**یافتن مشتریان وفادار با استفاده از تکنیک های یادگیری ماشین و پیش بینی مبتنی بر یادگیری**

**چکیده :**

طبق اصول مشتری مداری، وفاداری برای هر ســازمانی اهمیت استراتژیک دارد. اﻣﺮوزه، ﺷﻨﺎﺧﺖ و پیش بینی نیازﻫﺎی ﻣﺸﺘﺮﻳﺎن ﺑﺮای سازمان‌ها از اهمیت ﺧﺎﺻـی ﺑﺮﺧﻮردار اﺳﺖ. جمع آوری و دستیابی به اطلاعات مشتریان از یک سو ، کاوش ، تجزیه-تحلیل و تبدیل آن به بینش [[1]](#footnote-1)از سویی دیگر از برجسته ترین امور در دنیای امروزه است . یکی از راهکارهایی که کاوش داده می‌تواند مزیتی رقابتی در کسب و کار بدست آورد ، یافتن مشتریان وفادار می‌باشد. روش های دستیابی به اطلاعات ، داده کاوی و تجزیه-تحلیل آنها با استفاده از تکنیک های یادگیری ماشین و ساخت مدل های پیش بین از موضوعات جذاب و چالش برانگیز است . هدف اصلی این مقاله بکارگیری روش های یادگیری پیشگویانه و پیش‌بینی اینکه کدامین مشتری‌ برای خرید مجددا اقدام می‌نماید است. در واقع در این مقاله معیار ریزش و نوع وفاداری مشتری درنظر گرفته نشده و تنها میزان احتمال وفاداری با روش های پیش نهادی تعیین می گردد ، و نوع وفاداری که با چه احتمالی خریدار از نوع وفادار با ثبات ، نسبی و بی ثبات است، هدف این مقاله نمی‌باشد . به طور کلی پیش بینی مشتریان وفادار با توسعه یک مدل پیش گو آغاز می شود. در این مقاله روش پارتو - توزیع دو جمله ای منفی [[2]](#footnote-2)، روش بتا ژئومتریک - توزیع دو جمله ای منفی[[3]](#footnote-3) و مدل مرتبه ای بیز [[4]](#footnote-4)با یکدیگر مقایسه گردیده و بر روی داده های مسابقه وب سایت کاگل [[5]](#footnote-5) که مختص برگزاری مسابقات چالشی در سراسر دنیا می‌باشد ، آزمایش گردیده است . برای یادگیری و تست مدل ، کدهای برنامه به زبان برنامه نویسی پایتون و در محیط کاری لینوکس با استفاده از کتابخانه VW یا همان Vowpal Wabbit بر روی مجموعه های بدست آمده استفاده گردیده شده .

**کلمات کلیدی :**

**مقدمه :**

طی سال های اخیر ، برنامه های وفاداری به شکل فزاینده‌ایی تبدیل به ابزارهای لازمه بازاریابی در کسب و کارها شده اند. باوجود اینکه جذب مشتریان از طریق تبلیغات همواره عنصری اصلی آمیخته شده با بازاریابی بوده است ، تلاش هایی که برای حفظ مشتریان می شوند هم اهمیت یکسانی دارند .

برنامه های موفق وفاداری مشتری بخاطر اصالت و سابقه شان در دنیای خرده فروشی امروزه گسترش یافته اند . به طور میانگین از کیف پول های مصرف کنندگان میشه دریافت که علاوه بر کارت های هویتی و اعتباری ، به احتمال زیاد چندین کارت پاداش هم دراختیار دارند. مطالعه اخیری توسط بازاریابی وفاداری مارتیز نشان می‌دهد که یک بزرگسال معمولی به طور متوسط همواره عضوی از 4/7 برنامه وفاداری است و این درصورتی است که 71% از پاسخ دهندگان نظرسنجی، برای ثبت نام در برنامه های بیشتری اعلام آمادگی کرده اند.

ﻣﺸﺘﺮی ﺑﻪ ﻋﻨﻮان عاملی ﻛلیدی و ﻣﺤﻮری در ﺑﻘﺎی کسب و کار اﻳﻔﺎی ﻧﻘﺶ می ﻛﻨﺪ وجهت گیری کلیه اهداف، اﺳﺘﺮاﺗﮋی ﻫﺎ و ﻣﻨﺎﺑﻊ ﺣﻮل ﻣﺤﻮر ﺟﺬب و ﻧﮕﻬﺪاری مشتری می ﺑﺎﺷﺪ. ﺣﻔﻆ و ﺗﻘﻮﻳﺖ وفاداری ﻣﺸﺘﺮﻳﺎن ﺑﺮای ﺷﺮﻛﺖ هایی ﻛﻪ دﻏﺪﻏﻪ ﺣﻔﻆ و ﺗﻮﺳﻌﻪ ﺟﺎﻳﮕﺎه رﻗـﺎبتی ﺧﻮﻳﺶ را در ﺑﺎزار دارﻧﺪ، ﭼﺎﻟش اﺳﺘﺮاﺗﮋﻳﻚ تلقی می ﺷﻮد؛ چالشی که با پیش بینی مشتری های وفادار می تواند تا حدی هموار گردد. بی تردید، مهم ترین وجه تمایز یک مشتری وفادار از یک خریدار عادی از دید فروشنده این است که مشتری برای خرید همواره به او مراجعه نماید. در این مقاله بحث بر روی روش ها و تکنیک های یادگیری ماشین و پیش بینی مبتنی بر یادگیری برای دستیابی به این مهم است.

**10 راهکاری که کاوش داده می‌تواند مزیتی رقابتی در کسب و کار بدست آورد**

شرکت‌های بی‌شماری هستند که داده‌های حجیم و مناسبی از مشتریان خود در اختیار دارند ، و متاسفانه هیچ بهره‌ای از آن نمی‌برند. و این واقعا جای تعجب دارد ، چراکه داده‌ایی را دراختیار دارند که معدن طلایی از بینش است .

