы уже имеете некоторое представление об электронных таблицах и пользуетесь ими. Если же вы никогда не сталкивались с расчетами по формулам, вам поначалу придется нелегко. Перед нашим совместным погружением в Excel с головой я бы рекомендовал проштудировать «Excel для чайников» или другую подобную литературу вводного уровня.

Но даже если вы – заслуженный мастер по работе с Excel, все равно в моем тексте иногда будут возникать упоминания о таких возможностях программы, которыми вы никогда не пользовались. Так, в настоящей главе вы встретите много небольших приемов с простыми функциями. Некоторые наверняка покажутся вам немного странными. Не зацикливайтесь на непонятном – двигайтесь дальше. Вы всегда сможете вернуться к недочитанной главе позже.

Отличия в разных версиях Excel

Как я уже упоминал во введении, для этой книги подходят Excel 2007, 2009, 2011 для MacOS и LibreOffice. К сожалению, в каждой новой версии Excel разработчики Microsoft перемещают инструменты и функционал как им угодно.

Например, элементы из вкладки «Разметка» в версии 2011 года находятся во вкладке «Вид» во всех остальных версиях. «Поиск решения» в версиях 2010 и 2013 одинаковый, но реализован он лучше в 2007 и 2011, несмотря на гротескный интерфейс в версии 2007.

Снимки с экрана в этой книге будут делаться с Excel 2011. Если у вас более новая или старая версия, вам придется действовать по-другому, особенно когда дело касается положения элементов управления во вкладках меню. Я очень постараюсь найти и учесть все различия. Если что-то мной упущено, поисковый инструмент Excel и Google всегда к вашим услугам.

А вот то, что, несомненно, должно вас обрадовать: «табличная часть электронной таблицы» всегда неизменна.

Несколько слов о LibreOffice. Если вы решили пользоваться открытыми источниками программного обеспечения, рискну предположить, что вы – человек, склонный до всего доходить самостоятельно. И хотя я не буду напрямую обращаться к интерфейсу LibreOffice, вы этого попросту не заметите. Они с Excel похожи как две капли воды.

Немного данных для примера

Заметка:

Рабочая тетрадь Excel, используемая в этой главе «Concessions.xlsx», доступна для загрузки на сайте книги www.wiley.com/go/datasmart

Представьте себе, что вам жутко не везет по жизни. Даже став взрослым, вы до сих пор живете с родителями и работаете в киоске на баскетбольных матчах в своей старой школе. (Клянусь, это только наполовину автобиографично!).

У вас есть электронная таблица о вчерашних продажах, и выглядит она примерно как рис. 1–1.

На рис. 1–1 показана каждая продажа: что именно продано, к какому типу еды или напитков относится проданный товар, цена и процент прибыли с продажи.

Быстрый просмотр с помощью кнопок управления

Если хотите ознакомиться с записями – промотайте список колесиком мышки, пальцем (если у вас сенсорный экран) или стрелками клавиатуры. Пока вы просматриваете записи, приятно иметь строку заголовков в поле зрения – тогда вы точно не забудете, что в какой колонке записано. Для этого выберите «Закрепить области» или «Закрепить верхнюю строку» во вкладке «Вид» в Windows (вкладка «Разметка» в MacOS2011, как показано на рис. 1–2).

Чтобы быстро переместиться в конец документа и посмотреть, сколько всего у вас продаж, выберите значение в одном из заполненных столбцов и нажмите Ctrl+? (Command+? на Mac). Вас отбросит прямо к последней заполненной ячейке этого столбца. В этой таблице последняя строка – 200. Заметьте также, что кнопка Ctrl/Command со стрелками даст вам возможность точно так же перемещаться по всему документу, в том числе вправо и влево.

Если вы хотите узнать среднюю прибыль за единицу проданного за вечер, то под столбцом с ценами (столбцом C) можно вбить формулу:

=AVERAGE(C2:C200)/

=СРЗНАЧ(C2:C200)

Средняя прибыль получается $2,83, так что отдыхать от трудов праведных вам, увы, еще не время. Подсчет можно произвести и другим способом: переместиться на последнюю ячейку в столбце, C200 и удерживать Shift+Ctrl+?, чтобы выделить весь столбец доверху, а затем выбрать «Среднее значение» на нижней панели справа (рис. 1–3). В Windows нужно кликнуть на этой панели для того, чтобы выбрать среднее значение, скажем, вместо суммы, стоящей там по умолчанию. В MacOS, если нижняя панель отключена, нужно нажать на меню «Вид» и выбрать «Строку состояния», чтобы включить ее.

Быстрое копирование формул и данных

Пожелав видеть свою прибыль в фактических долларах, а не в процентах, вы можете добавить что-то вроде заголовка в столбец E, который назовем «Фактическая прибыль». В ячейке E2 нужно просто перемножить соответствующие значения из столбцов с ценой и прибылью, чтобы получить такую формулу:

=C2*D2

Для строки с пивом результат будет равным $2. Не нужно переписывать формулу для каждой строки. Excel позволяет переносить формулы из ячейки перетаскиванием за правый нижний угол куда вам угодно. Значения в столбцах C и D будут меняться в зависимости от того, куда скопирована формула. Если, как в случае с данными о продажах, столбец слева полностью заполнен, дважды кликните на правом нижнем углу ячейки с формулой, и Excel заполнит значениями весь столбец, как это показано на рис. 1–4. Попробуйте этот двойной клик сами, потому что я буду использовать его во всей книге. Освоив его сейчас, вы избавите себя от огромных неудобств в будущем.

