

داده‌کاوی برای تحلیل خودکار کسب و کار

مفاهیم، فنون و کاربردها در R

گالیت شمولی، پیتر سی. بروس، اینبال یاهاو، نیتین آر. پاتل،

کنت سی لیختندال

www.ketab.ir

ترجمه

محمد رضا فقیهی حبیب آبادی





۷۸۳

مرکز چاپ و انتشارات دانشگاه شهید بهشتی

داده‌کاوی برای تحلیل خودکار کسب‌وکار: مفاهیم، فنون و کاربردهای R
گالیت شمولی، پیتروسی بروس، اینبال یاهاو، نیتین آر. پاتل، کنت سی لیختندال
ترجمه محمدرضا فقیهی حبیب‌آبادی

Galit Shmueli, Peter C. Bruce, Inbal Yahav, Nitin R. Patel, Kenneth C. Lichtendahl, Jr., *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*, John Wiley & Sons, Inc., 2018.

ویراستار: محمود ایرانی فرد
صفحه‌آرا: سمانه تات
طراح جلد: امیرشاهرخ فریوسفی
ناظر چاپ: نوید سیفان
چاپ: ۱۴۰۰
شمارگان: ۵۰۰
قیمت: ۲,۲۵۰,۰۰۰ ریال

کلیه حقوق برای دانشگاه شهید بهشتی محفوظ است.

عنوان و نام پدیدآور:	داده‌کاوی برای تحلیل خودکار کسب‌وکار: مفاهیم، فنون و کاربردهای R / گالیت شمولی... و دیگران؛ ترجمه محمدرضا فقیهی حبیب‌آبادی؛ ویراستار محمود ایرانی فرد. تهران: دانشگاه شهید بهشتی، مرکز چاپ و انتشارات، ۱۴۰۰. بیست و دو، ۷۳۲ ص: مصوره جدول، نمودار. دانشگاه شهید بهشتی، مرکز چاپ و انتشارات؛ ۷۸۳. ۹۷۸ ۹۶۴ ۴۵۷ ۵۲۰ ۴
مشخصات نشر:	فبا
مشخصات ظاهری:	عنوان اصلی: Data mining for business analytics: concepts, techniques, and ... 2018
فروست:	نویسندگان گالیت شمولی، پیتروسی بروس، اینبال یاهاو، نیتین آر پاتل، کنت سی لیختندال
شابک:	کسب و کار -- داده‌پرداز، Business - Data processing
وضعیت فهرست‌نویسی:	داده‌کاوی، Data mining، آر (زبان برنامه‌نویسی کامپیوتر)، R (Computer program language)
یادداشت:	ریاضیات بازرگانی -- برنامه‌های کامپیوتری، Business mathematics - Computer programs
یادداشت:	شمولی، گالیت، ۱۹۷۱-م.؛ Shmueli, Galit, 1971-
موضوع:	فقیهی حبیب‌آبادی، محمدرضا، ۱۳۳۵- مترجم
موضوع:	دانشگاه شهید بهشتی، مرکز چاپ و انتشارات
موضوع:	Shahid Beheshti University. Printing and Publishing Center
شناسه افزوده:	HF۵۵۴۸/۲
شناسه افزوده:	۶۵۸/۴
شناسه افزوده:	۷۵۶۹۶۹۲
شناسه افزوده:	
رده‌بندی کنگره:	
رده‌بندی دیویی:	
شماره کتاب‌شناسی ملی:	

کد ناشر ۱۰۰۱۷۳۴

www.pub.sbu.ac.ir
unipress@mail.sbu.ac.ir

فهرست مطالب

سیزده	پیشگفتار مترجم
پانزده	پیشگفتار گرت جیمز
هفده	پیشگفتار راوی بابنا
نوزده	پیشگفتار ویراست R

بخش اول مقدمات

۱

۳	۱ مقدمه
۳	۱.۱ تحلیل خودی چیست؟
۶	۲.۱ داده‌کاوی چیست؟
۶	۳.۱ داده‌کاوی و واژه‌های مرتبط
۸	۴.۱ داده‌های بزرگ
۱۰	۵.۱ علم داده‌ها
۱۰	۶.۱ چرا روش‌های متفاوت بسیار زیادی وجود دارند؟
۱۱	۷.۱ واژه‌شناسی و نمادها
۱۴	۸.۱ نقشه راه این کتاب

۲ نگاهی کلی به فرایند داده‌کاوی

۱۹

۱۹	۱.۲ مقدمه
۲۰	۲.۲ ایده‌های اصلی در داده‌کاوی
۲۴	۳.۲ گام‌های داده‌کاوی
۲۷	۴.۲ گام‌های مقدماتی
۲۳	۵.۲ توان پیشگویی و پیش‌برازش
۵۱	۶.۲ مساحت یک مدل پیشگویی
۵۶	۷.۲ استفاده از R برای داده‌کاوی روی یک ماشین در محل