بینش‌هایی که می‌تواند منجر شود به :

* افزایش وفاداری مشتریان
* بازنمودن سودآوری پنهان
* کاهش ریزش مشتریان

اما این داده‌ها را باید استفاده نمود ؟ دراینجا 10 راهکار عملی را به صورت لیست‌وار مطرح نموده و بر‌روی طرح وفاداری مشتری تمرکز خواهیم نمود . و در ادامه به بررسی روش‌ها ، ابزارها و الگوریتم‌های یافتن مشتریان وفادار و آزمایش دقت هریک از آنها خواهیم پرداخت .

1. **تجزیه-تحلیل سبد خرید**

این مورد برمبنای فرضیاتی است که طبق رفتار مشتری از داده های عملکردی گذشته آن پیش‌بینی شده است ، که این داده‌ها شامل خریدها و صلایق آنان است . و این شیوه در موارد زیر هم قابل استفاده می‌باشد :

1. **پیش بینی فروش**

تمرکز این مورد براین است که مشتری چه هنگام خرید کرده و پیش‌بینی می‌گردد خرید مجدد ایشان چه زمانی باشد . به علاوه اطلاعات خرید تعدادی از مشتریان را مورد بررسی قرارمی‌دهد تا پیش‌بینی نماید که میزان واقعی فروش چه مقدار بوده است .

1. **بازاریابی پایگاه داده**

با آزمایش الگوهای خرید مشتریان ، موقعیت جغرافیایی آنان و روانشانختی مشتریان می‌توان برای هریک پرونده‌ای ایجاد نمود که محصولاتی را تولید کرد که برای فروش یا ارائه خدمت به آنان طراحی شده است .

1. **برنامه‌ریزی تبلیغاتی**

این مورد برای شرکت‌های آنلاین یا آفلاین مفید است . برای شرکتی آفلاین ، که به دنبال افزودن فروشگاهی جدید است ، میزان تبلیغات دقیقی که نیاز دارند طبق اطلاعات فروشگاه‌های موجود بررسی می‌شود. و برای کسب‌وکاری آنلاین برنامه‌ریزی تبلیغاتی می‌تواند در تعیین گزینه‌های انبارداری و موجودی انبار کمک کننده باشد.

1. **کارت بازاریابی**

اگر کسب‌وکاری شامل صدور کارت‌های اعتباری است ، می‌توان با جمع‌آوری اطلاعات میزان استفاده ، مشتریان را بخش‌بندی نمود و بسته به اطلاعات این بخش‌بندی‌ها برنامه‌هایی را ایجاد کرد که به رشد ابقاء مشتری ، بهبود جذب ، هدف گذاری قیمت و طراحی محصولات کمک نمایند .

1. **تجزیه-تحلیل جزئیات سوابق تماس‌ها**

اگر شرکتی وابسته به ارتباطات تلفنی است ،داده‌های رسیده را برای کشف الگوهای استفاده شده و ایجاد پرونده‌های مشتریان از این الگوها می‌توان به‌کار گرفت و پس از آن درجهت به حداکثر رساندن سود‌آوری ، ساختاری طبقه‌بندی شده‌ قیمت‌گذاری را ایجاد نمود . یا پیشنهاداتی را منطبق با داده‌های تحلیل شده ارائه نمود .

1. **وفاداری مشتری**

در دنیایی که جنگ قیمت‌ها رخ می‌دهد ، با مشتریانی مواجه هستید که درهر لحظه به دنبال کمترین قیمت تغییر جهت می‌دهند . و با استفاده از داده کاوی می‌توان این ریزش مشتری را به حداقل رساند .

1. **بخش بندی بازار**

یکی از بهترین استفاده‌های داده کاوی بخش‌بندی مشتریان است و اینکار بسیار ساده است . توسط داده‌ها میتوان آنها را به بخش های معناداری مثل سن ، درآمد ، شغل یا جنسیت بخش‌بندی کرد .

1. **تولید محصول**

همچنین داده‌کاوی برای ایجاد محصولاتی سفارشی برای بخش‌های بازار مناسب می‌باشد . درواقع ، می‌توان پیش‌بینی نمود چه ویژگی‌هایی را مشتریان انتظار دارند . اگرچه نوآوری در تولید محصول همواره در دستور کار خواهد بود .

1. **گارانتی**

درنهایت ، کاوش پایگاه داده اجازه پیش‌بینی این را می‌دهد که چند نفر به گارانتی‌ایی که راه‌اندازی می‌شود واقعا نیاز خواهند داشت. این درمورد ضمانت هم صادق است . برای نمونه طرح گارانتی برای محصولی طراحی شود که وقتی پایگاه‌داده آن تجزیه-تحلیل شده است ، میزان تعداد واقعی افراد برای مراجعه به گارانتی را تخمین بزند .

**وفاداری مشتری**

**هدف : پیدا کردن مشتری‌های وفادار**

کاتلر[[6]](#footnote-6) و آرمسترانگ [[7]](#footnote-7)، رفتار تصمیم گیری خرید مشتری را در یک مدل محرک –پاسخ و مفهوم جعبه سیاه تعبیر می‌کنند .و این رفتار تحت تاثیر عوامل تحریک کننده و انگیزشی به همراه عوامل محرک های بازاریابی (آمیخته بازاریابی) وارد جعبه سیاه مصرف کننده می‌شوند و او واکنش های مشخصی از خود نشان می دهد. جعبه سیاه از دو بخش تشکیل شده است : 1- ویژگی های شخصی خریدار 2- فرآیند تصمیم گیری خریدار .

**ابعاد وفاداری**

مطالعه کیم زینس [[8]](#footnote-8)سه جنبه وفاداری را مشخص ساخت :

1. نگرشی
2. رفتاری
3. ترکیبی ( نگرشی و رفتاری )

هدف مسئله پیش‌بینی این است که کدام مشتری‌ها برای خرید مجددا اقدام می‌نمایند. توجه شود که در این مقاله مقصود تعیین نوع وفاداری مشتری نیست ، بلکه هدف بررسی وفادار بودن یا نبودن یک مشتری می‌باشد .

