خلاصه کتاب

هوش مصنوعی برای تحلیلگری کسب و کار: الگوریتم‌ها، پلتفرم‌ها و سناریوهای کاربردی

Artificial Intelligence for Business Analytics

Algorithms, Platforms and Application Scenarios

تهیه شده توسط گروه تحلیل‌گری عظیم‌داده و کسب و کار

https://bdbanalytics.ir


بسم الله

تیم تهیه خلاصه کتاب

گروه تحلیل‌گری عظیم‌داده و کسب و کار

سرپرست گروه:

دکتر سعید روحانی (عضو هیئت علمی دانشگاه تهران)

همکاران:

آرش قاضی سعیدی

میثم عسگری آهی

علی محمدی

فاطمه مصلحی

صبا بزرگی

فاطمه مظفری

روشنک آقاباقری

محمدرضا مرادی

حمید جمالی

زهرا رفیعی پور

احسان نگهدار

امین صالح‌نژاد

سمانه عبادی

فهرست مطالب

۱ هوشمندی و تحلیل کسب و کار

فصل اول کتاب هوش مصنوعی برای تحلیل کسب و کار: الگوریتم ، پلتفرم و سناریوهای کاربردی به بررسی مفاهیم اولیه کسب و کار و لایه¬های تحلیلی مختلفی از هوشمندی که در هر سازمان باید مورد توجه قرار گیرد می پردازد.
سه مشخصه اصلی عظیم داده؛ حجم، تنوع و سرعت پیشران های اصلی برای نیاز کسب و کارها به تصمیمات مبتنی بر داده می باشد. بنابراین آنچه که امروزه در تحلیل کسب و کار مورد توجه است به کارگیری افراد مناسب برای استخراج داده های مناسب و ارائه تحلیل درست در زمان مناسب می باشد تا بدین ترتیب فرآیند تبدیل داده خام به تصمیمات نهایی مبتنی بر داده منجر شود.

زمانی که صحبت از تصمیمات داده محور می‌شود نیاز است تا سه لایه تحلیلی بر اساس مقطع زمانی مورد بررسی قرار گیرند:

گذشته: چه اتفاقاتی در گذشته رخ داده است؟ چگونه و به چه دلیل اتفاق است؟

حال: در حال حاضر چه اتفاقاتی در جریان هستند و بهترین اقدام برای مرحله بعدی چه تصمیماتی می‌باشد؟

آینده: چه اتفاقی در آینده رخ خواهد داد؟ چه سناریوهایی را می توان برای اتفاقات محتمل در نظر گرفت؟

زمانیکه صحبت از هوشمندی کسب و کار می شود رویکرد گذشته نگر بیشتر مدنظر می باشد درحالیکه در تحلیل داده محور کسب و کار علاوه بر رویکرد گذشته نگر، رویکرد آینده نگر نیز مورد توجه قرار می گیرد.
بر اساس همین تقسیم بندی زمانی سه نوع رویکرد تحلیلی کسب و کار مطرح می شود:

تحلیل توصیفی: این رویکرد با استفاده از تکنیک های آماری و تحلیل داده بر اتفاقات گذشته متمرکز است تا به صورت تجمعی و خلاصه شده، بینشی را در مورد اتفاقات گذشته ارائه دهد. تنظیم و محاسبه شاخص های کلیدی کسب و کار در بخش های مختلف عملیات، مالی، فروش ، بازاریابی و ... در این لایه قرار می گیرند چرا که با محاسبه این سنجه ها بر اساس داده های تاریخی می توان به تحلیل هوشمندانه ای از گذشته رسید.

تحلیل پیش گویانه: لایه بعدی تحلیل داده در سازمان مربوط به تحلیل هایی از جنس پیش بینی دارد. در این لایه با استفاده از الگوریتم های داده کاوی، هدف یافتن الگو و پیش بینی آینده می باشد. بدین منظوری مدل های مختلفی بر اساس داده ها استخراج شده است و سپس با مقایسه مدل ها بر اساس شاخص های صحت و جامعیت مدل بهترین مدل پیش بینی برای داده‌ها انتخاب می شود.

منابع داده: با توجه به وجود دپارتمان های مختلف و همچنین تبادل داده ای که خارج از سازمان قرار می گیرد همواره منابع داده ای مختلفی وجود دارد که به عنوان خوراک اولیه تصمیمات داده محور خواهند بود.
آماده سازی داده: با توجه به ناسازگاری که بین منابع مختلف وجود دارد و همچنین موضوع کیفیت داده ها نیاز است تا اقداماتی در مورد اماده سازی و پیش پردازش اولیه داده ها صورت گیرد تا نهایتا تصمیمات بر اساس داده های مورد اطمینان باشند.
ذخیره سازی داده: منظور از ذخیره سازی در این چارچوب، ذخیره کردن داده های تحلیلی می-باشد که عموما در انباره داده و دریاچه داده صورت می گیرد.
تحلیل داده: قسمت تحلیل داده اجزای شکسته شده بخش قبلی می باشند که با توجه به هدف کسب و کار و راحتی تحلیل به بخش های معناداری مثل دیتامارت ها تبدیل شدند.

استفاده و دسترسی داده: آخرین لایه این چارچوب مربوط به نحوه دسترسی و استفاده از این داده ها می باشد. این موضوع در هر کسب و کار با توجه به دسترسی های مختلف متفاوت می باشد.
این چارچوب فناوری تحلیل کسب و کار در کنار مراحلی که در تحلیل کسب و کار باید طی کرد، دید کلی را برای طراحی فناوری های مورد نیاز برای هر سازمان ایجاد خواهد کرد. علاوه بر این بستری برای پیاده سازی هوش مصنوعی خواهد بود که موضوع فصول بعدی کتاب می‌باشند.

۲ هوش مصنوعی

در فصل دوم کتاب به بررسی مفاهیم اولیه و کاربرد هوش مصنوعی در تحلیل کسب و کار پرداخته می شود. به طور کلی هوش مصنوعی شامل شاخه های متنوعی همچون پردازش زبان طبیعی، رباتیک، سیستم های خبره، روش های فرا ابتکاری حل مساله و یادگیری ماشین می باشد. آن ساحت از هوش مصنوعی که در موضوعات تحلیلگری کسب و کار مورد توجه قرار می گیرد روش های یادگیری ماشین هست که در این فصل به آن پرداخته می شود.

یادگیری ماشین در هوش مصنوعی به روش ها و الگوریتم های آماری اشاره دارد که با هدف استخراج الگوهای پنهان از مجموعه داده ها به کار گرفته می شوند. به طور کلی الگوریتم های یاد گیری ماشین را در سه طبقه می توان دسته بندی کرد:

- **یادگیری با نظارت:**

در این روش در مجموعه داده ها، کلاس یا متغیری در مجموعه داده ها وجود دارد که به دنبال پیش بینی آن می باشیم. یادگیری در تحت نظارت مدل شامل ایجاد تابعی است که با استفاده از مجموعه داده های آموزشی آموزش داده می شود و سپس می تواند باشد به داده های جدید اعمال می شود. در این روش مجموعه داده آموزشی حاوی رکوردهای برچسب گذاری شده (برچسب) است به طوری که نگاشت به نتیجه دلخواه با توجه به ورودی مجموعه از قبل شناخته شده است. هدف ساختن این تابع به گونهای است که امکان تعمیم برای فراتر از داده های اولیه وجود داشته باشد، یعنی داده های ناشناخته را به نتیجه صحیح اختصاص دهید.