Обязательно ли значение в ячейках, упомянутых в формуле, должно меняться в зависимости от того, куда вы ее перетаскиваете или копируете? Нет, конечно. Хотите оставить что-то неизменным – просто поставьте перед ним $.

К примеру, если вы измените формулу в E2 таким образом:

=C$2?D$2

В этом случае при копировании формулы на все последующие ячейки в ней ничего не меняется. Формула продолжает обращаться ко 2-й строке.

Если скопировать формулу вправо, то C заменится на D, D на E и т. д. Если вам не нравится такое «поведение», добавьте $ также перед ссылками на столбцы в формуле. Это называется абсолютной ссылкой, в противоположность относительной ссылке.

Форматирование ячеек

Excel предлагает статические и динамические опции для форматирования содержимого ячеек. Взгляните на столбец E с фактической прибылью, который вы только что создали. Выделите его, кликнув на серый заголовок колонки. Затем кликните на выбранном столбце правой клавишей и выберите «Формат ячеек».

В этом меню вы можете выбрать формат содержимого ячеек столбца E. В нашем случае нужен денежный формат. Также можно указать число знаков после запятой при округлении. Оставьте 2 знака после запятой, как показано на рис. 1–5. Также в меню «Формат ячеек» доступны такие опции, как цвет шрифта, заливка ячейки, выравнивание текста, границы и т. д.

Но есть нюанс. Допустим, нужно отформатировать только те ячейки, которые содержат определенные значения или диапазон значений, и это форматирование должно меняться в зависимости от значений.

Такой вид форматирования называется условным форматированием, и оно повсеместно используется в этой книге.

Закройте меню «Формат ячеек» и переместитесь во вкладку «Главная». В разделе «Стили» («Формат» в MacOS) вы найдете «Условное форматирование» (рис. 1–6). При нажатии на него выпадает меню. Самое используемое условное форматирование в этой книге – цветовые шкалы. Выберите шкалу для столбца E и посмотрите, как изменился цвет каждой ячейки в зависимости от величины значения в ней.

Чтобы очистить условное форматирование, используйте опцию «Удалить правила» меню условного форматирования.

Специальная вставка

Конечно, гораздо удобнее работать, если формулы не путаются у вас под рукой, как в колонке E на рис. 1–4. А если это еще и формулы вроде RAND()/СЛЧИС(), генерирующей случайные числа, которые меняют свое значение при каждом автопересчете таблицы, то ваше раздражение вполне справедливо. Решение проблемы – в копировании этих ячеек и вставке их обратно в таблицу в виде постоянных величин.

Чтобы перевести формулы в цифры, просто выделите и скопируйте столбец E, заполненный формулами, и вставьте его обратно с помощью опции «Специальная вставка» (находится во вкладке «Главная» под опцией «Вставить» в Windows и в меню «Редактировать» в MacOS). В окне «Специальная вставка» выберите вставку в качестве значений (рис. 1–7). Замечу, что это меню при вставке позволяет также транспонировать данные из вертикали в горизонталь и наоборот. Это свойство очень пригодится вам в дальнейшем.

Вставка диаграмм

Методичка, посвященная торговле с лотка, включает в себя графу «Калории» с малюсенькой табличкой. В ней указано, сколько калорий содержится в каждом напитке или закуске, которые продаются в киоске. Вы тоже легко можете сделать такую диаграмму в Excel. Во вкладке «Вставка» («Диаграммы» в MacOS) есть раздел, в котором находятся различные варианты отображения, такие как столбчатая гистограмма, линейный график и круговая диаграмма.

Заметка

В этой книге мы в основном будем пользоваться столбчатыми гистограммами, линейными графиками и графиками рассеяния. Никогда не пользуйтесь круговыми диаграммами! И особенно круговыми 3D-диаграммами, которые вам предлагает Excel. Не вздумайте ослушаться, иначе мой призрак будет мучить вас после моей смерти! Круговые диаграммы уродливы, плохо соотносятся с данными, эстетически их 3D-эффект примерно таков же, как у картинок из ракушек на стене кабинета моего стоматолога.

Выделяя столбцы A и B в листе «Calories» вы можете выбрать столбчатую диаграмму с группировкой для отображения данных. Поиграйте с графикой. Нажимайте на разделы правой клавишей мыши, чтобы увидеть меню форматирования. Например, щелчок правой клавиши на столбцах диаграммы вызовет меню, в котором можно выбрать «Формат рядов данных». Под ним вы сможете поменять цвет столбцов с синего по умолчанию на любой оттенок, который вам по вкусу, например черный.

В наличии легенды по умолчанию тоже нет никакого смысла, так что советую выделить ее и нажать «Удалить». Также вам может понадобиться выделить разные текстовые подписи к диаграмме и увеличить размер шрифта (размер шрифта находится под вкладкой «Главная»). Таким образом получается диаграмма, показанная на рис. 1–8.