انواع مختلفی از وفاداری وجود دارد :

1. **وفاداری با‌ثبات :** مشتریانی که همیشه فقط یک محصول یکسان را خریداری می‌نمایند.
2. **وفاداری نسبی یا موقت :** خریدارانی که به دو یا سه محصول وفادار می‌مانند .
3. **وفاداری بی‌ثبات :** مشتریانی که پس از چند بار خرید از یک مارک یا محصول ، محصولی دیگر را خریداری می‌کنند.

در واقع در این مقاله تنها میزان احتمال وفاداری یک فرد با روش های موجود تعیین می گردد ، اما آنکه اگر مشتری وفادار است کدام یک از رفتار های وفاداری را داراست (و به عبارتی به چه احتمالی خریدار از نوع وفادار با ثبات ، نسبی و بی ثبات است، هدف این مقاله نمی‌باشد . در حالت کلی، این موضوع می تواند مسئله ای پیچیده تر و چالش بر انگیز تر از مسئله کنونی باشد و در حوزه مدل های توصیفی پیشگو [[9]](#footnote-9)قرار می گیرد.

**روش‌های یادگیری مدل در داده کاوی**

همانگونه که می‌دانیم مراحل کاری در داده کاوی شامل سه مرحله اصلی است : آماده سازی داده، یادگیری مدل و در نهایت ارزیابی و تفسیر مدل.

در مرحله یادگیری مدل با استفاده از الگوریتم‌های متنوع و با در نظر گرفتن ماهیت داده ، به شناسایی الگوهای منظم موجود در داده‌ها پرداخته می‌شود. بطور کلی روش‌های مختلف کاوش داده را به دو گروه روش‌های پیش بینی و روش‌های توصیفی طبقه بندی می‌کنند.

در روش‌های پیش بینی از مقادیر بعضی ویژگی‌ها برای پیش بینی کردن مقدار یک ویژگی مشخص استفاده می‌کنند. این روش‌ها در متون علمی با نام روش‌های نظارت شده[[10]](#footnote-10) نیز شناخته می‌شوند. الگوریتم‌های نظارت شده از دو مرحله با عنوان مرحله آموزش [[11]](#footnote-11)(یادگیری) و مرحله ارزیابی[[12]](#footnote-12) تشکیل شده اند.

در مرحله آموزش؛ با استفاده از مجموعه داده‌های آموزشی [[13]](#footnote-13)مدل ساخته می‌شود. شکل مدل ساخته شده به نوع الگوریتم یادگیرنده بستگی دارد.

در مرحله ارزیابی؛ از مجموعه داده‌های آزمایشی[[14]](#footnote-14) برای اعتبارسنجی و محاسبه دقت مدل ساخته شده استفاده می‌شود که در واقع در مرحله آموزش و ساخت مدل الگوریتم این مجموعه داده‌ها را در نظر نگرفته است [[15]](#footnote-15).

برای نمونه روش‌های دسته بندی [[16]](#footnote-16)، رگرسیون [[17]](#footnote-17)و تشخیص انحراف [[18]](#footnote-18)سه روش یادگیری مدل در داده کاوی با ماهیت پیش بینی هستند.

در روش‌های توصیفی همانطور که انتظار داریم الگوهای قابل توصیف از روابط حاکم بر داده‌ها بدون در نظر گرفتن هر گونه برچسب و یا متغیر خروجی بدست می‌آید. این روش‌ها در متون علمی با نام روش‌های نظارت نشده [[19]](#footnote-19)نیز شناخته می‌شوند. برای نمونه روش‌های خوشه بندی [[20]](#footnote-20)، کاوش قوانین انجمنی [[21]](#footnote-21)و کشف الگوهای ترتیبی [[22]](#footnote-22)سه روش یادگیری مدل در داده کاوی با ماهیت توصیفی هستند.

در ادامه به معرفی اجمالی هر کدام از این روش‌ها می‌پردازیم:

**دسته بندی:**

در الگوریتم‌های دسته بندی مجموعه داده اولیه به دو مجموعه داده با عنوان مجموعه داده‌های آموزشی و مجموعه داده‌های آزمایشی تقسیم می‌شود. می‌دانیم هر مورد شامل مجموعه ای از ویژگی‌هاست ، که یکی از این خصوصیات ویژگی دسته نامیده می‌شود. در مرحله آموزش؛ مجموعه داده‌های آموزشی به یکی از الگوریتم‌های دسته بندی داده می‌شود تا بر اساس سایر ویژگی‌ها برای مقادیر ویژگی دسته، مدل ساخته شود. پس از ساخت مدل، در مرحله ارزیابی؛ دقت مدل ساخته شده به کمک مجموعه داده‌های آزمایشی ارزیابی خواهد شد. در الگوریتم‌های دسته بندی از آنجا که ویژگی دسته مربوط به هر مورد مشخص است به صورت الگوریتم‌های نظارت شده محسوب می‌شوند. بدیهی است که تشخیص بر اساس دسته هایی است که مدل در مرحله آموزش با آنها روبرو شده است؛ بنابراین امکان تشخیص دسته جدید در کاربرد دسته بندی وجود نخواهد داشت.

**رگرسیون:**

رگرسیون در علوم آمار و شبکه‌های عصبی بطور وسیعی مورد بررسی و مطالعه قرار می‌گیرد. پیش بینی مقدار یک متغیر پیوسته بر اساس مقادیر سایر متغیرها بر مبنای یک مدل وابستگی خطی یا غیر خطی رگرسیون نامیده می‌شود. یک نوع خاصی از رگرسیون، پیش بینی سری‌های زمانی[[23]](#footnote-23) است؛ برای مثال تغییرات قیمت سهام شرکتی را به صورت نمودار داریم؛ می‌خواهیم ادامه روند این نمودار را برای مدتی مشخص پیش بینی کنیم. در مسائل سری‌های زمانی یکی از متغیرهای اصلی زمان می‌باشد. بدیهی است که رگرسیون لزوماً سری زمانی نیست و همانند دسته بندی کاربرد رگرسیون نیز از نوع پیش بینی نظارت شده است و بطور مشابه در رگرسیون هم دو مرحله آموزش و ارزیابی وجود دارد. مثال هایی از کاربرد رگرسیون می‌تواند شامل موارد زیر باشد : پیش بینی میزان فروش محصولی جدید، براساس میزان فروش محصولات در گذشته و یا براساس میزان تبلیغات صورت یافته ، همچنین مسائل مربوط به پیش بینی سری‌های زمانی از قبیل بورس و ... .