- **یادگیری بدون نظارت:**

یادگیری بدون نظارت یک مدل یادگیری ماشین که فاقد یک برچسب یا کلاس نظارتی است و راهی به مشابه یادگیری با نظارت برای سنجش کیفیت وجود ندارد. هدف این روش خوشه بندی داده ها بر اساس ویژگی های مشترک به دسته های مختلفی می باشد به گونه ای که درون هر خوشه بیشترین شباهت و بین خوشه های بیشترین اختلاف باشد. یادگیری بدون نظارت را در دو گروه خوشه بندی و استخراج قوانین قرار دارد.

- **یادگیری تقویتی:**

یادگیری تقویتی یک مدل یادگیری با توانایی یادگیری نه تنها پیش بینی خروجی بر اساس ورودی، بلکه چگونگی نگاشت متغیرهای ورودی به متغیرهای خروجی بر اساس وابستگی بین متغیرها می باشد. در این روش مفهومی به نام محیط یادگیری تعریف می شود که در ان امکان بررسی عمل و عکس العمل های مختلف بر اساس تنبیه و پاداش ممکن می باشد. در این روش تمامی متغیرهای ورودی بر اساس یک عامل هوشمند نظارت می شود و سپس بر اساس تابع هدف تعریف شده اقدامات مختلفی تعریف می شود تا بتوان تاثیر پیاده سازی اقدامات مختلف را در محیط یادگیری ارزیابی کرد. نهایتا با استفاده از مکانیزم پاداش و تنبیه این امکان برای یادگیری و بهبود مدل فراهم می شود.

علاوه بر سه دسته بندی ارائه شده برای مدل های یادگیری ماشین نیاز هست تا انواع مسائل و تکنیک های یادگیری ماشین نیز آشنا شویم که در ادامه فصل به آن پرداخته شده است.

- **طبقه بندی:**

فرآیند شناسایی، درک و دسته بندی داده‌ها یا اشیای از پیش تعیین شده است. الگوریتم طبقه بندی یک تکنیک یادگیری نظارت شده است. در این تکنیک، ماشین از مجموعه‌ای از داده‌ها یاد می‌گیرد و سپس آن‌ها را به تعدادی کلاس یا گروه طبقه‌بندی می کند به طور خلاصه در روش نظارت شده، از توابعی استفاده می شود که بتواند یک برچسب از قبل تعریف شده را برای مجموعه‌ای از ورودی‌ها پیش‌بینی کند. نکته مهم در روش طبقه بندی کلاس وابسته و یا برچسب مساله می باشد که به صورت متتغیر اسمی می باشند نه عددی پیوسته. در این دسته روش های مختلف و متنوعی همانند درخت تصمیم گیری، نایو بیز، شبکه های عصبی و نزدیک ترین همسایگی وجود دارد.

- **استخراج قوانین**

در این روش ها، برچسب و یا کلاس خاصی وجود ندارد و هدف شناسایی و کشف الگوهای وابستگی و الگوهای تکرار شونده بین متغیرها می باشد. الگوهای تکرار شونده به معنی وجود وابستگی در میان داده ها می باشد و به قوانینی که چنین روابطی را نشان می دهند، قوانین وابستگی یا قوانین انجمنی گفته می شود. چنین اطلاعاتی می تواند به عنوان مبنای تصمیماتی برای برخی از فعالیت های فروشگاهی همچون ارائه مناسب تخفیف برای محصولات یا قراردادن مناسب محصولات در کنار هم، مورد استفاده قرار گیرد. نقطه قوت این روش¬ها استخراج قوانین شفاف و تفسیر پذیر می باشد که در تحلیل کسب و کار بینش های ارزشمندی را ایجاد می کند.

- **خوشه بندی**

این روش زیرمجموعه‌ای از یادگیری نظارت نشده است که با بهره‌گیری از الگوریتم خوشه بندی، داده‌های بدون برچسب زیادی در اختیار آن قرار داده و اجازه می‌دهیم تا نمونه‌ها را در گروه‌های مجزا طبقه‌بندی کند. به هر یک از این گروه‌ها «خوشه» می‌گویند. یک خوشه، گروهی شامل نقاط داده شبیه به یک‌دیگر از نظر ارتباطشان با نمونه‌های مجاور است. در این نوع مسائل، در ابتدای کار، اطلاعات چندانی درباره مسئله نداریم و به همین خاطر، بهره‌گیری از الگوریتم های خوشه بندی نقطه شروع خوبی برای آشنایی با داده‌ها است.

- **پیش بینی و رگرسیون:**

یکی دیگر از مسائل متداولی که در یادگیری ماشین برای تحلیل کسب و کار با آن مواجه می شویم استفاده از رگرسیون برای پیش بینی می باشد. در روش های رگرسیون به دنبال یافتن ارتباط بین متغیر وابسته که یک متغیر عددی پیوسته می باشد با مجموعه متغیرهای ورودی هستیم. مکانیزم این روش مشابه روش های یادگیری با نظارت می باشد به این صورت که مجموعه داده به دو دسته آموزش و تست تقسیم می شود تا بتوان بهترین مدل پیش بینی را انتخاب کرد.

- **بهنیه سازی:**

بهینه سازی فرایندی است که در آن بهترین جواب (با توجه به مجموعه‌ای از معیارها) از میان مجموعه‌ای از جواب‌های ممکن، برای یک مسأله خاص انتخاب می‌شود. در هر مساله کسب و کاری که هدف بیشینه یا کمینه کردن تابع هدفی باشد می توان از روش های بهینه سازی استفاده کرد. نکته مهم در این مسائل شناسایی دقیق محدودیت ها و مدل سازی ریاضیاتی آن ها می باشد تا بتوان با در نظر گرفتن تمامی محدودیت های کسب و کاری بهترین جواب ممکن برای متغیرهای تصمیم گیری را گرفت.

- **سیستم های توصیه گر:**

سیستم های توصیه گر زیر مجموعه‌ای از سیستم‌های فیلتر اطلاعات محسوب می‌شوند. هدف سیستم های توصیه گر، پیشنهاد مناسب‌ترین آیتم‌ها به کاربران است. در روش های مبتنی بر پیشنهاد، داده‌های مرتبط با رفتار کاربران (در خرید کالا، دریافت داده‌ها، تماشای سرگرمی و سایر موارد) تحلیل و مدل‌سازی می‌شوند. سپس، از مدل تولید شده برای پیش‌بینی مناسب‌ترین آیتم برای کاربران استفاده می‌شود. بیشترین استفاده سیستم های توصیه گر، در پیاده‌سازی کاربردهای تجاری به ویژه، در حوزه محصولات و خدمات مصرفی کاربران است.

۳ هوش مصنوعی و پلتفرم‌های تحلیلگری کسب و کار

بخش اول از فصل سوم کتاب هوش مصنوعی برای تحلیلگری کسب و کار: الگوریتم‌ها، پلتفرم‌ها و سناریوهای کاربردی، ابتدا به مفاهیم و چارچوب‌های نرم افزاری پایه همچون مدیریت داده‌ها، انبارداده، و دریاچه داده پرداخته شده است و سپس به معرفی ابزارها و پلتفرم‌های مختلف مرتبط با پردازش و تحلیل داده‌های عظیم می‌پردازد.