Расположение меню поиска и замены

В этой книге вам частенько придется пользоваться функциями поиска и замены. В Windows это делается, как обычно, нажатием Ctrl+F для открытия окна поиска (и Ctrl+H для замены) или перемещением во вкладку «Главная», где в разделе «Правка» находится кнопка «Найти». В MacOS строка поиска расположена в верхнем правом углу листа (для замены нажмите либо стрелку вниз, либо Cmd+F для вызова меню поиска и замены).

Чтобы проверить прочитанное на практике, откройте меню замены на листе «Calories». Замените слово «Калории» на слово «Энергия» везде, где оно встречается (рис. 1–9), вбив эти слова в окно поиска и замены и нажав «Заменить все».

Формулы поиска и вывода величины

Если бы я не уточнил, что вам знакомы хотя бы некоторые простые формулы Excel (SUM, MAX, MIN, PERCENTILE / СУММ, МАКС, МИН, ПЕРСЕНТИЛЬ и т. д.), мы бы просидели здесь целый день. А я хочу начать анализировать данные. Вместе с тем я часто использую в этой книге формулы, с которыми вы могли ни разу не столкнуться, если до этого не погружались с головой в волшебный мир электронных таблиц. Эти формулы работают с поиском значения в ряду и выводом его положения или, наоборот, поиском положения в ряду и возвратом значения.

Я покажу это на примере листа «Calories».

Иногда хочется узнать положение элемента в столбце или строке. Первый он, второй или третий? Формула MATCH/ПОИСКПОЗ справляется с этим довольно неплохо. Под вашими данными о калориях назовите A18 Match/Поискпоз. Вы можете применить формулу к ячейке B18, чтобы найти, где в списке выше упоминается слово «Hamburger». Чтобы использовать эту формулу, необходимо указать в ней значение, которое нужно найти, границы поиска и 0, чтобы она вывела позицию самого слова:

=MATCH("Hamburger", A2:A15,0) /

=ПОИСКПОЗ("Hamburger", A2:A15,0)

Она выдает 6, так как «Hamburger» – шестая позиция в списке (рис. 1-10).

Следующая формула – INDEX / ИНДЕКС. Назовите ячейку A19 Index/Индекс.

Эта формула находит значение элемента по заданному положению в строке или столбце. Например, подставив в нее из нашей таблицы калорий A1:B15 и задав координаты поиска «3 строка, 2 столбец», мы получим количество калорий в бутылке воды:

=INDEX(A1:B15,3,2) /

=ИНДЕКС(A1:B15,3,2)

Мы видим количество калорий, равное 0, как и предполагалось (рис. 1-10).

Другая формула, которая часто встречается в нашем тексте, – это OFFSET/СМЕЩ. Назовем же ячейку A20 Offset/Смещ и поиграем с формулой в B20.

С помощью этой формулы вы задаете промежуток, который перемещаете, подобно курсору, по сетке из столбцов и строк (точно так же, как INDEX/ИНДЕКС ищет единственную ячейку, если только в нем не упомянут 0). Например, можно задать функции OFFSET/СМЕЩ рамки от верхней левой ячейки листа A1 и затем растянуть ее на 3 ячейки вниз, создавая ряд из 3 строк и 0 столбцов:

=OFFSET(A1,3,0) /

=СМЕЩ(A1,3,0)

Эта формула возвращает значение третьего элемента списка – «Chocolate Bar» (рис. 1-10).

Последняя формула, о которой я хочу сказать в этом разделе, – SMALL/НАИМЕНЬШИЙ (у него есть двойник – LARGE/НАИБОЛЬШИЙ, который работает точно так же). Если у вас есть список значений и вы хотите выбрать, скажем, третье наименьшее из них, данная функция делает это за вас. Назовите

ячейку A21 Small/Наименьший, а в B21 напишите следующую формулу, содержащую границы поиска и параметр 3:

=SMALL(B2:B15,3)/

=НАИМЕНЬШИЙ(B2:B15,3)

Эта формула возвращает значение 150, которое является третьим наименьшим после 0 (бутылка воды) и 120 (газировка), как показано на рис. 1-10.

И, наконец, еще одна формула для поиска значений, похожая на MATCH/ПОИСКПОЗ, употребившую стероиды. Это VLOOKUP/ВПР (и ее горизонтальный двойник HLOOKUP/ГПР). Им я уделю целый раздел, ибо это монстры.

Использование VLOOKUP/ВПР для объединения данных

Перейдем обратно к листу продаж на баскетбольных матчах. При этом мы в любое время можем обратиться предыдущему листу с калориями, просто указав его название и поставив перед номером ячейки «!». Например, Calories!B2 является отсылкой к количеству калорий в пиве, несмотря на то, что вы в данный момент работаете с другим листом.

Предположим, вы захотите увидеть количество калорий на листе продаж для каждого наименования товара. Вам нужно будет каким-то образом найти содержание калорий в каждом товаре и поместить его в колонку, следующую за прибылью. Что ж, оказывается, и для этого есть отдельная функция под названием VLOOKUP/ВПР.

