

УДК 330.47
ББК 65.051.03
Ф79

Переводчик А. Соколова
Редактор Л. Мамедова

Форман Дж.

Ф79 Много цифр: Анализ больших данных при помощи Excel / Джон Форман ; Пер. с англ. А. Соколовой. — М. : Альпина Пабlishер, 2016. — 461 с.

ISBN 978-5-9614-5032-3

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.

УДК 330.47
ББК 65.051.03

Все права защищены. Никакая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами, включая размещение в сети Интернет и в корпоративных сетях, а также запись в память ЭВМ для частного или публичного использования без письменного разрешения владельца авторских прав. По вопросу организации доступа к электронной библиотеке издательства обращайтесь по адресу tylib@alpina.ru

© John Wiley & Sons, Inc., Indianapolis, Indiana, 2014
All Rights Reserved. This translation published under license with the original publisher John Wiley & Sons, Inc.
© Издание на русском языке, перевод, оформление.
ООО «Альпина Пабlishер», 2016
© Фотография на обложке. Jason Travis /
Courtesy of John W. Foreman

ISBN 978-5-9614-5032-3 (рус.)
ISBN 978-1-118-66146-8 (англ.)

Содержание

Введение	11
1 Все, что вы жаждали знать об электронных таблицах, но боялись спросить	21
Немного данных для примера.....	22
Быстрый просмотр с помощью кнопок управления	23
Быстрое копирование формул и данных	24
Форматирование ячеек.....	25
Специальная вставка.....	27
Вставка диаграмм	28
Расположение меню поиска и замены.....	29
Формулы поиска и вывода величины.....	30
Использование VLOOKUP/ВПР для объединения данных	32
Фильтрация и сортировка	33
Использование сводных таблиц	37
Использование формул массива.....	40
Решение задач с помощью «Поиска решения».....	41
OpenSolver: хотелось бы обойтись без него, но это невозможно.....	46
Подытожим	47
2 Кластерный анализ, часть I: использование метода k-средних для сегментирования вашей клиентской базы	49
Девочки танцуют с девочками, парни чешут в затылке.....	51
Реальная жизнь: кластеризация методом k-средних в электронном маркетинге	56
Оптовая Винная Империя Джоуи Бэг О'Донатса	56
Исходный набор данных	57
Определяем предмет измерений	58

Начнем с четырех кластеров.....	61
Евклидово расстояние: измерение расстояний напрямик	62
Расстояния и принадлежность к кластеру для всех!	65
Поиск решений для кластерных центров	67
Смысл полученных результатов	70
Рейтинг сделок кластерным методом	71
Силуэт: хороший способ позволить разным значениям k посостязаться	75
Как насчет пяти кластеров?.....	82
Поиск решения для пяти кластеров	83
Рейтинг сделок для всех пяти кластеров.....	84
Вычисление силуэта кластеризации по пяти средним	87
K-медианная кластеризация и асимметрическое измерение расстояний.....	89
Использование k-медианной кластеризации	89
Переходим к соответствующему измерению расстояний.....	90
А теперь все то же самое, но в Excel.....	92
Рейтинг сделок для 5-медианных кластеров.....	94
Подытожим	98

3 Наивный байесовский классификатор

и неопишуемая легкость бытия идиотом	101
Называя продукт Mandrill, ждите помех вместе с сигналами	101
Самое быстрое в мире введение в теорию вероятности	104
Суммируем условную вероятность	104
Совместная вероятность, цепное правило и независимость	105
Что же с зависимыми событиями?.....	106
Правило Байеса	107
Использование правила Байеса для создания моделирования	108
Высококласные вероятности часто считаются равными.....	110
Еще немного деталей классификатора	111
Да начнется Excel-вечеринка!	113
Убираем лишнюю пунктуацию.....	113
Разное о пробелах	114
Подсчет жетонов и вычисление вероятностей.....	118
У нас есть модель! Воспользуемся ею	121
Подытожим	127

4 Оптимизационное моделирование:

этот «свежевыжатый апельсиновый сок» не смешает себя сам	129
Зачем ученым, работающим с данными, нужна оптимизация?	130
Начнем с простого компромисса.....	131
Представим проблему в виде политопа.....	132

