

داده‌کاوی برای تحلیل خودکار کسب و کار
مفاهیم، فنون و کاربردها در پایتون

گالیت شمولی، پیتر سی. بروس، پیتر گِدک، نیتین آر. پاتل

www.ketab.ir

ترجمه

محمد رضا فقیهی حبیب آبادی
سمانه تات



۱۴۰۲



۸۷۲

مرکز چاپ و انتشارات دانشگاه شهید بهشتی

داده‌کاوی برای تحلیل خودکار کسب و کار: مفاهیم، فنون و کاربردها در پایتون

گالیت شمولی، پیتر سی. بروس، پیتر گدک، نیتین آر. پاتل

Galit Shmueli, Peter C. Bruce, Peter Gedeck, Nitin R. Patel, *Data Mining for Business Analytics: Concepts, Techniques, and Applications in Python*, John Wiley & Sons, Inc., 2020.

ترجمه دکتر محمدرضا فقیهی حبیب‌آبادی و دکتر سمانه تات

ویراستار: محمود ایرانی فرد

حروف‌نگار و صفحه‌آرا: سمانه تات

ناظر چاپ: نوید سیفان

چاپ: ۱۴۰۲

شمارگان: ۲۰۰

قیمت: ۳,۹۷۰,۰۰۰ ریال

کلیه حقوق برای دانشگاه شهید بهشتی محفوظ است.

«قیمت تمام کتاب به‌عهده مترجمان است.»

عنوان و نام پدید آور: داده‌کاوی برای تحلیل خودکار کسب و کار: مفاهیم، فنون و کاربردها در پایتون / گالیت شمولی... او دیگران... ترجمه محمدرضا فقیهی حبیب‌آبادی، سمانه تات؛ ویراستار محمود ایرانی فرد.

تهران: دانشگاه شهید بهشتی، مرکز چاپ و انتشارات، ۱۴۰۲.

بیست و پنج، ۷۶۹ ص.

مرکز چاپ و انتشارات دانشگاه شهید بهشتی؛

۹۷۸-۹۶۴-۴۵۷-۵۹۸-۳

نویسندگان گالیت شمولی، پیتر سی. بروس، پیتر گدک، نیتین آر پاتل

عنوان اصلی: *Data mining for business analytics: concepts, techniques and applications in Python*, c 2020.

کتاب حاضر با عنوان "داده‌کاوی با Python" با ترجمه مهدی اسماعیلی، سیدمهدی وحیدی‌پور توسط انتشارات آتی‌نگر و وینا در سال ۱۴۰۰ منتشر شده است.

موضوع: ریاضیات بازرگانی -- برنامه‌های کامپیوتری، - Business mathematics Computer programs

کسب و کار -- داده‌پردازی، Business - Data processing

موضوع: داده‌کاوی، Data mining: پایتون (زبان برنامه‌نویسی کامپیوتر)، Python (Computer program language)

شمولی، گالیت، ۱۹۷۱-م.؛ Shmueli, Galit, 1971-

فقیهی حبیب‌آبادی، محمدرضا، ۱۳۳۶- مترجم: تات، سمانه، ۱۳۶۵- مترجم

دانشگاه شهید بهشتی، مرکز چاپ و انتشارات

Shahid Beheshti University. Printing and Publishing Center

HF554A2

۶۵۰/۰۷۲۷

۹۴۱۱۷۹۹

مشخصات نشر:

مشخصات ظاهری:

فروست:

شابک:

یادداشت:

یادداشت:

یادداشت:

موضوع:

موضوع:

موضوع:

شناسه افزوده:

شناسه افزوده:

شناسه افزوده:

شناسه افزوده:

رده‌بندی کنگره:

رده‌بندی دیویی:

شماره کتاب‌شناسی ملی:

کد ناشر ۱۰۰۱۷۳۴

unipress.sbu.ac.ir; press.sbu.ac.ir
unipress@mail.sbu.ac.ir

فهرست مطالب

هفت	پیشگفتار مترجمان
نه	پیشگفتار گرت جیمز
یازده	پیشگفتار راوی باپنا
سیزده	پیشگفتار ویراست پایتون
هفده	قدردانی