Назовем колонку F в нашем листе «Calories / Калории». Ячейка F2 будет содержать количество калорий из таблицы в товаре из первой строки – пиве. Используя эту формулу, можно указать в названии товара из ячейки A2 ссылку на таблицу Calories!$A$1:$B$15 и номер столбца, из которого следует выбирать значения. В нашем случае он второй по счету:

=VLOOKUP(A2,Calories!$A$1:$B$15,2,FALSE) /

=ВПР(A2,Calories!$A$1:$B$15,2,ЛОЖЬ)

FALSE/ЛОЖЬ в конце формулы означает, что вам не подходят приблизительные значения «Beer». Если функция не может найти «Beer» в таблице калорий, она возвращает ошибку.

После ввода формулы вы увидите, что 200 калорий считались из таблицы в листе «Calories». Поставив $ в формуле перед ссылками на таблицу, вы можете скопировать формулу вниз по колонке двойным щелчком на нижнем правом углу ячейки. Оп-ля! У вас есть количество калорий для каждой позиции, как показано на рис. 1-11.

Фильтрация и сортировка

Отразив в листе продаж калорийность ваших товаров, задайтесь целью видеть, например, только товары из категории «Замороженные продукты» – иными словами, отфильтровать ваш лист. Для этого сначала выберите данные в рамках A1:F200. Наведите курсор на A1 и нажмите Shift+Ctrl+?, а затем ?. Есть способ еще проще – кликнуть наверху столбца и, удерживая клавишу мышки нажатой, переместить курсор к столбцу F, чтобы выделить все 6 столбцов.

Затем, чтобы применить автофильтрацию к этим шести колонкам, нажмите кнопку «Фильтр» из вкладки «Данные». Она похожа на серую воронку, как на рис. 1-12.

Если автофильтрация включена, можно кликнуть на выпадающем меню, которое появляется в ячейке B1, и выбрать для показа только определенные категории (в данном случае отобразятся товары из категории «Замороженные продукты»), как на рис. 1-13.

После фильтрации выделение столбцов данных позволяет нижней панели показывать краткую информацию об этих ячейках. Например, отфильтровав только замороженные продукты, можно выделить значения в столбце E и использовать нижнюю панель, чтобы быстро узнать сумму прибыли только по этой категории товара, как на рис. 1-14.

Автофильтрация позволяет также производить сортировку. К примеру, если вы хотите рассортировать прибыль, просто кликните на меню автофильтрации в ячейке Profit/Прибыль (D1) и выберите сортировку по возрастанию (или убыванию), как на рис. 1-15.

Чтобы убрать все фильтры, которые вы применяли, либо вернитесь в меню фильтрации по категориям и отметьте другие категории, либо отключите кнопку «Фильтр» во вкладке «Данные», нажатую в самом начале. Вы увидите, что, несмотря на возвращение всех ваших данных на свои места, «Замороженные продукты» остаются в том порядке, который был определен фильтром.

Excel также предлагает интерфейс для выполнения более сложных сортировок, чем те, на которые способна автофильтрация. Чтобы использовать его, выделите данные для сортировки (снова выберите A: F) и нажмите «Сортировка» в разделе

«Сортировка и фильтр» во вкладке «Данные». На экране появится меню сортировки. В MacOS для вызова этого меню нужно нажать стрелку вниз на кнопке сортировки и выбрать настройку.

В меню сортировки, показанном на рис. 1-16, независимо от наличия заголовка у столбцов с данными, можно выбрать колонки для сортировки по названию.

И теперь самая потрясающая часть этого сортировочного интерфейса, скрытая под кнопкой «Параметры». Нажмите ее, чтобы отсортировать данные слева направо вместо сортировки по колонкам. Это как раз то, чего не может автофильтрация. От начала до конца этой книги вам придется сортировать данные различным образом – и по столбцам, и по строкам, в чем вам очень поможет интерфейс сортировки. А сейчас просто выйдите из этого меню – ведь данные уже рассортированы так, как вам хотелось.

Использование сводных таблиц

Предположим, вам нужно знать количество проданного товара каждого типа или общую сумму выручки по определенному товару.

Эти задачи сродни запросам «aggregate» или «group by», используемым в традиционных базах данных SQL. Но наши данные – еще не база. Это – электронная таблица. И здесь нам на помощь приходят сводные таблицы.

После фильтрации вы начинаете с выделения данных, которыми хотите манипулировать. В нашем случае – данных о продажах в области A1:F20. Во вкладке «Вставить» (вкладка «Данные» в MacOS) выберите «Сводная таблица» и создайте ее на новом листе. Несмотря на то, что новые версии Excel позволяют вставлять сводную таблицу в существующий лист, ее, как правило, помещают на отдельном, если нет явной причины сделать иначе.

На новом листе конструктор сводных таблиц будет расположен справа от таблицы (в MacOS он перемещается). Он позволяет брать столбцы из листа с выделенными данными и использовать их как фильтры отчета, заголовки столбцов и строк для группировки или как значения. Фильтр отчета делает все то же самое, что и фильтр из предыдущего раздела, – позволяет вам выбрать определенный набор данных вроде «Замороженных продуктов». Заголовки столбцов и строк наполняют содержимое отчета сводной таблицы различными значениями из выделенных столбцов.