**تشخیص انحراف:**

از کاربردهای متداول تشخیص انحراف، می‌توان به کشف کلاهبرداری کارت‌های اعتباری [[24]](#footnote-24) اشاره کرد. در مواقعی از این کاربرد استفاده می‌شود که تنها نمونه هایی با یک برچسب یکسان که معمولاً وضعیت نرمال را نشان می‌دهند در دسترس باشد و امکان دراختیار داشتن داده‌ها با تمامی برچسب‌های موجود به دلایل مختلف وجود ندارد. بنابراین چون فقط نمونه‌های دسته نرمال در اختیار است، الگوریتم برای وضعیت نرمال و با توجه به یک آستانه[[25]](#footnote-25) مشخص ، مدل را می‌سازد و هر گونه تخطی از آن آستانه را بعنوان وضعیت غیرنرمال در نظر می‌گیرد. توجه شود روش‌های دسته بندی تنها قادر به شناسائی دسته هایی هستند که در مرحله آموزش، نمونه ای از آنها به الگوریتم ارائه شده است، بنابراین امکان تشخیص هیچ گونه کلاهبرداری توسط روش‌های دسته بندی وجود ندارد.

**خوشه بندی:**

در این مسائل از آنجا که بر خلاف دسته بندی هیچ گونه دسته خاصی وجود ندارد، بنابراین براساس معیار شباهت داده‌ها گروه بندی و خوشه بندی صورت می‌گیرد. بدین ترتیب مورد‌هایی که بیشترین شباهت را به یکدیگر دارند در یک خوشه قرار می‌گیرند، به بیان دیگر موردهای موجود در خوشه‌های متفاوت ، کمترین شباهت را به یکدیگر خواهند داشت. بدیهی است که خوشه بندی براساس ویژگی ورودی نمونه‌ها انجام می‌گیرد و از آنجائی که برای این الگوریتم‌ها ویژگی دسته تعریف نمی‌شود و موردها برچسب خاصی ندارند، جزء الگوریتم‌های بدون ناظر محسوب می‌شوند. در واقع هدف در تمامی الگوریتم‌های خوشه بندی کمینه کردن فاصله درون خوشه ای[[26]](#footnote-26) و بیشینه نمودن فاصله بین خوشه ای[[27]](#footnote-27) است و عملکرد خوب یک الگوریتم خوشه بندی زمانی محرز می‌شود که تا حد امکان خوشه‌ها را از یکدیگر دورتر کند و در ضمن موردهای موجود در یک خوشه بیشترین شباهت را به یکدیگر داشته باشند.

**کشف قوانین انجمنی:**

قوانین وابستگی یا انجمنی، اتفاق و وقوع یک شیء را براساس وقوع سایر اشیاء توصیف می‌کنند، برای مثال در یک سوپر مارکت هدف در کاوش قوانین انجمنی؛ یافتن نظم حاکم بر سبد خرید می‌باشد، در این کاربرد به ازای هر سبد؛ یک قانون پیدا می‌شود و بررسی خواهد شد که این قانون در چه تعداد از سبدها صدق می‌کند و در نهایت مجموعه قوانینی که در اکثر سبدها صدق می‌کند به عنوان مجموعه قوانین انجمنی خروجی ارائه می‌شود. به بیان دیگر در این کاربرد به دنبال پیدا کردن مجموعه‌ایی از قوانین وابستگی هستیم تا براساس آن قوانین بتوانیم نتیجه گیری کنیم وجود کدامیک از مجموعه اشیاء[[28]](#footnote-28) بر وجود چه مجموعه اشیاء دیگری تاثیر گذار است.

**کشف الگوهای ترتیبی:**

در این کاربرد به دنبال کشف الگوهایی هستیم که وابستگی‌های ترتیبی محکمی را در میان وقایع مختلف نشان می‌دهند. این کاربرد مشابه کاوش قوانین انجمنی می‌باشد با این تفاوت که در کاوش قوانین انجمنی زمان و ترتیب زمانی مطرح نیست، اما در کشف الگوهای ترتیبی ، زمان و ترتیب آن اهمیت ویژه ای دارند برای مثال می‌توان به دنباله‌های تراکنش‌های فروش اشاره نمود.

**مدل‌های مورد بررسی**

مدل‌دهی پیشگویانه از آمار برای پیش‌بینی نتایج استفاده می‌کند . در اکثر مواقع رویدادی که قراراست پیش‌بینی شود در آینده وجود دارد ، ولی مدل‌دهی پیشگویانه می‌تواند برروی هرنوعی از وقایع ناشناخته اعمال شود ، بدون درنظر گرفتن اینکه چه زمان اتفاق افتاده است . برای نمونه ، اغلب مدل‌های پیشگویانه برای تشخیص جرم و تعیین مضمون پس از ارتکاب به جرم ، به کار برده شده است .

در بسیاری از موارد ، مدل برمبنای تئوری تشخیص انتخاب شده است تا با مقادیر داده ورودی که به آن داده شده است سعی بر حدس زدن احتمال نتیجه‌ کند ، برای نمونه بکارگیری آنها برای برای تعیین اینکه یک نامه الکترونیکی چقدر شبیه اسپم تشخیص داده می شود .

مدل‌ها می‌توانند یک یا چندین طبقه‌بند [[29]](#footnote-29) را در تعیین احتمال اینکه مجموعه‌ایی از داده‌ها متعلق به مجموعه ایی دیگر باشند را به کار بگیرند .

بسته به مرزهای مشخص شده ، مدل‌های پیش بینی به طور گسترده‌ای با رشته یادگیری ماشین هم‌ رده هستند یا همپوشانی دارند ، همانطور که اغلب بیشتر در تحصیلات دانشگاهی یا زمینه تحقیقاتی و توسعه اشاره شده است . از زمانیکه به صورت تجاری گسترده شد ، مدل دهی پیش بین اغلب به عنوان تجزیه-تحلیل پیش بین اشاره شده است .