۸.۲. خودکارسازی راه‌حل‌های داده‌کاوی ۵۷

بخش دوم اکتشاف داده‌ها و کاهش بُعد ۷۱

۷۳	۳ تصویرسازی داده‌ها
۷۳	۱.۳. فواید تصویرسازی
۷۶	۲.۳. مثال‌های داده‌ها
	۳.۳. نمودارهای پایه: نمودارهای میله‌ای، نمودارهای خطی، و نمودارهای
۷۸	پراکنش
۸۸	۴.۳. تصویرسازی چندبُعدی
۱۰۶	۵.۳. تصویرسازی‌های خاص
	۶.۳. خلاصهٔ مطالب: تصویرسازی‌ها و عملیات‌های عمده، در ارتباط
۱۱۱	با هدف داده‌کاوی

۱۲۱	۴ کاهش بُعد
۱۲۱	۱.۴. مقدمه
۱۲۲	۲.۴. مشق بُعدچندی
۱۲۳	۳.۴. ملاحظات عملی
۱۲۳	۴.۴. خلاصهٔ داده‌ها
۱۳۰	۵.۴. تحلیل همبستگی
۱۳۲	۶.۴. کاهش تعداد رسته‌ها در متغیرهای رسته‌ای
۱۳۳	۷.۴. تبدیل یک متغیر رسته‌ای به متغیری پیوسته
۱۳۴	۸.۴. تحلیل مؤلفه‌های اصلی
۱۴۷	۹.۴. کاهش بُعد با استفاده از مدل‌های رگرسیون
۱۴۷	۱۰.۴. کاهش بُعد با استفاده از درختان رده‌بندی و رگرسیون

بخش سوم ارزیابی عملکرد ۱۵۳

۱۵۵	۵ ارزیابی عملکرد پیشگویانه
-----	----------------------------

۱۵۶	مقدمه	۱.۵
۱۵۶	ارزیابی عملکرد پیشگویانه	۲.۵
۱۶۳	قضاوت در مورد عملکرد رده‌بندی‌کننده	۳.۵
۱۸۰	قضاوت در مورد عملکرد رتبه‌بندی	۴.۵
۱۸۶	بیش‌نمونه‌گیری	۵.۵

بخش چهارم روش‌های پیشگویی و رده‌بندی ۱۹۹

۲۰۱	۶ رگرسیون خطی چندگانه	
۲۰۱	مقدمه	۱.۶
۲۰۲	مدل‌بندی پیشگو در برابر مدل‌بندی تبیینی	۲.۶
۲۰۴	برآورد معادله رگرسیون و پیشگویی	۳.۶
۲۰۹	انتخاب متغیر در رگرسیون خطی	۴.۶
۲۲۹	۷ k -نزدیک‌ترین همسایه (k -NN)	
۲۲۹	رده‌بندی‌کننده k -NN (برآمد رده‌بندی)	۱.۷
۲۳۹	k -NN برای یک برآمد عددی	۲.۷
۲۴۲	مزایا و معایب الگوریتم‌های k -NN	۳.۷
۲۴۷	۸ رده‌بندی‌کننده بیز ناپخته	
۲۴۷	مقدمه	۱.۸
۲۵۰	به‌کارگیری رده‌بندی‌کننده بیزی کامل (دقیق)	۲.۸
۲۶۱	مزایا و معایب رده‌بندی‌کننده بیز ناپخته	۳.۸
۲۷۱	۹ درخت‌های رده‌بندی و رگرسیون	
۲۷۲	مقدمه	۱.۹
۲۷۳	درختان رده‌بندی	۲.۹
۲۸۳	ارزیابی عملکرد یک درخت رده‌بندی	۳.۹
۲۸۷	پرهیز از بیش‌برازش	۴.۹
۲۹۴	قواعد رده‌بندی از درختان	۵.۹

- ۶.۹. درختان رده‌بندی برای بیش از دو رده ۲۹۷
- ۷.۹. درختان رگرسیون ۲۹۸
- ۸.۹. بهبودبخشیدن پیشگویی: جنگل‌های تصادفی و درخت‌های تقویت‌شده ۳۰۰
- ۹.۹. مزایا و ضعف‌های یک درخت ۳۰۳

۱۰ رگرسیون لوژستیک ۳۱۱

- ۱.۱۰. مقدمه ۳۱۲
- ۲.۱۰. مدل رگرسیون لوژستیک ۳۱۳
- ۳.۱۰. مثال: پذیرش وام شخصی ۳۱۵
- ۴.۱۰. ارزیابی عملکرد رده‌بندی ۳۲۴
- ۵.۱۰. مثالی از یک تحلیل کامل: پیشگویی تأخیر پروازها ۳۲۸
- ۶.۱۰. پیوست: رگرسیون لوژستیک برای نیم‌رخ‌سازی ۳۴۰

۱۱ شبکه‌های عصبی ۳۵۷

- ۱.۱۱. مقدمه ۳۵۷
- ۲.۱۱. مفهوم و ساختار یک شبکه عصبی ۳۵۸
- ۳.۱۱. برازش یک شبکه به داده‌ها ۳۵۹
- ۴.۱۱. ورودی الزامی کاربر ۳۷۶
- ۵.۱۱. اکتشاف ارتباط بین پیشگوها و برآمد ۳۷۷
- ۶.۱۱. مزیت‌ها و ضعف‌های شبکه‌های عصبی ۳۷۹