پلتفرم‌هایی که می‌توانند برای پیاده‌سازی دریاچه داده استفاده شوند عبارتند از:
فایل سیستم توزیع شده هدوپ (HDFS)
سرویس ذخیره‌سازی ساده آمازون (S3)، ذخیره‌سازی ابری گوگل، ذخیره‌سازی دریاچه داده Azure
دریاچه‌های داده همچنین می‌توانند با انبارداده کلاسیک، HBase یا پایگاه داده NoSQL (مانند MongoDB) ترکیب شوند.

پردازش جریان داده و صف پیام

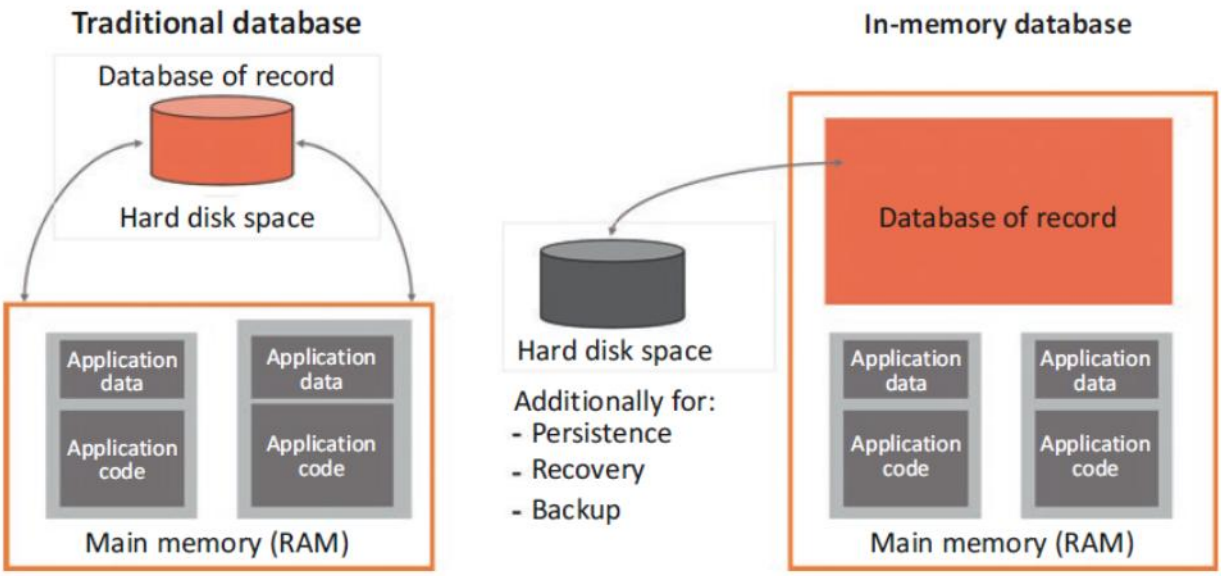
پردازش جریانی اصطلاحاً پردازش داده‌های در حرکت است. به بیان دیگر، تحلیل داده‌ها به صورت مستقیم در لحظه تولید یا دریافت آن‌ها. این الگو این واقعیت را در نظر می‌گیرد که اکثر داده‌ها به صورت جریان‌های پیوسته تولید می‌شوند: رویدادهای سنسورها، فعالیت کاربر در یک وب‌سایت یا داده‌های معاملات مالی. همه این داده‌ها به صورت مجموعه‌ای از رویدادها در طول زمان تولید می‌شوند. داده‌های عظیم، ارزش بینش‌هایی را از پردازش داده‌ها تعیین کرده است. چنین بینش‌هایی همه یکسان نیستند. برخی از بینش‌ها اندکی پس از وقوع ارزشمندتر هستند، زیرا ارزش آن‌ها با گذشت زمان بسیار سریع کاهش می‌یابد. پردازش جریان، چنین سناریوهایی را امکان‌پذیر می‌سازد و بینش‌ها را سریع‌تر، اغلب در عرض میلی‌ثانیه تا ثانیه پس از ایجاد ارائه می‌دهد .محاسبات جریان همچنین می‌توانند چندین جریان را با هم پردازش کنند و هر محاسبه روی جریان رویداد می‌تواند جریان‌های رویداد دیگری تولید کند. در مقابل پردازش جریانی پردازش دسته‌ای قرار دارد. پردازش دسته ای برای محاسبه پرس‌وجوهای دلخواه روی مجموعه داده‌های مختلف قابل استفاده است. معمولاً نتایج حاصل از تمام داده‌های تحت پوشش را محاسبه می‌کند و امکان تحلیل عمیق مجموعه داده‌های عظیم را فراهم می‌کند. سیستم‌های مبتنی بر MapReduce مانند Amazon EMR نمونه‌هایی از پلتفرم‌هایی هستند که از چنین پردازش‌هایی پشتیبانی می‌کنند.

در مقابل، پردازش جریان نیازمند گرفتن یک دنباله از داده‌ها و به‌روزرسانی تدریجی سنجه‌ها، گزارش‌ها و آمار خلاصه در پاسخ به هر مجموعه داده ورودی است. بنابراین، پردازش جریان برای نظارت و پاسخگویی بلادرنگ مناسب‌تر است.

امروزه بسیاری از شرکت‌ها با ترکیب دو رویکرد و ایجاد یک لایه بلادرنگ (جریان) و یک لایه دسته ای به طور همزمان یا سری، بر روی مدل‌های ترکیبی کار می‌کنند. داده‌ها ابتدا توسط یک پلتفرم داده جریانی برای ارائه بینش‌های بلادرنگ پردازش می‌شوند و سپس به یک مخزن داده بارگذاری می‌شوند که در آنجا می‌توان آن را تبدیل و برای انواع مختلف موارد استفاده پردازش دسته‌ای استفاده کرد.

سیستم مدیریت پایگاه داده

یک سیستم مدیریت داده in-memory (IMDBMS) یک سیستم مدیریت پایگاه داده است که عمدتاً برای ذخیره‌سازی، مدیریت و دستکاری داده‌ها به حافظه اصلی متکی است. این کار تأخیر و سربار ذخیره‌سازی دیسک را حذف می‌کند و مجموعه دستورالعمل‌های مورد نیاز برای دسترسی به داده‌ها را کاهش می‌دهد. برای ذخیره‌سازی و دسترسی کارآمدتر، داده‌ها می‌توانند در یک قالب فشرده ذخیره شوند.



آپاچی هدوپ

محیط‌های عظیم داده معمولاً نه تنها شامل داده‌های عظیم بلکه انواع مختلفی از داده‌های تراکنش ساختاریافته تا انواع نیمه‌ساختاریافته و بدون ساختار اطلاعات مانند سوابق کلیک، وب سرور و لاگ‌های برنامه‌های موبایل، پست‌های رسانه‌های اجتماعی، ایمیل‌های مشتری و داده‌های سنسور از اینترنت اشیا می‌شوند. فناوری که در ابتدا فقط با نام آپاچی هدوپ شناخته می‌شد، به عنوان بخشی از یک پروژه متن‌باز در بنیاد نرم‌افزار آپاچی (ASF) در حال توسعه است. توزیع تجاری هدوپ در حال حاضر توسط چهار ارائه دهنده اصلی پلتفرم داده‌های عظیم ارائه می‌شود: فناوری‌های آمازون وب سرویس (AWS)، کلودرا، Hortonworks و MapR. علاوه بر این، گوگل، مایکروسافت و سایر فروشندگان خدمات مدیریت مبتنی بر ابر را بر اساس هدوپ و فناوری‌های مرتبط ارائه می‌دهند.