بخش اول مقدمات

۱	مقدمه
۳	۱.۱ تحلیل خودکار کسب و کار چیست؟
۳	۲.۱ داده‌کاوی چیست؟
۶	۳.۱ داده‌کاوی و واژه‌های مرتبط
۶	۴.۱ داده‌های بزرگ
۸	۵.۱ علم داده‌ها
۱۰	۶.۱ چرا روش‌های متفاوت بسیار زیادی وجود دارند؟
۱۰	۷.۱ واژه‌شناسی و نمادها
۱۱	۸.۱ نقشه راه کتاب
۱۴	۲ نگاهی کلی به فرایند داده‌کاوی
۱۹	۱.۲ مقدمه
۱۹	۲.۲ ایده‌های اصلی در داده‌کاوی
۲۰	۳.۲ گام‌های داده‌کاوی
۲۴	۴.۲ گام‌های مقدماتی

۳۰۲	ارزیابی عملکرد یک درخت رده‌بندی	۳۰۹
۳۰۸	پرهیز از بیش‌برازش	۴۰۹
۳۱۶	قواعد رده‌بندی از درختان	۵۰۹
۳۱۷	درختان رده‌بندی برای بیش از دو رده	۶۰۹
۳۱۷	درختان رگرسیون	۷۰۹
۳۲۲	بهبودبخشیدن پیشگویی: جنگل‌های تصادفی و درخت‌های تقویت‌شده	۸۰۹
۳۲۵	مزایا و ضعف‌های یک درخت	۹۰۹

۱۰ رگرسیون لوژستیک

۳۳۵		۳۳۵
۳۳۶	مقدمه	۱۰۱۰
۳۳۸	مدل رگرسیون لوژستیک	۲۰۱۰
۳۴۱	مثال: پذیرش وام شخصی	۳۰۱۰
۳۴۸	ارزیابی عملکرد رده‌بندی	۴۰۱۰
۳۵۲	رگرسیون لوژستیک برای چندرده‌ای	۵۰۱۰
۳۵۸	مثالی از یک تحلیل کامل: پیشگویی تأخیر پروازها	۶۰۱۰
۳۷۲	پیوست: استفاده از Statmodels	۷۰۱۰

۱۱ شبکه‌های عصبی

۳۷۹		۳۷۹
۳۸۰	مقدمه	۱۰۱۱
۳۸۱	مفهوم و ساختار یک شبکه عصبی	۲۰۱۱
۳۸۱	برازش یک شبکه به داده‌ها	۳۰۱۱
۳۹۷	ورودی الزامی کاربر	۴۰۱۱
۳۹۸	اکتشاف ارتباط بین پیشگوها و برآمد	۵۰۱۱
۳۹۹	یادگیری عمیق	۶۰۱۱
۴۰۶	مزایا و معایب شبکه‌های عصبی	۷۰۱۱

۱۲ تحلیل تشخیصی

۴۱۱		۴۱۱
۴۱۲	مقدمه	۱۰۱۲
۴۱۴	فاصله یک ثبت از یک رده	۲۰۱۲
۴۱۷	توابع رده‌بندی خطی فیشر	۳۰۱۲

۴۲۱. عملکرد رده‌بندی تحلیل تشخیصی
۴۲۳. احتمال‌های پیشین
۴۲۳. هزینه‌های بدرده‌بندی نامساوی
۴۲۴. رده‌بندی بیش از دو رده
۴۲۸. مزایا و معایب

۴۳۵ روش‌های تلفیقی: کلیت‌ها و مدل‌سازی تعالی بخش ۱۳

۴۳۶. کلیت‌ها ۱.۱۳
۴۴۴. مدل‌سازی تعالی بخش (اقتاعی) ۲.۱۳
۴۵۲. خلاصه ۳.۱۳

۴۵۹ بخش پنجم کاوش ارتباط‌ها در میان ثب‌ها

- ۴۶۱ قواعد پیوند و پالایش گروهی ۱۴
۴۶۲. قواعد پیوند ۱.۱۴
۴۷۹. پالایش گروهی ۲.۱۴
۴۹۲. خلاصه ۳.۱۴