В Windows появляющаяся сводная таблица будет по умолчанию пустой, в то время как в MacOS она оказывается уже частично заполненной значениями из первого выделенного столбца в первом столбце и значениями из второго столбца во всех остальных. Если вы пользуетесь MacOS, то просто уберите все галочки в окнах конструктора и работайте дальше с пустой таблицей.

Допустим, теперь вам нужно знать объем выручки за каждый товар. Для этого перетащите ссылку Item/«Товар» в конструкторе сводных таблиц в поле строк, а ссылку Price/«Цена» – в поле данных. Это значит, что вы будете работать с доходом, сгруппированным по названию товара.

Изначально сводные таблицы создавались для простого подсчета количества записей о ценах внутри группы. Например, на рис. 1-17 есть 20 строк о пиве.

Нужно изменить подсчет количества записей на сумму значений. Чтобы это сделать в Windows, используйте выпадающее меню из ссылки «Цена» в поле данных и выберите в нем «Параметры поля данных». В MacOS нужно нажать маленькую кнопочку «i». В этом меню среди множества других опций можно выбрать сумму.

А что, если вам захотелось разбить эти суммы по категориям? Для этого в конструкторе перетащите ссылку «Категории» в поле столбцов. В итоге получается таблица, показанная на рис. 1-18. Заметьте, что сводная таблица на рисунке автоматически суммирует для вас строки и столбцы.

Если же вы хотите избавиться от каких-либо данных в своей таблице, просто снимите галочку в конструкторе или вытащите ссылку из поля, в котором она находится, будто решили ее выбросить. Избавьтесь, к примеру, от ссылки «Категория».

Когда требуемый отчет появился у вас в виде сводной таблицы, вы всегда можете выделить значения и вставить их в другой лист для дальнейшей работы. В нашем примере можно скопировать таблицу (A5:B18 в MacOS) и с помощью «Специальной вставки» перенести значения на новый лист под названием «Прибыль по каждой позиции» (рис. 1-19).

Поперемещайте разные заголовки строк и столбцов, пока вам не станет ясна процедура. К примеру, попробуйте подсчитать калорийность всех проданных позиций по категориям с помощью сводных таблиц.

Использование формул массива

В методичке по торговле с лотка есть графа под названием «Комиссия». Оказывается, тренер О'Шонесси позволяет вам торговать с лотка, только если вы отправляете ему некую часть своей прибыли (возможно, чтобы облегчить его затраты на носки без пяток, которые он привык покупать). Графа «Комиссия» отображает, сколько процентов прибыли он забирает с каждой продажи.

Как же узнать, сколько вы ему должны после вчерашнего матча? Чтобы ответить на этот вопрос, умножьте общую прибыль по каждой позиции из сводной таблицы на процент комиссии и затем сложите результаты.

Есть замечательная функция как раз для этой операции, которая умножит и сложит все, что надо, одним махом. Называется она довольно интересно – SUMPRODUCT/СУММПРОИЗВ. В ячейку E 1 листа с прибылью по каждой позиции добавьте заголовок «Общая комиссия Тренера». В C2 поместите SUMPRODUCT/СУММПРОИЗВ для расчета прибыли и процентов от нее:

=SUMPRODUCT(B2:B15,'Fee Shedule'!B2:O2) /

=СУММПРОИЗВ(B2:B15,'Комиссия'!B2:O2)

Ого, да здесь ошибка! В ячейке видно только #Value/#Значение. Что же случилось?

Несмотря на то, что вы выбрали две области данных одинакового размера и применили к ним функцию SUMPRODUCT/СУММПРОИЗВ, формула не видит их равенства, потому что одна область вертикальная, а другая – горизонтальная.

К счастью, в Excel есть функция для расположения массивов в нужном направлении. Она называется TRANSPOSE/ТРАНСП. Нужно написать такую формулу:

=SUMPRODUCT(B2:B15,TRANSPOSE('FeeSchedule'!B2:O2)) /

=СУММПРОИЗВ(B2:B15,ТРАНСП('Комиссия'!B2:O2))

Но нет! Все еще возникает ошибка.

Причина, по которой это происходит, такова: каждая функция Excel по умолчанию возвращает результат в виде единственного значения. Даже если TRANSPOSE/ТРАНСП вернет первое значение в транспонированный массив. Если вы хотите видеть в результате целый массив, нужно включить TRANSPOSE/ТРАНСП в «формулу массива». Формулы массива действительно возвращают результат в виде массива, а не единственного значения.

Вам не нужно писать SUMPRODUCT/СУММПРОИЗВ как-то иначе, чтобы все получилось. Все, что следует сделать – это вместо клавиши «Enter» после ввода формулы нажать «Ctrl+Shift+Enter». В MacOS нужное сочетание –

«Command+Return».

Победа! Как видно на рис. 1-20, результатом вычислений является 57,60 долларов. Но я бы округлил его до 50 – неужели Тренеру нужно столько носков?

Решение задач с помощью «Поиска решения»

Многие техники, которым вы научитесь в этой книге, могут быть сведены к моделям оптимизации. Проблема оптимизации – из разряда тех, для которых нужно подобрать лучшее решение (наилучший способ инвестирования, минимизировать траты вашей компании и т. д.). Применительно к моделям оптимизации приходится часто пользоваться словами «минимизировать» и «максимизировать».