تحلیل داده و زبان‌های برنامه‌نویسی

علاوه بر الگوریتم‌ها، تحلیل داده همچنین نیازمند پیاده‌سازی و یکپارچه‌سازی با اپلیکیشن‌های موجود است. زبان‌های برنامه‌نویسی مانند پایتون و R در این راستا مورد استفاده قرار می‌گیرند. R نه تنها توسط کاربران دانشگاهی استفاده می‌شود، بلکه بسیاری از شرکت‌های بزرگ از جمله اوبر، گوگل، ایربی‌ان‌بی، فیس‌بوک و غیره نیز از R استفاده می‌کنند.

همچنین اسکالا (Scala) یک زبان برنامه‌نویسی سطح بالا، چند پارادایمی و همه منظوره است. این زبان یک زبان برنامه‌نویسی شیءگرا است که از رویکرد برنامه‌نویسی تابعی نیز پشتیبانی می‌کند. هیچ داده اولیه‌ای وجود ندارد، زیرا همه چیز در اسکالا یک شیء است. اسکالا برای بیان الگوهای رایج برنامه‌نویسی به شیوه‌ای دقیق، مختصر و type-safe طراحی شده است. برنامه‌های اسکالا می‌توانند به بایت کد تبدیل شوند و روی ماشین مجازی جاوا (JVM) اجرا شوند.

جولیا (Julia) در سال ۲۰۰۹ ایجاد شد و در سال ۲۰۱۲ به عموم معرفی شد. جولیا با هدف رفع کمبودهای پایتون و سایر زبان‌ها و برنامه‌های محاسبات علمی و پردازش داده طراحی شده است. جولیا از metaprogramming پشتیبانی می‌کند. برنامه‌های جولیا می‌توانند برنامه‌های جولیای دیگر را تولید کرده و حتی کد خود را به روشی شبیه به زبان‌هایی مانند Lisp تغییر دهند.

چارچوب های هوش مصنوعی

در گذشته برای پیاده سازی مفاهیم و الگوریتم های هوش مصنوعی نیاز بود تا منطق و فرمول های ریاضیاتی از ابتدا و با جزئیات نوشته شوند که کار نسبتا سخت و پیچیده ای بود. اما در حال حاضر با توجه به با توجه به پیشرفت زیرساخت ها و نرم افزارهای متن باز این امکان برای متخصصان هوش مصنوعی وجود دارد که با فراخوانی چارچوب های قدرتمند هوش مصنوعی الگوریتم های مورد نیاز خود را از صفر بازنویسی کنند. به عبارتی دیگر این چارچوب ها با دریافت پارامترهای ریاضی مختلف و فیت شدن روی دیتاست های مختلف می توانند خروجی های مورد نظر برای تحلیلگران را فراهم کنند تا در وقت و انرژی صرفه جویی قابل توجهی داشته باشند. از میان تمامی چارچوب های هوش مصنوعی ۵ چارچوب محبوبیت و کاربردهای زیادی دارند که در ادامه به بررسی آن ها می پردازیم:

TensorFlow.¬1

تنسور فلو ابتدا توسط گوگل برای استفاده داخلی توسعه داده شد و در سال ۲۰۱۵ تحت مجوز اپن سورس آپاچی ۲ منتشر شد. گوگل همچنان از این کتابخانه برای خدمات مختلفی مانند تشخیص گفتار، جستجوی تصاویر و پاسخ‌های خودکار در جیمیل استفاده می‌کند. محبوبیت گوگل و مدل‌های گراف جریان داده‌ای که تنسور برای ایجاد مدل‌ها استفاده می‌کند، باعث جذب تعداد زیادی از مشارکت‌کنندگان به این کتابخانه شده است. این امر به دسترسی عمومی همراه با مستندات و آموزش‌های دقیق منجر شده که ورود به دنیای شبکه‌های عصبی را آسان می‌کند. تنسور فلو یک ابزار پایتون برای بررسی شبکه‌های عصبی عمیق و محاسبات پیچیده ریاضی است و حتی از یادگیری تقویتی نیز پشتیبانی می‌کند. این کتابخانه از گراف‌های جریان داده تشکیل شده که شامل گره‌ها (عملیات ریاضی) و لبه‌ها (آرایه‌های عددی یا تنسورها) می‌شود. انعطاف‌پذیری این چارچوب به دلیل امکان استفاده از آن برای تحقیقات و وظایف تکراری یادگیری ماشین است. کاربران می‌توانند از APIسطح پایین هسته های تنسورفلو استفاده کنند که کنترل کامل روی مدل‌ها و داده‌ها فراهم می‌کند. همچنین مدل‌های از پیش آموزش دیده‌ای هم وجود دارند که به کاربران اجازه می‌دهند از APIهای سطح بالا استفاده کنند.

همچنین این چارچوب به دلیل نیاز به مهارت‌های کدنویسی گسترده و دانش دقیق از علم داده، ممکن است برای افراد تازه‌کار مناسب نباشد و بسیاری آن را کتابخانه‌ای پیچیده می‌دانند. در نتیجه، استفاده از آن بیشتر در شرکت‌های بزرگ با دسترسی به متخصصان یادگیری ماشین معمول است. به عنوان مثال، سوپرمارکت آنلاین Ocado در بریتانیا از تنسورفلو برای اولویت‌بندی ایمیل‌ها و پیش‌بینی تقاضا استفاده کرده و شرکت بیمه جهانی Axaاز آن برای پیش‌بینی ادعاهای بزرگ مشتریان بهره می‌برد.

Theano.2

یک کتابخانه محاسبات علمی سطح پایین مبتنی بر پایتون است که برای وظایف یادگیری عمیق مرتبط با تعریف، بهینه‌سازی و ارزیابی عبارات ریاضی استفاده می‌شود. با اینکه این کتابخانه قدرت محاسباتی چشمگیری دارد، کاربران از رابط کاربری غیرقابل دسترسی و پیام‌های خطای غیرمفید آن شکایت دارند. به همین دلیل، عمدتاً در ترکیب با کتابخانه‌های رابط کاربری سطح بالا مثل Keras، Lasagne, Blocks که برای نمونه‌سازی سریع و تست مدل‌ها استفاده می‌شوند، به کار می‌رود. برای Theano مدل‌های عمومی وجود دارد، اما هر چارچوب دیگری نیز که استفاده شود، بسیاری از آموزش‌ها و مجموعه‌داده‌های پیش‌آموزش‌دیده در دسترس است. به عنوان مثال، Kerasمدل‌های موجود و آموزش‌های دقیقی را در مستندات خود ذخیره می‌کند.