۵۰۱ تحلیل خوشه‌ای ۱۵

۵۰۲. مقدمه ۱.۱۵
۵۰۶. اندازه‌گیری فاصله بین دو ثب ۲.۱۵
۵۱۳. اندازه‌گیری فاصله بین دو خوشه ۳.۱۵
۵۱۶. خوشه‌بندی سلسله‌مراتبی (تراکمی) ۴.۱۵
۵۲۶. خوشه‌بندی غیرسلسله‌مراتبی: الگوریتم k - میانگین ۵.۱۵

۵۳۹ بخش ششم پیش‌بینی سری‌های زمانی

- ۵۴۱ نحوه انجام سری زمانی ۱۶
۵۴۲. مقدمه ۱.۱۶

۵۴۳. مدل‌سازی توصیفی در مقابل مدل‌سازی پیشگویانه
۵۴۴. روش‌های متداول پیش‌بینی در کسب و کار
۵۴۵. مؤلفه‌های سری زمانی
۵۵۱. افراز داده‌ها و ارزیابی عملکرد

۱۷ پیش‌بینی رگرسیون مینا ۵۶۱

۵۶۲. یک مدل با روند
۵۷۰. یک مدل با فصلی بودن
۵۷۴. یک مدل با روند و فصلی بودن
۵۷۵. خودهمبستگی و مدل‌های آریما

۱۸ روش‌های هموارسازی ۶۰۱

۶۰۲. مقدمه
۶۰۲. میانگین متحرک
۶۰۹. هموارسازی نمایی ساده
۶۱۱. هموارسازی نمایی پیشرفته

بخش هفتم تحلیل خودکار داده‌ها ۶۲۹

۱۹ تحلیل خودکار شبکه اجتماعی ۶۳۱

۶۳۲. مقدمه
۶۳۳. شبکه‌های سودار و بی‌سو
۶۳۵. تصویری سازی و تحلیل شبکه‌ها
۶۴۰. سنج‌های داده‌های اجتماعی و تاکسونومی
۶۴۵. استفاده از سنج‌های شبکه در پیشگویی و رده‌بندی
۶۵۴. گردآوری داده‌های شبکه اجتماعی با پایتون
۶۵۵. مزایا و معایب

۲۰ متن‌کاوی ۶۵۹

۶۶۰. مقدمه

۶۶۱	۲.۲۰. نمایش جدولی متن: ماتریس واژه-سند و «کیسه کلمه‌ها»
۶۶۱	۳.۲۰. کیسه کلمه‌ها در مقابل استخراج معنی در سطح سند
۶۶۳	۴.۲۰. پیش‌پردازش متن
۶۷۳	۵.۲۰. پیاده‌سازی روش‌های داده‌کاوی
۶۷۳	۶.۲۰. مثال: بحث‌های برخط درباره خودرو و الکترونیک
۶۷۸	۷.۲۰. خلاصه

۶۸۳	۲۱ موارد
۶۸۳	۱.۲۱. کانون کتاب چارلز
۶۹۱	۲.۲۱. اعتبار آلمان
۶۹۶	۳.۲۱. فروشنده کاتالوگی نرم‌افزار تایکو
۷۰۰	۴.۲۱. اقناع سیاسی
۷۰۵	۵.۲۱. لغو درخواست‌های تاکسی
۷۰۷	۶.۲۱. بخش بندی مصرف‌کنندگان صابون حمام
۷۱۱	۷.۲۱. جمع‌آوری اعلیه با دست مستقیم
۷۱۴	۸.۲۱. فروش متقابل کاتالوگی
۷۱۶	۹.۲۱. پیشگویی ورشکستگی
۷۱۹	۱۰.۲۱. مورد سری زمانی: پیش‌بینی تقاضای حمل‌ونقل عمومی

۷۲۳	منابع
۷۲۷	واژه‌نامه فارسی-انگلیسی
۷۳۹	واژه‌نامه انگلیسی-فارسی
۷۴۷	نام‌نامه
۷۵۱	نمایه