В рамках науки о данных многие приемы, чего бы они ни касались – искусственного интеллекта, извлечения и анализа данных или прогнозирования – на деле состоят из некоторой подготовки данных и стадии подбора модели, которая на самом деле и является моделью оптимизации. Так что, по моему мнению, имеет смысл сначала научиться оптимизации. Но просто взять и выучить все об оптимизации невозможно. Мы проведем углубленное исследование оптимизации в главе 4, после того, как вы немного «поиграете» с проблемами машинного самообучения в главах 2 и 3. Но чтобы заполнить пробелы, неплохо было бы немного попрактиковаться в оптимизации уже сейчас – для пробы.

В Excel проблемы оптимизации решаются с помощью встроенного модуля под названием «Поиск решения».

• В Windows «Поиск решения» можно подключить, пройдя во вкладку «Файл» (в Excel 2007 это верхняя левая кнопка Windows) ? Параметры ? Надстройки. Нажав «Доступные надстройки» в выпадающем меню, отметьте «Поиск решения».

• В MacOS «Поиск решения» добавляется из меню «Инструменты», в котором следует выбрать «Надстройки», а затем Solver.xlam.

Кнопка «Поиск решения» появится в разделе «Анализ» вкладки «Данные» в любой версии Excel.

Отлично! Включив «Поиск решения», можете приступать к оптимизации. Представьте, что вам велели потреблять 2400 ккал в день. Какое минимальное количество покупок нужно сделать в вашем киоске, чтобы набрать дневную норму? Очевидно, самый простой выход – купить 10 сэндвич-мороженых по 240 ккал в каждом, но можно ли набрать норму, совершив меньше 10 покупок?

«Поиск решения» знает ответ!

Для начала сделайте копию листа «Calories»/«Калории», назовите его «Калории – решение» и удалите из копии все, кроме таблицы калорийности. Чтобы сделать копию листа в Excel, просто кликните правой клавишей мышки на заголовке листа, который хотите скопировать (внизу), и выберите в появившемся меню «Переместить» или «Копировать». Так вы получите новый лист, как на рис. 1-21.

Чтобы заставить «Поиск решения» искать решение, нужно задать ему пределы ячеек, в которых следует вести поиск. В нашем случае мы хотим узнать, сколько и чего нужно купить. Поэтому следующий за калорийностью столбец C назовите «Сколько?» (или как вам больше нравится) и разрешите «Поиску решения» хранить свои решения в нем.

Excel считает значения всех пустых ячеек равными нулю, так что вам не нужно заполнять этот столбец перед началом работы. «Поиск решения» сделает это за вас.

В ячейке C16 просуммируйте количество покупок таким образом:

=SUM(C2:C15) /

=СУММ(C2:C15)

Под данной формулой можно подсчитать количество килокалорий в этих покупках (которая должна, по вашему разумению, равняться 2400), используя формулу SUMPRODUCT/СУММПРОИЗВ:

=SUMPRODUCT(B2:B15,C2:C15) /

=СУММПРОИЗВ(B2:B15,C2:C15)

Таким образом получается лист, изображенный на рис. 1-22.

Теперь вы готовы к построению модели, так что запускайте «Поиск решения», нажав кнопку «Поиск решения» во вкладке «Данные».

Заметка

Окно поиска решений в Excel 2011, показанное на рис. 1-23, выглядит примерно так же, как и в Excel 2010 и 2013. В Excel 2007 интерфейс немного другой, но единственное существенное отличие заключается в отсутствии окна выбора алгоритма. Зато можно выбрать «Линейную модель» в параметрах поиска решений. Обо всех этих элементах мы поговорим позже.

Основные элементы, которые нужны «Поиску решения» для решения проблемы, как показано на рис. 1-23, – это ячейка для результата, направление оптимизации (минимализация или максимализация), несколько условных переменных, которые «Поиск решения» может изменять, и какие-либо условия.

Наша цель – минимизировать количество покупок в ячейке C16. Ячейки, значение которых может меняться, находятся в пределах C2:C15. Условие же состоит в том, что значение C17 – общего количества килокалорий – должно равняться 2400. Также нужно добавить условие, что результат должен быть положительным и целым – мы ведь считаем покупки в штуках, так что придется отметить галочкой опцию «Неотрицательные значения» в меню параметров поиска решения Excel 2007 и добавить целочисленность как условие решения. Так или иначе, мы не можем купить 1,7 бутылок газировки. (Всю глубину условия целочисленности вы познаете в главе 4).

Чтобы добавить условие общего количества килокалорий, нажмите «Добавить» и задайте ячейке C17 значение 2400, как показано на рис. 1-24.

Точно так же можно добавить условие целочисленности для C2:C15, как показано на рис. 1-25.

Нажмите OK.

В Excel 2010, 2011 и 2013 убедитесь, что метод решения установлен на «Поиск решения линейных задач симплекс-методом». Это наиболее подходящий для нашей задачи метод, так как она линейна. Под линейностью я подразумеваю, что для решения проблемы нужны только линейные комбинации значений из C2:C15 (суммы, произведения значений и констант количества килокалорий и т. д.).