Решение путем сдвигания линии уровня функции.....	134
Симплекс-метод: все по углам	135
Работа в Excel	137
Монстр в конце главы	147
Свежий, из сада — прямо в стакан...	
с небольшой остановкой на модель смешивания.....	148
Вы используете модель для смешивания	149
Начнем с характеристик.....	150
Возвращаемся к консистенции	151
Вводим данные в Excel	152
Постановка задачи «Поиску решения»	155
Снижаем стандарты	158
Удаление дохлых белок: правило минимакс	161
«Если... то» и ограничение «Большого М»	164
Еще больше переменных: добьем до 11	167
Моделируем риски	175
Нормальное распределение данных	176
Подытожим	184
5 Кластерный анализ, часть II: сетевые графы и определение сообществ....	187
Что такое сетевой граф?.....	188
Визуализируем простой граф	189
Краткое введение в Gephi.....	192
Установка Gephi и подготовка файлов	192
Визуализация графа	194
Степень вершины	197
Приятная картинка	200
Прикосновение к данным графа	200
Строим граф из данных об оптовой торговле вином	202
Создание матрицы близости косинусов.....	204
Построение графа N-соседства	207
Числовое значение ребра: очки и штрафные в модулярности графа.....	212
Кто же такие «очки» и «штрафные»?	212
Подготовка к итоговому подсчету.....	216
Переходим к кластеризации!	219
Деление 1	219
Деление 2: электролатино!	225
И... деление 3: возмездие	227
Кодируем и анализируем группы	228
Туда и обратно: история Gephi	233
Подытожим	238

6	Бабушка контролируемого искусственного интеллекта — регрессия	241
	Погоди, ты что — беременна?	241
	Не обольщайтесь!	242
	Определение беременных покупателей РитейлМарта с помощью линейной регрессии	243
	Набор отличительных признаков	244
	Сборка обучающих данных	245
	Создание фиктивных переменных	247
	Мы сделаем свою собственную линейную регрессию!	250
	Статистика линейной регрессии: R-квадрат, критерии Фишера и Стьюдента	259
	Делаем прогнозы на основании новых данных и измеряем результат	270
	Предсказание беременных покупателей РитейлМарта с помощью логистической регрессии	281
	Первое, что нам нужно — это функция связи	281
	Присоединение логистической функции и реоптимизация	282
	Создание настоящей логистической регрессии	286
	Выбор модели: сравнение работы линейной и логистической регрессий	287
	Дополнительная информация	291
	Подытожим	292
7	Комплексные модели: огромная куча ужасной пиццы	293
	Используем данные из главы 6	294
	Бэггинг: перемешать, обучить, повторить	296
	Одноуровневое дерево решений — неудачное название «неумного» определителя	296
	А мне не кажется, что это глупо!	297
	Нужно еще сильнее!	300
	Обучим же ее!	300
	Оценка бэггинговой модели	310
	Бустинг: если сразу не получилось, бустингуйте и пробуйте снова	315
	Обучаем модель: каждому признаку — шанс	315
	Оценка модели бустинга	324
	Подытожим	327
8	Прогнозирование: дышите ровно, выиграть невозможно	329
	Торговля мечами начата	330
	Знакомство с временной последовательностью данных	331
	Медленный старт с простым экспоненциальным сглаживанием	333
	Настраиваем прогноз простого экспоненциального сглаживания	335

Возможно, у вас есть тренд	341
Экспоненциальное сглаживание Холта с корректировкой тренда	344
Настройка холтовского сглаживания с коррекцией тренда в электронной таблице ...	346
Мультипликативное экспоненциальное сглаживание Холта–Винтерса	360
Установка исходных значений уровня, тренда и сезонности	362
Приступим к прогнозу	367
И наконец... оптимизация!	372
Пожалуйста, скажите, что это все!!!	373
Создаем интервал прогнозирования вокруг прогноза	374
И диаграмма с областями для пушкого эффекта	378
Подытожим	381
9 Определение выбросов: выделяющиеся не значит важные	383
Выбросы тоже (плохие?) люди!	384
Захватывающее дело Хадлум против Хадлум	384
Границы Тьюки	386
Применение границ Тьюки в таблице	386
Ограничения этого нехитрого метода	388
Ни в чем не ужасен, плох во всем	390
Подготовка данных к отображению на графе	391
Создаем граф	394
Вычисляем k ближайших соседей	397
Определение выбросов на графе, метод 1: полустепень захода	398
Определение выбросов на графе, метод 2: нюансы k-расстояния	401
Определение выбросов на графе, метод 3: факторы локальных выбросов – это то, что надо	403
Подытожим	409
10 Переходим от таблиц к программированию	411
Налаживаем контакт с R	412
Пошевелим пальцами	413
Чтение данных в R	421
Настоящая научная работа с данными	423
Сферическое k-среднее винных данных в нескольких линиях	423
Построение моделей ИИ для данных о беременных	430
Прогнозирование в R	439
Определение выбросов	443
Подытожим	448
Заключение	451
Благодарности	459