**مدل های استفاده شده**

تقریبا هر مدل رگرسیونی می‌تواند برای اهداف پیش‌بینی استفاده شود . به طور کلی دو دسته مدل پیش‌بینی داریم : پارامتریک [[30]](#footnote-30)و غیر پارامتریک [[31]](#footnote-31). سومین دسته ، مدل های نیمه پارامتریک [[32]](#footnote-32)، شامل ویژگی هایی از هر دو دسته هستند . مدل‌های پارامتریک مفروضات خاصی را با توجه به یک یا چندین پارامتر تراکمی که توزیع های وابسته به آن را شکل می‌دهند ، می‌سازند ، درصورتیکه رگرسیون‌های غیر پارامتریکی مفروضات کمتری را نسبت به همتایان پارمتریک خودشان ایجاد می‌نمایند .

در اینجا بدلیل اینکه از مسیر اصلی این مقاله خارج نشویم و تمرکز اصلی را بر روی مدل های پیشنهادی خود بگذاریم ، فقط به ذکر نام انواع مدل‌های موجود بسنده می‌کنیم .

1. Group method of data handling
2. Naive Bayes
3. k-nearest neighbor algorithm
4. Majority classifier
5. Support vector machines
6. Random forests
7. Boosted trees
8. CART (Classification and Regression Trees)
9. MARS [[33]](#footnote-33)
10. Neural Networks
11. ACE and AVAS
12. Ordinary Least Square
13. Generalized Linear Models (GLM)
14. Logistic regression
15. Generalized additive models
16. Robust regression
17. Semiparametric regression

به طور کلی دو مدل توصیفی و پیشگو در حوزه تجزیه-تحلیل‌های پیشگویانه قرار دارند. به طور کلی پیش بینی مشتریان وفادار با توسعه یک مدل پیش گو آغاز می شود. اولین گام در توسعه چنین مدلی جمع آوری داده هایی خاص در مورد هر مشتری است که ماهیت این داده ها در ادامه توضیح داده خواهد شد. و این داده ها برای برآورد نمودن احتمال خرید مجدد مشتری در یک بازه زمانی معین در آینده از فروشنده ای خاص ، کفایت می‌کند.

**کارهای مرتبط**

تا به این جا سعی شد مقدمه کاملی از مسئله پیش روی ، هدف و اهمیت آن تشریح گردد. همچنین مدل مورد بررسی برای حل مسئله بیان شد. اکنون به معرفی روش های مختلفی می پردازیم که مسئله مورد نظر به کمک آنها حل خواهد شد.

**اصل پارتو [[34]](#footnote-34)**

یکی از مهمترین اصول ، اصل پارتو است ، که با نام‌های قانون ۸۰–۲۰ ، قانون وابستگی حیات اندک[[35]](#footnote-35) و اصل عامل تُنُکی[[36]](#footnote-36) نیز شناخته می‌شود بیان می‌کند که ۸۰ درصد رخدادها از ۲۰ درصد دلایل بوجود می‌آید. جوزف ام جوران [[37]](#footnote-37)که شغل مشاور مدیریتی داشته این اصل را ارائه داده و نام آن‌را از نام ویلفردو پارتو[[38]](#footnote-38)، اقتصاددان ایتالیایی گرفته که در سال ۱۹۰۶ ، دریافت۸۰ درصد زمین‌های ایتالیا در دست ۲۰ درصد مردم آن کشور است؛ همچنین پارتو دریافته بود که ۸۰ درصد نخودفرنگی‌های باغچه‌اش در ۲۰ درصد غلاف‌های نخودفرنگی قرار دارند.

اصل پارتو یک قاعده سرانگشتی رایج در تجارت است؛ برای مثال «۸۰ درصد فروش‌ها مربوط به ۲۰ درصد مشتریان است.» از لحاظ ریاضی اصل ۸۰–۲۰ از توزیع قانون توان با مجموعه پارامترهای ویژه پیروی می‌کند (با نام توزیع پارتو نیز شناخته می‌شود) و بسیاری از پدیده‌های طبیعی از این توزیع پیروی می‌کنند.

اصل پارتو با کارایی پارتو نیز در ارتباطی نزدیک است. پارتو از هردوی این مفاهیم در مفهوم توزیع درآمد و ثروت میان جمعیت استفاده کرد.

**روش پارتو - توزیع دو جمله ای منفی**

پیچیده ترین و مرسوم ترین روشی که برای یافتن مشتریان وفادار مورد استفاده قرار گرفته است ، چارچوبی به نام پارتو- توزیع دو جمله ای منفی است. این روش در سال 1987 توسط شخصی به نام اشمیت لین[[39]](#footnote-39) معرفی گردید . اگرچه این روش از قدرتمند ترین چارچوب ها برای انجام پیش بینی است، اما پیاده سازی این مدل در زمان تخمین پارامتر ها بسیار دشوار و چالش برانگیز است.

**بررسی روش**

در این چارچوب رفتار خرید مشتری مدل می شود ولی ریزش مشتری نادیده گرفته می شود. بدین معنا که در این چارچوب فرض بر این است که مشتریان در یک بازه زمانی مشخص با یک نرخ ثابت اما به صورت تصادفی خرید می کنند و سپس غیر فعال می شوند. منظور از غیر فعال شدن مشتری در این روش آن است که فرآیند خرید با یک نرخ ثابت از این پس ادامه پیدا نمی کند. بدیهی است که در مقابل فعال بودن مشتری به معنای آن است که مشتری از دست نرفته است و هنوز عمل خرید را با یک نرخ معینی انجام می دهد.