پیشگفتار مترجمان

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 یافت. در آن زمان، نیاز به کشف و استخراج این الگوها موجب تأسیس رشته داده‌کاوی و به‌کارگیری آن در سازمان‌ها و شرکت‌های بزرگ جهان شد. رشد نمایی 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 کامپیوتر نقش دارند. امروزه با اینکه رشته علوم داده‌ها در جهان در مرکز توجه مدیران و مراکز دانشگاهی قرار گرفته اما هسته اصلی کار یک متخصص علوم داده‌ها، استخراج دانایی از داده‌هاست که اغلب با فنون داده‌کاری انجام می‌شود.

از آنجاکه داده‌کاوی از دو تخصص آمار و علوم کامپیوتر بهره می‌برد و بدون توجه به این دو نمی‌توان حداکثر استفاده را از آن برد، وجود آمیگی 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، اما غالب آن‌ها از جامعیت و رویکرد مناسبی برخوردار نیستند. وضعیت در منابع فارسی به مراتب بدتر است، زیرا بررسی‌ها نشان می‌دهد محدود کتاب‌های موجود فارسی، قابلیت استفاده در آموزش دانشگاهی را ندارند. به این دلیل، کتاب حاضر که ویژگی‌های خوب پُرشماری دارد پس از نسخه مربوط به R برای ترجمه انتخاب شد. با توجه به محبوبیت روزافزون پایتون در میان متخصصان تجزیه و تحلیل، این کتاب می‌تواند پاسخگوی نیاز کسانی باشد که زبان برنامه‌نویسی پایتون را به زبان R ترجیح می‌دهند یا نیاز هم‌زمان به هر دوی آن‌ها دارند.

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

قوی و تعامل بی‌دردسر با چند بسته محاسباتی، نیاز دارند. بیشتر متون آماری، بدون تأکید زیاد بر کاربردهای عملی، جدای از کسب و کار، بر آموزش مجرد با روش‌های کلاسیک تمرکز دارند.

این کتاب دارای جامع‌ترین مرور روش‌های تحلیل خودکار کسب و کار است که من تا حالا دیده‌ام، همه چیز را پوشش می‌دهد، از رویکردهای کلاسیک مانند رگرسیون خطی و لوژستیک، تا روش‌های مدرن مثل شبکه‌های عصبی، بسته‌بندی و تقویت‌کردن و حتی شیوه‌های مختص کسب و کار مانند تحلیل شبکه‌های اجتماعی و متن‌کاوی. این کتاب اگر نگویم انجیل، حداقل راهنمای قطعی این موضوع است. در عین حال درست به اندازه فهرست مطالب، اینکه همگی به سبکی کاربردی و با استفاده از کاربردهای کسب و کار ارائه شده‌اند اهمیت دارد. در واقع آخرین فصل کاملاً به ۱۰ مورد جداگانه که رویکردهای تحلیل خودکار کسب و کار می‌تواند به‌کار گرفته شود، اختصاص یافته است.

در این آخرین ویرایش، نویسندگان پشتیبانی از پایتون، یک زبان برنامه‌نویسی که به سرعت در میان دانشمندان داده محبوبیت یافته است، اضافه کرده‌اند. کتاب توضیحات دقیق و کدهای مربوط به برنامه‌های کاربردی پایتون را در بسیاری از تنظیمات تجاری ارائه داده و اطمینان می‌دهد که خواننده واقعاً بتواند دانش خود را برای حل مشکلات زندگی واقعی اعمال کند. من مطمئن‌ام این کتاب یک ابزار ضروری برای هر دوره تحلیل تجاری با استفاده از پایتون خواهد بود.

ما اخیراً یک درس تحلیل خودکار کسب و کار برای دروس الزامی برنامه درسی MBA معرفی کردیم و من قصد دارم از این کتاب برای تدوین سرفصل، استفاده فراوان کنم و مطمئن‌ام که ابزاری صرف‌نظر نکردنی برای چنین درس‌هایی خواهد بود.

گرت جیمز

دانشکده کسب و کار مارشال، دانشگاه کالیفرنیا جنوبی، ۲۰۱۹