همچنین با استفاده از Lasagne یا Keras به عنوان یک رابط کاربری سطح بالا، شما به انواع مختلفی از آموزش‌ها و مجموعه داده‌های پیش‌آموزش‌دیده دسترسی دارید. سادگی و بلوغ تئانو به تنهایی از نکات مهمی است که باید در این تصمیم‌گیری مد نظر قرار داد. این چارچوب به عنوان یک استاندارد صنعتی برای تحقیقات و توسعه یادگیری عمیق محسوب می‌شود و در ابتدا برای پیاده‌سازی الگوریتم‌های یادگیری عمیق پیشرفته توسعه داده شد. با این حال، از آنجایی که افراد به ندرت به طور مستقیم از تئانو استفاده می‌کنند، کاربردهای زیادی از آن به عنوان پایه‌ای برای کتابخانه‌های دیگر گسترش یافته است: تشخیص دیجیتال و تصویر، مکان‌یابی اشیاء، و حتی چت‌بات‌ها.

Torch. 3

این چارچوب اغلب به عنوان ساده‌ترین ابزار یادگیری عمیق برای مبتدیان شناخته می‌شود. این ابزار از یک زبان اسکریپت‌نویسی ساده به نام Luaاستفاده می‌کند و جامعه‌ای فعال دارد که مجموعه‌ای چشمگیر از آموزش‌ها و بسته‌ها را برای تقریباً هر هدف یادگیری عمیق ارائه می‌دهد. با وجود اینکه زبان پایه Luaکمتر رایج است، خود Torchبه طور گسترده‌ای استفاده می‌شود .شرکت‌هایی مانند فیسبوک، گوگل و توییتر از آن در پروژه‌های هوش مصنوعی خود استفاده می‌کنند. فهرستی از مجموعه‌داده‌های محبوب برای استفاده در این چارچوب را می‌توان در صفحه گیت‌هاب پیدا کرد. علاوه بر این، فیسبوک کد رسمی برای پیاده‌سازی شبکه‌های عصبی عمیق باقی‌مانده (ResNets) با مدل‌های پیش‌آموزش‌دیده را همراه با دستورالعمل‌هایی برای تنظیم دقیق مجموعه‌داده‌های شما منتشر کرده است. صرف نظر از تفاوت‌ها و شباهت‌ها، انتخاب همیشه به زبان پایه بستگی دارد زیرا تعداد توسعه‌دهندگان با تجربه در Luaهمیشه کمتر از پایتون خواهد بود. با این حال، به طور قابل‌توجهی خواناتر است و این موضوع در سینتکس ساده Torchمنعکس می‌شود

Jupyter Notebook.5

چارچوب آخری که قصد معرفی آن را داریم یک برنامه وب منبع‌باز است که به شما امکان می‌دهد اسنادی حاوی کد زنده، معادلات، تجسم‌ها و متن ایجاد و به اشتراک بگذارید. این چارچوب که توسط تیم پروژه Jupyter پشتیبانی می‌شود یک پروژه جانبی از پروژه IPythonاست که قبلاً خود پروژه IPython Notebook را داشت.این چارچوب برای انواع مختلف پروژه‌ها در روش‌های مختلف مناسب است:

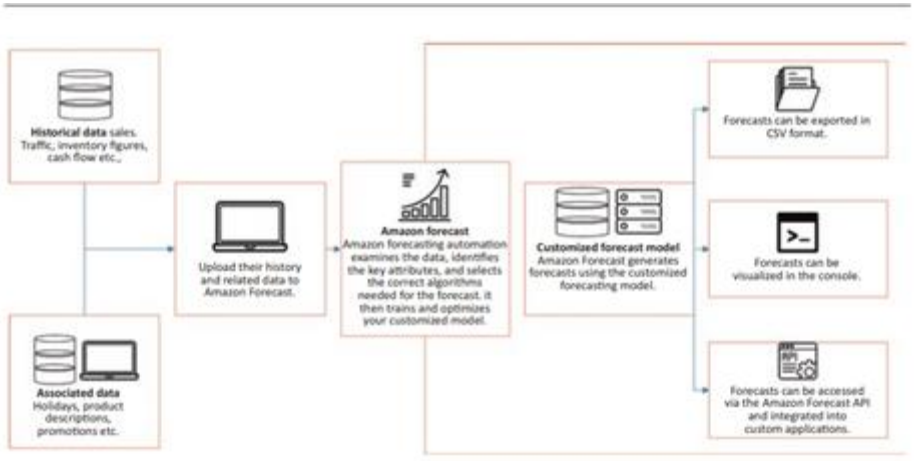
- تجسم داده‌ها: اکثر افراد اولین مواجهه خود با Jupyter Notebookرا از طریق یک تجسم داده تجربه می‌کنند، یک دفترچه به اشتراک‌گذاشته‌شده که شامل نمایش یک مجموعه‌داده به صورت نمودار است. Jupyter به شما امکان می‌دهد تا تجسم‌ها را انجام دهید، آن‌ها را به اشتراک بگذارید و تغییرات تعاملی را در کد و مجموعه‌داده مشترک اعمال کنید.
- به اشتراک‌گذاری کد: سرویس‌های ابری مانند GitHub روش‌هایی برای به اشتراک‌گذاری کد ارائه می‌دهند، اما این روش‌ها تا حد زیادی غیرتعاملی هستند. با Jupyter Notebook، می‌توانید کد را مشاهده کنید، آن را اجرا کنید و نتایج را مستقیماً در مرورگر وب خود ببینید.
- تعاملات کد زنده: کد Jupyter Notebook ثابت نیست؛ می‌توان آن را ویرایش کرد و در زمان واقعی به‌صورت تدریجی دوباره اجرا کرد، با بازخورد مستقیم در مرورگر. دفترچه‌ها همچنین می‌توانند کنترل‌های کاربر (مانند اسلایدرها یا فیلدهای ورودی متن) را جاسازی کنند که می‌توانند به‌عنوان نقاط ورودی برای کد استفاده شوند.
- مستندسازی مثال‌های کد – اگر قطعه‌ای از کد دارید و می‌خواهید آن را خط به خط با بازخورد زنده توضیح دهید، می‌توانید آن را در Jupyter Notebook جاسازی کنید.

یکی از چالش های اصلی پیاده سازی الگوریتم¬ های یادگیری ماشین ایجاد زیرساخت های لازم و سرمایه گذاری برای یادگیری این متدها می باشد. استفاده از سرویس های ابری برای یادگیری ماشین این چالش را تا حدی برطرف می¬کند.
مفهوم یادگیری ماشین به عنوان سرویس تمام ابعاد یادگیری ماشین همچون زیرساخت ها، پردازش داده، مدل های یادگیری و ارزیابی را در بر می¬گیرد. پلتفرم های ابری متنوعی همچون سرویس Amazon, Azure, IBM, Googleدر این راستا وجود دارند که در ادامه به مختصر توضیحی پرداخته می شود. علاوه بر موضوع یادگیری ماشین، مفهوم ذخیره و نگه داری داده در این سرویس ها نیز حائز اهمیت می باشد چرا که پایه اولیه برای تمامی الگوریتم های یادگیری ماشین وجود بستر و یک زیرساخت مناسب برای ذخیره و نگه داری حجم انبوهی از داده می باشد.