لازم به ذکر است که در این روش زمان از دست رفتن یا ریزش با استفاده از مدل زمانی پارتو (ترکیب نمایی-گاما ) بیان می شود. همیچنین رفتار خرید مکرر به هنگام فعال بودن با استفاده از مدل شمارشی توزیع دوجمله ای منفی ( ترکیب پواسن- گاما ) شبیه سازی می گردد. به بیان ساده تر، در این روش برای آنکه مشخص شود آیا مشتری از دست رفته است یا خیر از یک سکه استفاده می کنیم و سپس از یک تاس تا مشخص شود که یک مشتری چه تعداد از اجناس یا اقلام را سفارش خواهد داد.

**فرض ها**

مدل پارتو – توزیع دوجمله ای منفی مبتنی بر پنچ فرض اساسی است که درک آنها کمک زیادی به حل مسئله خواهد کرد:

1. تا زمانی که یک مشتری فعال است، تعداد تراکنش هایی که در یک بازه زمانی به طول t انجام می دهد یک توزیع پواسن با میانگین λt است.
2. ناهمگنی[[40]](#footnote-40) در نرخ تراکنش λ در بین مشتریان دارای یک توزیع گاما با پارامتر شکل[[41]](#footnote-41) r و پارامتر مقیاس[[42]](#footnote-42) α است.
3. هر مشتری طول عمری برابر با τ دارد. این نقطه از زمان که مشتری در آن غیر فعال می شود یک توزیع نمایی با نرخ ریزش ای برابر μ است.
4. ناهمگنی در نرخ های ریزش در بین مشتریان دارای یک توزیع گاما با پارمتر شکل s و پارامتر مقیاس β است.
5. نرخ تراکنش λ و نرخ ریزش μ به طور مستقل از هم در بین مشتریان تغییر می کنند.

**شرح روش**

همانطور که اشاره شد، در هر مدلی از جمله مدل پارتو- توزیع دو جمله ای منفی دانستن آنکه در یک بازه معین، زمان آخرین تراکنش چه زمانی بوده و نیز چه تعداد تراکنش در آن بازه انجام شده است الزامی است.

نمادی که برای نشان دادن این موضوع استفاده می شود به صورت است که در آن x تعداد تراکنش های انجام شده در بازه زمانی می باشد و زمان انجام آخرین تراکنش است . با در دست داشتن این اطلاعات می توان موارد زیر را به بدین صورت تعریف کرد و آنها را محاسبه نمود:

1. مقدار برابر است با تعداد تراکنش های مورد انتظار در یک بازه زمانی به طول t
2. مقدار برابر است با احتمال آنکه در یک بازه زمانی به طول t ، تعداد تراکنش هایی برابر x انجام شود. در واقع احتمال آنکه در زمان x ،t تراکنش انجام شود.
3. مقدار برابر است با تعداد تراکنش های مورد انتظار در یک بازه زمانی به طول برای یک نفر با رفتار

در ادامه با در دست داشتن مقادیر فوق و انجام محاسبات ریاضی مربوطه ، تابع احتمال مدل پارتو - توزیع دو جمله ای منفی بصورت رابطه زیر بدست خواهد آمد:

توجه نمایید که هر یک از پارامتر های موجود در رابطه فوق، در قسمت فرض ها بیان شده است. اکنون لازم است تا ضریب ​​ مشخص شود. در صورتی که آنگاه ضریب ​​ به صورت زیر خواهد بود :

همچنین در صورتی که آنگاه ضریب ​​ به صورت زیر می باشد :

بدیهی است که در حالت تساوی متغیر های α و β ، هر دو رابطه مربوط به محاسبه ضریب ​​ با یکدیگر برابر خواهند شد. در هر یک از روابط معادله 1 و معادله 2 تابع ​​ ، تابع فوق هندسی گاوس [[43]](#footnote-43) است که در معادله 1 بیان شده است. همانطور که مشاهده می کنید، تابع احتمال مدل پارتو - توزیع دو جمله ای منفی بسیار پیچیده و شامل چندین تقریب از تابع هندسه فضایی گاوس است. به همین دلیل امروزه، عموما از این روش در عمل استفاده نمی شود. در واقع، اگر چه این روش یکی از قدرتمند ترین روش ها برای انجام پیش بینی های لازم در زمینه وفاداری مشتریان است، اما پیاده سازی این مدل به دلیل پیچیدگی فرمول هایش بسیار دشوارتر از سایر روش های موجود است. در این میان، برخی از این روش ها می توانند با همان دقت و در عین حال با پیچیدگی کمتری خروجی را تولید کنند. در ادامه چنین روش هایی بیان خواهند شد.

شایان ذکر است که در همه ی روش هایی که در ادامه مطرح می شود، بسیاری از فرض های موجود در روش پارتو - توزیع دو جمله ای منفی باز هم صادق است و بنابراین از تکرار توضیح آنها پرهیز می شود. در واقع، در هر روش ضمن تبیین نکات متمایز، نکات مشابه نیز بیان می شود اما تنها موارد اختلافی توضیح داده خواهند شد.

**روش بتا ژئومتریک - توزیع دو جمله ای منفی**

روش بتاژئومتریک – توزیع دو جمله ای منفی به طور کلی بیشتر جنبه های پارتو – توزیع دو جمله ای منفی را داراست و تنها تغیرات کوچکی اعمال شده است. اما همین تغییرات کوچک باعث می شود که فرمول ها بسیار ساده تر از روش قبل گردند و پارامتر های مختلف به آسانی بدست آیند. با این حال عموما نتایج گرفته شده بسیار به روش قبل نزدیک اند که شاید همین موضوع دلیلی بر ترجیج این روش بر روش قبلی باشد.

**بررسی روش**

تنها تفاوت با روش پارتو- توزیع دو جمله ای منفی در این جاست که چگونه و چه زمانی مشتری ها غیر فعال می شوند. همان طور که پیشتر نیز توضیح داده شد، منظور از غیر فعال شدن مشتری در یک زمان خاص آن است که از آن زمان به بعد مشتری به انجام خرید با یک نرخ ثابت ادامه نمی دهد چرا که اصولا در این دو روش فرض بر این است که همواره مشتری با یک نرخ معین و ثابت به طور مداوم ( هر چند به صورت تصادفی ) اقدام به خرید می کند. در واقع، در مدل پارتو-توزیع دو جمله ای منفی فرض می شود که ریزش در هر نقطه ای از زمان، مستقل از زمان رخداد واقعی خرید ها، می تواند روی دهد. حال اگر در عوض فرض کنیم که ریزش بلافاصله پس از هر خرید روی می دهد می توانیم این فرآیند را با استفاده از یک مدل بتا ژئومتریک شبیه سازی کنیم.