۱۲ تحلیل تشخیصی ۳۸۵

- ۱.۱۲. مقدمه ۳۸۵
- ۲.۱۲. فاصله یک ثبت از یک رده ۳۸۷
- ۳.۱۲. توابع رده‌بندی خطی فیشر ۳۹۰
- ۴.۱۲. عملکرد رده‌بندی تحلیل تشخیصی ۳۹۴
- ۵.۱۲. احتمال‌های پیشین ۳۹۷
- ۶.۱۲. هزینه‌های بدرده‌بندی نامساوی ۳۹۷
- ۷.۱۲. رده‌بندی بیش از دو رده ۳۹۸
- ۸.۱۲. مزیت‌ها و ضعف‌ها ۴۰۱

در اواخر قرن بیستم، با افزایش سرعت محاسبات رایانه‌ای، حجم داده‌ها رو به فزونی نهاد و در نتیجه تحلیل داده‌ها و به‌ویژه استخراج الگوهای پنهان در آنها اهمیت بسیاری پیدا کرد. در آن زمان، نیاز به کشف و استخراج این الگوها موجب تأسیس رشته داده‌کاوی شد که در سازمان‌ها و شرکت‌های بزرگ گوناگون جهان به‌کار گرفته شد. رشد نمایي حجم داده‌ها و سرعت رو به فزونی پردازش آنها به‌ویژه در دهه دوم قرن بیست و یکم، رشته تخصصی علوم داده‌ها را شکل داده است که در آن علوم آمار، ریاضی، علوم و مهندسی رایانه نقش دارند. امروزه با اینکه رشته علوم داده‌ها در جهان در مرکز توجهات مدیران و مراکز دانشگاهی قرار گرفته است، اما هسته اصلی کار یک متخصص علوم داده‌ها، استخراج دانایی از داده‌هاست، که اغلب با فنون داده‌کاوی انجام می‌شود.

از آنجا که داده‌کاوی از دو تخصص آمار و علوم رایانه بهره می‌برد و بدون توجه به هر دو موضوع نمی‌توان حداکثر استفاده را از آن برد، وجود منبعی که با عنایت به هر دو موضوع، خواننده را به اجرای درست طرح‌های داده‌کاوی رهنمون سازد، بسیار بااهمیت است. کتاب‌های بسیاری در جهان در زمینه داده‌کاوی منتشر شده‌اند که غالب آنها از جامعیت و رویکرد مناسبی برخوردار نیستند. در عین حال وضعیت در منابع فارسی به مراتب بدتر است زیرا بررسی‌ها نشان می‌دهد که معدود کتاب‌های موجود فارسی، هیچ‌کدام قابلیت استفاده در آموزش دانشگاهی را ندارند. به این دلیل کتاب حاضر که ویژگی‌های خوب پُرشماری دارد برای ترجمه انتخاب شد. لازم به یادآوری است که داده‌کاوی در بسیاری از حوزه‌ها به‌کار گرفته می‌شود، یکی از مهم‌ترین آنها تحلیل خودکار کسب و کار (Business Analytics) است که در هر سازمان، نهاد و شرکتی کاربرد دارد. این کتاب، داده‌کاوی را برای تحلیل خودکار کسب و کار تشریح می‌کند، اما خواننده با درک مفاهیم و فنون آن قادر خواهد بود در همه زمینه‌ها داده‌کاوی را به‌کار گیرد.

این کتاب برای تمامی دروس داده‌کاوی در مقاطع کارشناسی و کارشناسی ارشد

می‌تواند استفاده شود. برای درس داده‌کاوی مقطع کارشناسی فصول ۱ تا ۹، ۷ تا ۱۱ و ۱۵ گزینه‌های مناسبی‌اند، اما براساس رشته تحصیلی دانشجویان، برای درس داده‌کاوی در مقاطع گوناگون تحصیلی، مجموعه فصول دیگری می‌تواند برای تدریس انتخاب شود.

در تهیه کتاب بر خود لازم می‌دانم از زحمات خانم سمانه تات در اعمال نظرات ویراستار محترم و همچنین صفحه‌آرایی کتاب قدردانی کنم. همچنین از ویراستار کتاب آقای ایرانی که به‌طور قطع نظراتشان موجب بهتر شدن کتاب شد، تشکر می‌کنم. از خوانندگان این کتاب تقاضا می‌شود هرگونه خطا یا ابهامی را به اطلاع مترجم برسانند تا در ویراست‌های بعدی از آنها استفاده شود. در پایان، امیدوارم عرضه این کتاب به متخصصان کشور عزیزمان ایران، بتواند به استفاده درست از فنون داده‌کاوی کمک کند.

محمدرضا فقیهی حبیب‌آبادی

عضو هیئت علمی دانشگاه شهید بهشتی

بهار ۱۴۰۰

www.ketab.ir