سرویس ابری آمازون

سرویس ابری آمازون بر اساس حق عضویت خدمات مختلفی در زمینه ذخیره، پردازش و یادگیری ماشین به کسب و کارها ارائه می دهد. طبیعتا برای پشتیبانی از این موضوع تمام منابع موردنیاز کامپیوتر همانند گرافیک، رم ، شبکه و هر آنچه که مورد نیاز است را در مقیاس بزرگ پشتیبانی می کند تا موضوع مدیریت منابع نیز در سمت آمازون مدیریت شود. بر این مبنا تمامی کاربران آمازون وب سرویس می توانند از درگاه وب موردنظر به سیستم تحلیل داده ای که نیاز دارند متصل شوند و نیازهای خودشون را مرتفع کنند. سرویس ابری آمازون، امکانات و ابزارهای تحلیلی مختلفی را برای صاحبان کسب و کار جهت کاربردهای تحلیلی فراهم می کند. نکته مثبت در مورد این سرویس، وجود ماژول های متنوع برای هوش مصنوعی و تحلیل کسب و کار می باشد که اینکار با ارائه دو راه حل تحلیل¬های پیش گویانه برای تحلیل گران داده و ابزارهای تحلیلی برای دانشمندان داده پوشش داده می شود. طبیعتا راه حل های ارائه شده آمازون بدون فضای ذخیره سازی مناسب میسر نخواهد بود در این راستا سرویس داده ای آمازون ارائه می شود تا مباحث مربوط به ذخیره و نگه داری داده در این بستر تسهیل شود. سرویس داده آمازون ۱۲ ابزار مختلف برای ذخیره سازی داده ارائه می کند تا تمامی نیازهای ذخیره سازی را پوشش دهد. سرویس داده آمازون از مدل های پایگاه داده ای مرسوم که در بیشتر کسب و کارها استفاده می شود پشتیبانی می کند.

یکی دیگر از امکانات آمازون وب سرویس که در تحلیل داده بسیار قابل توجه می شود سرویس یادگیری ماشین می باشد. این سرویس در کنار سرویس داده آمازون این امکان را فراهم می کند تا ضمن اتصال به تمامی داده های کسب و کارها، استفاده و پیاده سازی تمامی الگوریتم های یادگیری ماشین در مقیاس های بزرگ نیز امکان پذیر باشد. بعد از پیاده سازی جریان داده و الگوریتم های یادگیری ماشین می¬توان فلو موردنظر را اتوماتیک کرد تا این پروسه به صورت اتوماتیک از دریافت و پردازش داده تا لایه تحلیل پیش رود.



یکی از قدرت های سرویس آمازون مدل¬های پیش بینی است که در تصویر مشاهده می شود. ایجاد مدل¬های پیش بینی اختصاصی این امکان را فراهم می کند تا در قالب های مختلفی همچون داشبورد، تعبیه سازی در اپلیکیشن استفاده شود.

سرویس ابری گوگل

یکی دیگر از معروف ترین سرویس های ابری در زمینه تحلیل داده و هوش مصنوعی سرویس ابری گوگل می باشد. سرویس ابری گوگل مشابه آمازون دارای زیرساخت های ذخیره سازی، پردازش و اپلیکیشن های مختلف می باشد. علاوه بر سرویس های متداول، سرویس های پردازش عظیم داده و اینترنت اشیا در این پلتفرم قابل استفاده می باشد. سرویس داده ای گوگل در قالب دیتابیس فایر بیس ارائه می شود و این امکان را فراهم می آورد تا تمامی نیازمندی ها ذخیره سازی؛ نگه داری و پردازش به صورت یکپارچه در یک پایگاه داده ذخیره شود.

یکی از ابزارهای تحلیلی معروف سرویس ابری گوگل، بیگ کوئری می باشد که منزله انباره داده عمل می کند تا با استفاده از کوئری های استاندارد مختلف بتوان به تحلیل و پردازش داده های موردنیاز کسب و کار پرداخت. مزیت مثبت این سرویس در قیاس با پایگاه های داده ای لایه تحلیلی می باشد که امکان بازخوانی حجم انبوهی از داده را فراهم می کند. سرویس ابری گوگل از ابزارهای پیشرفته و اتوماتیک یادگیری ماشین پشتیبانی می کند که امکان دریافت، ذخیره سازی و پردازش انواع داده های ساختاریافته و غیرساختاریافته را دارد.

سرویس آژور مایکروسافت

زمانیکه صحبت از یادگیری ماشین به عنوان یک سرویس می شود سرویس آژور به علت کاربردهای متنوعی که دارد. دیتابیس هایی که سرویس آژور پشتیبانی می کند شامل هر دو نوع دیتابیس های رابطه ای و غیر رابطه ای می باشد تا امکان پردازش و تحلیل عظیم داده¬ها را داشته باشد. این سرویس به دلیل مقیاس پذیر بودن امکان استفاده از زبان های برنامه نویسی مثل پایتون و ویژال استودیو را فراهم می کند تا در صورت نیاز امکان سفارشی سازی کردن الگوریتم ها متناسب با نیازهای کسب و کار فراهم باشد.

پلتفرم تکنولوژی SAP

یکی دیگری از پلتفرم هایی که ارائه محصولات نرم افزاری به عنوان سرویس را پشتیبانی می کند مربوط به شرکت SAP می باشد. این پلتفرم تنها برای بحث تحلیلی و یادگیری ماشین نمی باشد و سرویس های دیگری برای توسعه نرم افزار، موبایل و تجربه کاربری را نیز پوشش می دهد اما آنچه که در این فصل مورد تمرکز می باشد سرویس داده ای این پلتفرم می باشد. این سرویس با پشتیبانی از پایگاه داده های مختلفی امکان ذخیره سازی نگه داری و پردازش را فراهم می کند. ساختار این پلتفرم به نحوی می باشد که امکان طراحی پایپ لاین داده را به منظور اتوماتیک سازی فرآیند ممکن می کند. علاوه بر ساختار ذخیره سازی و نگه داری، سرویس یادگیری ماشین این پلتفرم با ایجاد مدل های مختلف پیش بینی، امکان پیاده سازی انواع مختلف مدل ها را در تحلیل کسب و کار فراهم می کند.

۴ موارد مطالعاتی از به کارگیری تحلیلگری کسب و کار مبتنی بر هوش مصنوعی

فصل چهارم کتاب هوش مصنوعی برای تحلیلگری کسب و کار: الگوریتم‌ها، پلتفرم‌ها و سناریوهای کاربردی، به بررسی مورد مطالعه تحلیل احساسات مشتری به صورت بلادرنگ با استفاده از تحلیلگری جریان داده‌ها می‌پردازد.

در دنیای رقابتی خرده فروشی، رضایت مشتریان نقش کلیدی در موفقیت کسب و کارها ایفا می‌نماید. این مطالعه موردی بر تحلیل احساسات مشتریان به صورت بلادرنگ با استفاده از تحلیل جریان داده‌ها متمرکز است. هدف اصلی، توسعه یک شاخص برای رضایت مشتریان است که بتواند بازخوردهای مشتریان را به صورت لحظه‌ای پردازش و تحلیل نماید.
یکی از بزرگترین شرکت‌های خرده فروش در آلمان به بررسی راهکاری برای اندازه گیری و تحلیل لحظه‌ای رضایت مشتریان پرداخته است. این شرکت در این راستا به توسعه شاخص جریان احساسات مشتری (CSSI) پرداخته است که بتواند نظرات مشتریان را از شبکه‌های اجتماعی، شبکه‌های اجتماعی، تراکنش‌های فروشگاهی و سایر منابع اطلاعاتی جمع‌آوری و تحلیل کند.