Если бы в нашей модели встречались нелинейные комбинации (вроде квадратного корня из решения, логарифма или экспоненты), то мы могли бы

использовать какой-нибудь другой алгоритм, предлагаемый Excel. Подробно этот вариант рассматривается в главе 4.

В Excel 2007 обозначить задачу как линейную можно, нажав на «Линейную модель» внизу окна «Параметры поиска решений». В итоге должно получиться то, что изображено на рис. 1-26.

Отлично! Самое время нажать кнопку «Выполнить». Excel найдет решение практически мгновенно. Как явствует из рис. 1-27, результат равняется 5. Ваш Excel может выбрать какие-то другие 5 позиций, но их минимальное количество останется неизменным.

OpenSolver: хотелось бы обойтись без него, но это невозможно

Вообще эта книга писалась для работы исключительно со встроенным поиском решений Excel. Но по загадочным и необъяснимым причинам часть функционала была просто удалена из позднейших версий надстройки.

Это значит, что все описанное в книге работает для стандартного «Поиска решения» Excel 2007 и Excel 2011 на MacOS, но в Excel 2010 и 2013 встроенный поиск решения вдруг начинает жаловаться на то, что оптимизируемая линейная модель слишком велика для него (я заранее сообщу, какие модели из рассматриваемых в книге настолько сложны).

К счастью, существует превосходный бесплатный инструмент под названием OpenSolver, совместимый с версиями Excel для Windows, который восполняет этот недостаток. С ним можно строить модель в обычном интерфейсе «Поиска

решения», в который OpenSolver добавляет кнопку для использования симплекс-метода решения линейных задач, работающего буквально со скоростью света.

Для установки OpenSolver зайдите на http://opensolver.org (http://opensolver.org/) и загрузите оттуда архив. Распакуйте файл в папку и в любое время, когда понадобится решить «увесистую» модель, просто внесите ее в электронную таблицу и дважды кликните на файле opensolver.xlam, после чего во вкладке «Данные» появится новый раздел OpenSolver. Теперь нажмите на кнопку «Решить». Как показано на рис. 1-28, я применил OpenSolver в Excel 2013 к модели из предыдущего раздела, и он считает, что можно купить 5 кусков пиццы.

Подытожим

Вы научились быстро ориентироваться в Excel и выбирать области поиска, эффективно использовать абсолютные ссылки, пользоваться специальной вставкой, VLOOKUP/ВПР и другими функциями поиска ячейки, сортировкой и фильтрацией данных, создавать сводные таблицы и диаграммы, работать с формулами массива и поняли, как и когда прибегать к помощи «Поиска решения».

Но вот один грустный (или смешной, в зависимости от вашего нынешнего настроения) факт. Я знал консультантов по менеджменту в крупных компаниях, которые получали немаленькую зарплату за то, что я называю «двухшаговым консалтингом»/консультационным тустепом:

1. Разговор с клиентами обо всякой чепухе (о спорте, отпуске, барбекю… конечно, я не имею в виду, что жареное мясо – полная ерунда).

2. Сведение данных в Excel.

Вы можете не знать всего о школьном футболе (я определенно не знаю), но если вы усвоите эту главу, смело отправляйте второй пункт в нокаут.

Запомните: вы читаете эту книгу не затем, чтобы стать консультантом по менеджменту. Вы здесь для того, чтобы глубоко погрузиться в науку о данных. И это погружение произойдет буквально со следующей главы, которую мы начнем с небольшого неконтролируемого машинного самообучения.

2. Кластерный анализ, часть I: использование метода k-средних для сегментирования вашей клиентской базы

Я работаю в индустрии почтового маркетинга для сайта под названием MailChimp.com. Мы помогаем клиентам делать новостную рассылку для своей рекламной аудитории. Каждый раз, когда кто-нибудь называет нашу работу «почтовым вбросом», я чувствую на сердце неприятный холод.

Почему? Да потому что адреса электронной почты – больше не черные ящики, которые вы забрасываете сообщениями, будто гранатами. Нет, в почтовом маркетинге (как и в других формах онлайн-контакта, включая твиты, посты в Facebook и кампании на Pinterest) бизнес получает сведения о том, как аудитория вступает в контакт на индивидуальном уровне, с помощью отслеживания кликов, онлайн-заказов, распространения статусов в социальных сетях и т. д. Эти данные – не просто помехи. Они характеризуют вашу аудиторию. Но для непосвященного эти операции сродни премудростям греческого языка. Или эсперанто.

Как вы собираете данные об операциях с вашими клиентами (пользователями, подписчиками и т. д.) и используете ли их данные, чтобы лучше понять свою аудиторию? Когда вы имеете дело с множеством людей, трудно изучить каждого клиента в отдельности, особенно если все они по-разному связываются с вами. Даже если бы теоретически вы могли достучаться до каждого лично, на практике это вряд ли осуществимо.

Нужно взять клиентскую базу и найти золотую середину между «бомбардировкой» наобум и персонализированным маркетингом для каждого

отдельного покупателя. Один из способов достичь такого баланса – использование кластеризации для сегментирования рынка ваших клиентов, чтобы вы могли обращаться к разным сегментам вашей клиентской базы с различным целевым контентом, предложениями и т. д.