**فرض ها**

مدل بتاژئومتریک - توزیع دوجمله ای منفی مبتنی بر پنچ فرض اساسی زیر است:

1. تا زمانی که یک مشتری فعال است، تعداد تراکنش هایی که در یک بازه زمانی به طول t انجام می دهد یک توزیع پواسن با نرخ تراکنش λ است. این معادل آن است که فرض کنیم زمان بین تراکنش ها دارای توزیع نمایی با نرخ تراکنش λ است. یعنی:
2. ناهمگنی در λ دارای یک توزیع گاما به صورت زیر است:
3. پس از هر تراکنش، مشتری با احتمال p غیر فعال می شود. به عبارتی، احتمال آنکه بلافاصله پس از تراکنش j ام مشتری غیر فعال شود به صورت زیر است:
4. ناهمگنی در p دارای یک توزیع بتا به صورت زیر است:

که در آن B، همان تابع بتا است و برابر می باشد.

1. نرخ تراکنش λ و احتمال ریزش p به طور مستقل از هم در بین مشتریان تغییر می کنند.

**شرح روش**

فرض کنید یک مشتری x تراکنش در بازه زمانی انجام داده باشد به طوری که تراکنش ها به ترتیب در زمان های ​​ انجام شده باشند. شکل زیر به خوبی چنین موضوعی را نشان می دهد.



تراکنش ها در بازه زمانی مفروض

حال با درنظر گرفتن موارد زیر تابع احتمال ، به سادگی بدست خواهد آمد :

* احتمال آنکه اولین تراکنش در زمان ​​ رخ دهد، یک توزیع نمایی استاندارد به صورت است.
* احتمال وقوع دومین تراکنش در زمان ​​ برابر احتمال فعال باقی مانده در ​​، ضرب در توزیع نمایی استاندارد مربوطه است که برابر است با :

و این روند همین طور برای هر تراکنشی ادامه می یابد تا آنکه به تراکنش x ام می رسیم:

* احتمال آنکه تراکنش x ام در مکان ​​ روی دهد برابر احتمال فعال باقی مانده در ​​، ضرب در توزیع نمایی استاندارد مربوطه است که برابر است با:
* در نهایت، احتمال آنکه در بازه زمانی هیچ خریدی صورت نگیرد برابر احتمال آن است که یا مشتری در زمان ​​ غیر فعال شود و یا آنکه فعال بماند، اما در این بازه خریدی انجام ندهد که برابر است با:

در این روش نیز همانند روش پارتو - توزیع دو جمله ای منفی، فرض می شود که همه مشتریان در ابتدای کار فعال هستند لذا تابع احتمال برای خریداری که در بازه هیچ خریدی انجام نمی دهد به صورت تابع بقا [[44]](#footnote-44) نمایی استاندارد یعنی است. از این رو، طبق آنچه که گفتیم می توان نتیجه گرفت تابع احتمال در روش بتاژئومتریک - توزیع دو جمله ای منفی به صورت زیر خواهد بود :

در رابطه فوق برای هر داریم و در غیر این صورت مقدار ​​ صفر است.

**مدل مرتبه ای بیز**

مدل مرتبه ای بیز یک مدل آماری است که در چند سطح مختلف به صورت سلسله مراتبی نوشته می شود و اصولا هدفش آن است که پارامتر های توزیع شرطی پسین [[45]](#footnote-45)را به کمک تئوری بیز برآورد کند. در واقع در این روش، زیر مدل ها با یکدیگر ترکیب می شوند تا یک مدل سلسله مراتبی ایجاد گردد. سپس از الگوریتم بیز برای همگام سازی آنها با داده های مشاهده شده استفاده می شود که در نتیجه ی آن توزیع احتمال پسین ایجاد می گردد که گاه از آن تحت عنوان برآورد احتمال بروز رسانی شده[[46]](#footnote-46) نیز یاد می شود. ذکر این نکته نیز خالی از لطف نیست که عموما مدلسازی مرتبه ای هنگامی استفاده می شود که اطلاعات در سطوح مختلف واحد های مشاهده قرار داشته باشند. در چنین شرایطی، تجزیه و تحلیل سلسله مراتبی به درک بهتر مسئله به ویژه مسئله های چند پارامتری کمک خواهد کرد .

**مراحل**

به طور کلی برای نگرش بیزی به مسئله، نخست باید مراحل زیر در نظر گرفته شود:

* دانش موجود در مسئله را به طور احتمالاتی فرموله کنیم. برای این کار باید مقادیر کیفی دانش را به صورت توزیع احتمال، فرضیات استقلال و غیره مدل کرد. چنین مدلی دارای پارامتر های ناشناخته ای خواهد بود که برای هر یک از مقادیر ناشناخته توزیع احتمال اولیه ای در نظر گرفته می شود که بازگو کننده باور ما به محتمل بودن هر یک از این مقادیر بدون دیدن داده است. این توزیع احتمال اولیه را در اصطلاح توزیع احتمال پیشین [[47]](#footnote-47) می نامند.
* سپس با جمع آوری داده، بررسی و مشاهده آن مقدار توزیع احتمال ثانویه که توزیع احتمال پسین نیز نامیده می شود محاسبه می گردد.
* حال با استفاده از توزیع احتمال پسین به یک نتیجه گیری در مورد عدم قطعیت می رسیم و با میانگین گیری روی مقادیر احتمال ثانویه پیش بینی انجام می دهیم.

**اجزا**

مدل مرتبه ای بیز از دو مفهوم مهم در استخراج توزیع پسین یا ثانویه استفاده می کند:

* **ابر پارامتر[[48]](#footnote-48):** پارامتر موجود در توزیع پیشین را ابر پارامتر می نامند.
* **توزیع احتمال ابر پیشین[[49]](#footnote-49):** توزیع پارامتر موجود در توزیع پیشین را توزیع احتمال ابر پیشین گویند.