اهمیت رضایت مشتری در بخش خرده فروشی:

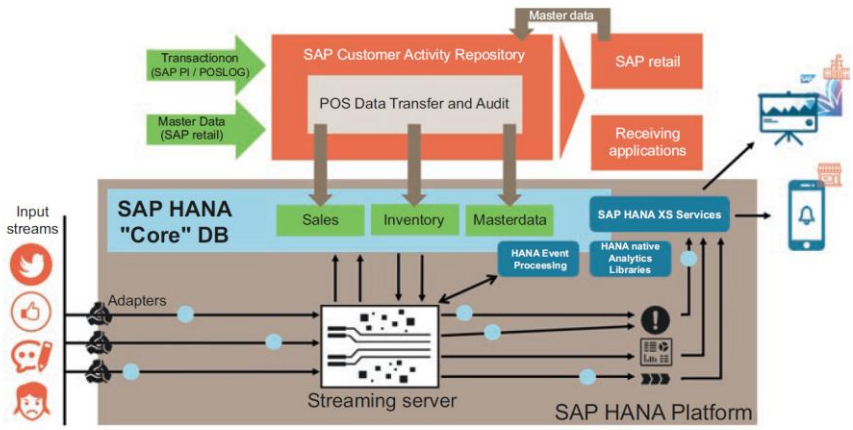
با افزایش رقابت در خرده‌فروشی و ظهور بازیگران دیجیتال، تفاوت بین فروشگاه‌های فیزیکی و آنلاین در حال کاهش است. مطالعات نشان داده‌اند که ارتباط مستقیمی بین رضایت مشتری و وفاداری به برند وجود دارد. خرده‌فروشان برای حفظ جایگاه خود نیاز به تحلیل عمیق بازخوردهای مشتریان دارند.

شاخص جریان احساسات مشتری (CSSI)

شاخص CSSI یک شاخص پویا است که با استفاده از جریان داده‌ها از تعاملات مشتریان در فضای دیجیتال محاسبه می‌شود. این شاخص شامل معیارهای مختلفی همچون میزان احساسات مثبت و منفی، مشکلات ثبت‌شده در فروشگاه و سایر پارامترهای کلیدی است.

فناوری و پذیرش داده‌های چندکاناله:

تحلیل احساسات در این مطالعه با استفاده از داده‌های به دست آمده از چندکانال شامل نظرات مشتریان در شبکه‌های اجتماعی، داده‌های فروشگاهی و سایر منابع اطلاعاتی انجام شده است. یک معماری داده‌ای بر پایه‌ی SAP HANA پیاده‌سازی شده است که امکان جمع‌آوری و پردازش داده‌ها به صورت بلادرنگ را فراهم می‌کند.
شکل زیر معماری این سیستم را بر اساس معماری مرجع خرده فروشی نشان می‌دهد.



بر اساس این معماری یک سیستم تحلیلی توسعه داده شده که داده‌های مختلف از جمله نظرات مشتریان، تراکنش‌های فروشگاهی، وضعیت موجودی کالاها و مدت زمان انتظار در صف صندوق را به هم مرتبط می‌کند. پردازش این اطلاعات در SAP HANA انجام شده و داشبوردی برای نمایش نتایج در اختیار مدیران فروشگاهی قرار گرفته است.

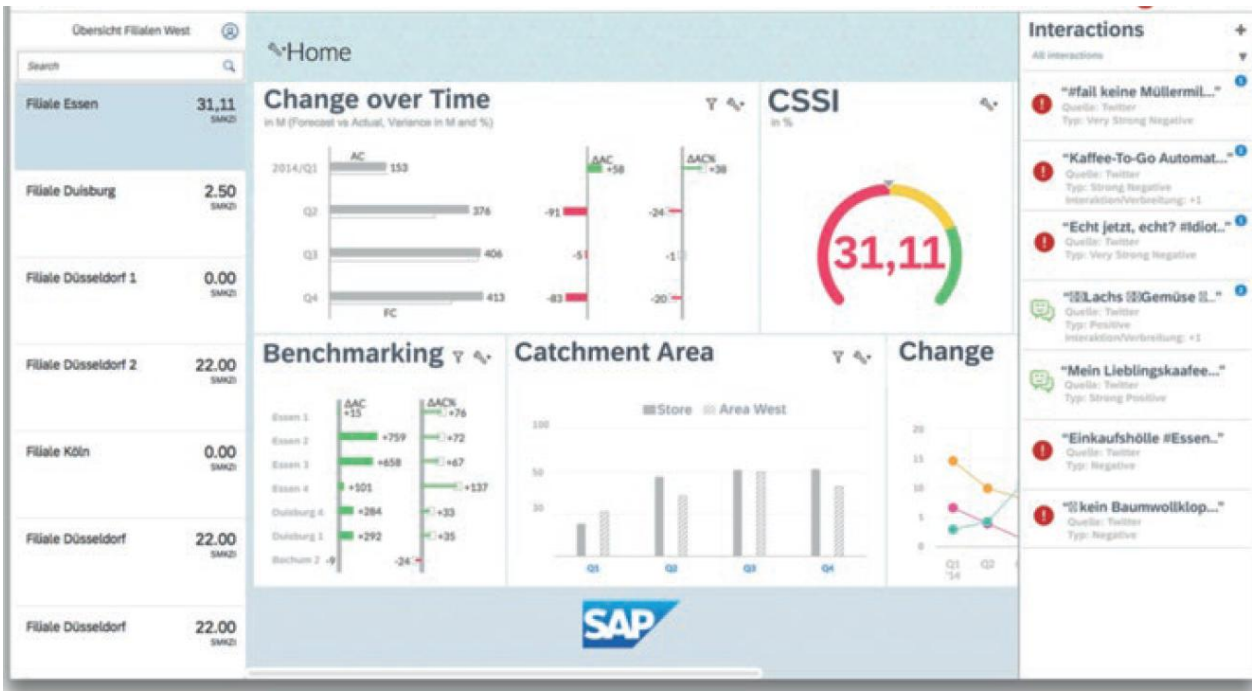
نتایج و مزایا:

در تست اولیه این سیستم، ۲۵۰,۰۰۰ توئیت در طی ۱۱ ماه در سال ۲۰۱۷ مورد بررسی قرار گرفتند. پردازش داده‌ها در سیستم SAP HANA نشان داد که از میان پیام‌های قابل ارزیابی:

۳۲٪ کاملاً مثبت، ۲۰٪ کاملاً منفی، ۱۲٪ تا حدی منفی، ۳٪ مربوط به مشکلات کوچک، و ۳٪ مربوط به مشکلات بزرگ بودند.

بر اساس این داده‌ها، شاخص CSSI محاسبه و در دسترس مدیران کسب و کار قرار گرفت. این اطلاعات از طریق داشبورد مدیریتی یا اعلان‌های مستقیم به کاربران ارائه شد. در سطح بالاتر سازمانی، امکان مقایسه عملکرد شعب مختلف نیز فراهم گردید.