Кластерный анализ – это сбор различных объектов и разделение их на группы себе подобных. Работая с этими группами – определяя, что у их членов общего, а что отличает их друг от друга – вы можете многое узнать о беспорядочном имеющемся у вас массиве данных. Это знание поможет вам принимать оптимальные решения, причем на более детальном уровне, нежели раньше.

В этом разрезе кластеризация называется разведочной добычей данных, потому что эти техники помогают «вытянуть» информацию о связях в огромных наборах данных, которые не охватишь визуально. А обнаружение связей в социальных группах полезно в любой отрасли – для рекомендаций фильмов на основе привычек целевой аудитории, для определения криминальных центров города или обоснования финансовых вложений.

Одно из моих любимых применений кластеризации – это кластеризация изображений: сваливание в кучу файлов изображений, которые «выглядят одинаково» для компьютера. К примеру, в сервисах размещения изображений типа Flickr пользователи производят кучу контента и простая навигация становится невозможной из-за большого количества фотографий. Но, используя кластерные техники, вы можете объединять похожие изображения, позволяя пользователю ориентироваться между этими группами еще до подробной сортировки.

Контролируемое или неконтролируемое машинное обучение?

В разведочной добыче данных вы, по определению, не знаете раньше времени, что же за данные вы ищете. Вы – исследователь. Вы можете четко объяснить, когда двое клиентов выглядят похожими, а когда разными, но вы не знаете лучшего способа сегментировать свою клиентскую базу. Поэтому «просьба» к компьютеру сегментировать клиентскую базу за вас называется неконтролируемым машинным обучением, потому что вы ничего не контролируете – не диктуете компьютеру, как делать его работу.

В противоположность этому процессу, существует контролируемое машинное обучение, которое появляется, как правило, когда искусственный интеллект попадает на первую полосу. Если я знаю, что хочу разделить клиентов на две группы – скажем, «скорее всего купят» и «вряд ли купят» – и снабжаю компьютер историческими примерами таких покупателей, применяя все нововведения к одной из этих групп, то это контроль.

Если вместо этого я скажу: «Вот что я знаю о своих клиентах и вот как определить, разные они или одинаковые. Расскажи-ка что-нибудь интересненькое», – то это отсутствие контроля.

В данной главе рассматривается самый простой способ кластеризации под названием метод k-средних, который ведет свою историю из 50-х годов и с тех пор стал дежурным в открытии знаний из баз данных (ОЗБД) во всех отраслях и правительственных структурах.

Метод k-средних – не самый математически точный из всех методов. Он создан, в первую очередь, из соображений практичности и здравого смысла – как афроамериканская кухня. У нее нет такой шикарной родословной, как у французской, но и она зачастую угождает нашим гастрономическим капризам. Кластерный анализ с помощью k-средних, как вы вскоре убедитесь, – это отчасти математика, а отчасти – экскурс в историю (о прошлых событиях компании, если это сравнение относится к методам обучения менеджменту). Его несомненным преимуществом является интуитивная простота.

Посмотрим, как работает этот метод, на простом примере.

Девочки танцуют с девочками, парни чешут в затылке

Цель кластеризации методом k-средних – выбрать несколько точек в пространстве и превратить их в k группы (где k – любое выбранное вами число). Каждая группа определена точкой в центре вроде флага, воткнутого в Луну и сигнализирующего: «Эй, вот центр моей группы! Присоединяйтесь, если к этому флагу вы ближе, чем к остальным!» Этот центр группы (с официальным названием кластерный центроид) – то самое среднее из названия метода k-

средних.

Вспомним для примера школьные танцы. Если вы сумели стереть ужас этого «развлечения» из своей памяти, я очень извиняюсь за возвращение таких болезненных воспоминаний.

Герои нашего примера – ученики средней школы Макакне, пришедшие на танцевальный вечер под романтическим названием «Бал на дне морском», – рассеяны по актовому залу, как показано на рис. 2–1. Я даже подрисовал в Photoshop паркет, чтобы было легче представить ситуацию.

А вот примеры песен, под которые эти юные лидеры свободного мира будут неуклюже танцевать (если вдруг вам захочется музыкального сопровождения, к примеру, на Spotify):

• Styx: Come Sail Away

• Everything But the Girl: Missing

• Ace of Base: All that She Wants

• Soft Cell: Tainted Love

• Montell Jordan: This is How We Do It

• Eiffel 65: Blue

Теперь кластеризация по k-средним зависит от количества кластеров, на которое вы желаете поделить присутствующих. Давайте остановимся для начала на трех кластерах (далее в этой главе мы рассмотрим вопрос выбора k). Алгоритм размещает три флажка на полу актового зала некоторым допустимым образом, как показано на рис. 2–2, где вы видите 3 начальных флажка, распределенных по полу и отмеченных черными кружками.

Конец ознакомительного фрагмента.

notes

Сноски

1

Популярный тип столбчатых диаграмм (гистограмм), который используется для иллюстрации плана, графика работ по какому-либо проекту. Является одним из методов планирования проектов. – Прим. ред.

2

Искусственный интеллект. – Прим. пер.