شکل زیر نمونه‌ای از داشبورد مربوط به شاخص CSSI را نشان می‌دهد.



در حالی که شخصی‌سازی در بازاریابی اولویت بسیاری از مدیران خرده‌فروشی است، در عمل کمتر اجرا می‌شود. کمبود داده و نبود فرآیندهای خودکار از موانع اصلی هستند. داده‌های فروش، قیمت، رقبا، و اطلاعات جمعیتی از منابع آنلاین و داخلی جمع‌آوری و در سطح جغرافیایی ۱ کیلومتر مربع مدل‌سازی شدند. با استفاده از الگوریتم Growing Neural Gas، فروشگاه‌ها خوشه‌بندی شده و بازاریابی محلی بر اساس این خوشه‌ها بهینه شد. این کار امکان تست‌های A/B و اتوماسیون بیشتر را فراهم کرده است. همچنین در مطالعه، اهمیت انتخاب مکان فروشگاه و تأثیر آن بر سیاست‌های بازاریابی بررسی شده است. در نهایت، با استفاده از مدل Segmenting-Targeting-Positioning (STP)، نحوه تقسیم بازار به بخش‌های کوچکتر برای بهبود تخصیص منابع و افزایش سود مورد تأکید قرار گرفته است.

روش‌های مختلف خوشه‌بندی داده‌ها برای تقسیم بازار بررسی می‌شوند. در ابتدا، به روش‌های سنتی مانند K-means، روش‌های سلسله‌مراتبی و نقشه‌های خودسازمان‌ده (SOM) اشاره می‌شود که هرکدام مزایا و معایب خاص خود را دارند. اما با رشد حجم داده‌ها و پیچیدگی آنها، این روش‌ها با مشکلاتی مانند انتخاب تعداد خوشه‌ها و ابعاد زیاد داده مواجه‌اند.

در ادامه، الگوریتم Growing Neural Gas و نوع پیشرفته‌تر آن یعنی Growing Neural Gas (GNG) معرفی می‌شود. GNG برخلاف روش‌های سنتی، به صورت افزایشی ساختار خود را تغییر می‌دهد و به مرور نرون‌های جدید اضافه می‌کند. این الگوریتم برای داده‌های بزرگ، پیچیده و نویزی مناسب‌تر است زیرا بدون نیاز به تعیین تعداد خوشه‌ها از پیش، خود را با ساختار داده تطبیق می‌دهد.

GNG در قالب شبکه‌ای بدون جهت عمل می‌کند که ارتباط بین نرون‌ها با توجه به فاصله وزنی آنها شکل می‌گیرد. گره‌هایی که خطای زیادی در خوشه‌بندی ایجاد می‌کنند شناسایی شده و نرون جدیدی در آن ناحیه اضافه می‌شود تا دقت خوشه‌بندی افزایش یابد. این الگوریتم در حوزه‌های متنوعی از جمله تصاویر نجومی و جریان داده نیز مورد استفاده قرار گرفته و اثربخشی آن در خوشه‌بندی داده‌های پیچیده اثبات شده است.

ساختار پروژه و اهداف اصلی پژوهش مطرح شده‌اند که در قالب سه سؤال تحقیقاتی بررسی می‌شوند:

- آیا امکان جمع‌آوری داده‌های کافی از وب برای خوشه‌بندی فروشگاه‌ها وجود دارد؟ این هدف به بررسی قابلیت اتکای داده‌های رایگان موجود در وب به‌عنوان جایگزینی برای داده‌های گران‌قیمت شرکت‌های تحقیقات بازار می‌پردازد. اگر این کار ممکن باشد، می‌تواند منجر به کاهش وابستگی، صرفه‌جویی در هزینه و مزیت رقابتی شود.
- آیا الگوریتم Neural Gas می‌تواند خوشه‌های مناسبی با شباهت درونی بالا و تفاوت بیرونی کم ایجاد کند؟

هدف دوم ارزیابی کارایی الگوریتم NG در مقایسه با روش‌های سنتی خوشه‌بندی است. علاوه بر مقاومت در برابر نویز و سازگاری با داده‌های متغیر، باید کیفیت خوشه‌ها بر اساس معیارهای استاندارد بررسی شود.

- آیا خوشه‌های به‌دست‌آمده برای بخش‌بندی فعالیت‌های بازاریابی و استفاده در اتوماسیون بازاریابی کاربردی هستند؟

خوشه‌بندی‌ها نباید صرفاً از منظر ریاضی مورد بررسی قرار گیرند، بلکه باید در عمل نیز در فرآیندهای بازاریابی خرده‌فروشی قابلیت استفاده و اثربخشی داشته باشند.

در رابطه با سؤال اول مقایسه با منابع علمی نشان می‌دهد که داده‌های خارجی (نظیر داده‌های منطقه‌ای) در این مطالعه گسترده‌تر از بسیاری منابع دیگر هستند، در حالی که داده‌های داخلی نادیده گرفته شده‌اند. با این حال، داده‌های داخلی را می‌توان در آینده از منابع سازمانی اضافه کرد. الگوریتم GNG قابلیت سازگاری با چنین توسعه‌ای را دارد، چراکه به راحتی با بردارهای داده‌ای گسترده‌تر تطبیق می‌یابد.

در رابطه با سؤال دومتست ریاضی و تجسم نتایج نشان می‌دهد که از بین ۳۷۶۱ فروشگاه، ۹ خوشه به دست آمده‌اند. هرچند در نگاه اول تعداد کمی به نظر می‌رسد، اما برای کاربردهای عملی بازاریابی کاملاً قابل استفاده است، چون تغییرات در ترکیب بازاریابی معمولاً نیازمند تغییرات فیزیکی در فروشگاه‌هاست. بنابراین خوشه‌ها از نظر کیفیت ریاضی و کاربرد عملی در خرده‌فروشی ایستا مناسب هستند.

در رابطه با سؤال سومارزیابی بر اساس معیارهای تقسیم‌بندی بازار نشان می‌دهد که خوشه‌ها با عوامل اقتصادی مرتبط‌اند که بر رفتار خرید مشتریان تأثیر دارند. این عوامل از منابع رسمی به‌دست آمده و قابل اندازه‌گیری و دسترسی هستند. با اینکه اجرای عملی خوشه‌ها در بازاریابی هنوز نیاز به آزمایش تجربی دارد، الگوریتم GNG در محیط‌های پویای خرده‌فروشی عملکرد قابل قبولی دارد. پیشنهاد شده است از تست‌های ساده A/B برای بررسی اثربخشی تغییرات در ترکیب بازاریابی استفاده شود تا فعالیت‌های بازاریابی شخصی‌سازی‌شده طراحی شوند.

راه‌های ارتباطی

گروه تحلیل‌گری عظیم‌داده و کسب و کار


https://t.me/BigData_BusinessAnalytics


https://www.instagram.com/dr.saeedrouhani/


https://chat.whatsapp.com/JPD9GYRQfMwFFLEB5YmmLv


https://www.linkedin.com/company/bdba/


https://www.aparat.com/bdbaanalytics.ir


https://bdbaanalytics.ir


bdbanalytic@gmail.com

Felix Weber

Artificial Intelligence for Business Analytics

Algorithms, Platforms and Application Scenarios

Springer