فرض کنید متغیر تصادفی Y دارای یک توزیع نرمال با میانگین θ و واریانس یک باشد. به عبارتی .حال هنگامی که می گوییم پارامتر θ یک توزیع پیشین دارد بدان معناست که دارای یک توزیع نرمال با میانگین μ و واریانس یک است یعنی . از طرفی پارامتر μ دارای یک توزیع دیگر است برای مثال توزیع نرمال . در چنین شرایطی پارامتر μ ابر پارامتر نامیده می شود و توزیع آن، که در این حالت همان است را توزیع احتمال ابر پیشین می گویند.

**چارچوب**

اکنون به تشریح چارچوب کلی مورد استفاده در مدل مرتبه ای بیز می پردازیم.

فرض کنید ​​یک مشاهده و ​​ پارامتر کنترل کننده فرآیند تولید ​​ باشد. همچنین فرض کنید به طور تصادفی و جابجا پذیر از یک جامعه آماری مشترک با توزیع کنترل شونده توسط ابر پارامتر انتخاب گردند. توجه شود که در اینجا پارامتر های و θ همانند هر پارامتر دیگری تنها دو متغیر تصادفی هستند. در این صورت مدل مرتبه ای بیز شامل سلسله مراتب زیر خواهد بود:

در ‌مراحل ‌فوق ‌احتمال همان احتمال مطلوب می باشد و منظور از همان توزیع احتمال پیشین یا اولیه مربوطه است. بدیهی است که تابع احتمال مطلوب تنها از طریق متغیر تصادفی به وابسته است. حال با استفاده از قضیه بیز می دانیم:

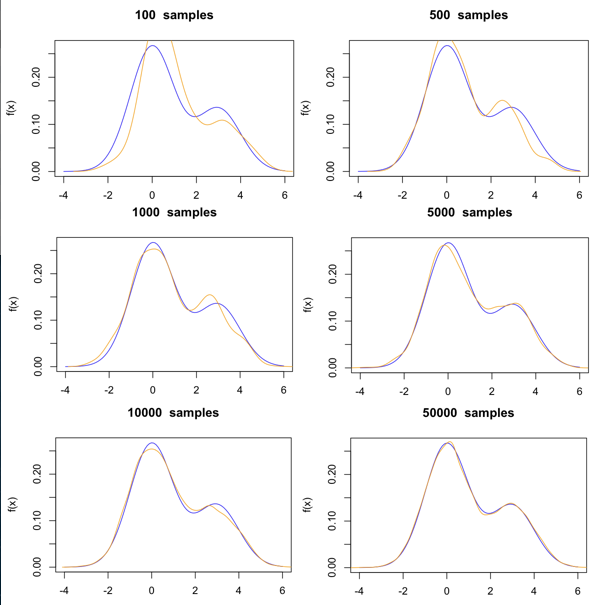
که در آن ابر پارامتر و توزیع احتمال ابر پیشین می باشد. لذا احتمال پسین یا ثانویه به صورت زیر خواهد بود:

اکنون که احتمال ثانویه بدست آمده است با میانگین گیری روی مقادیر احتمال آن می توان پیش بینی های لازم را انجام داد.

**بررسی و شرح روش**

پیش از آنکه ادامه دهیم، ابتدا لازم است با زنجیره مارکوف - مونت کارلو [[50]](#footnote-50) آشنا شویم.

زنجیره مارکوف – مونت کارلو روشی است که از آن برای نمونه برداری از یک توزیع احتمال استفاده می شود که در آن یک زنجیره مارکوف که دارای توزیع مورد نظر است طی چندین مرحله ساخته می شود. گفتنی است که این توزیع مورد نظر را در اصطلاح توزیع تعادل زنجیره می نامند. در نهایت حالت زنجیره پس از چندین مرحله به عنوان نمونه ای از توزیع مطلوب در نظر گرفته می شود و هر چه تعداد گام های بیشتری برای نمونه برداری طی شود، کیفیت تقریب افزایش خواهد داشت. در شکل زیر ، زنجیره مارکوف – مونت کارلو، توزیع آبی رنگ را با توزیع نارنجی تقریب می زند.

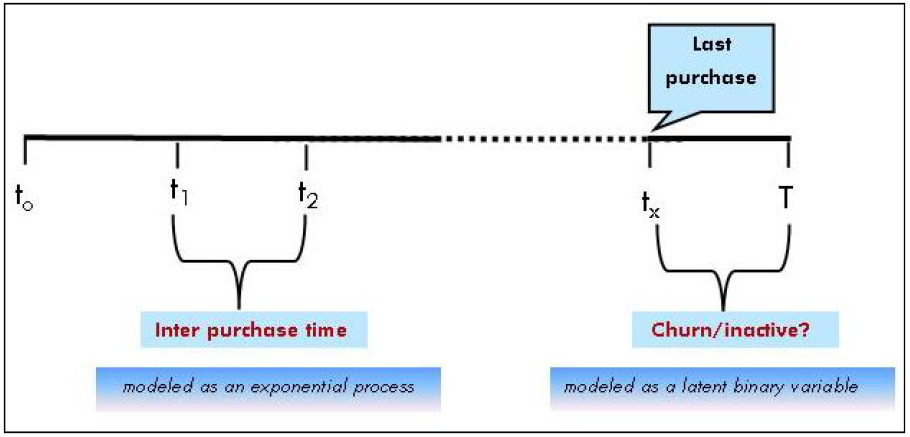


تقریب با استفاده از زنجیره مارکوف - مونت کارلو

همانگونه که مشاهده می شود، با افزایش تعداد نمونه ها، تعداد گام های بیشتری در زنجیره طی می شود و بنابراین خطای تقریب کاهش می یابد.

اکنون که با مفهوم زنجیره مارکوف - مونت کارلو آشنا شدیم، شرح مدل مرتبه ای بیز را ادامه می دهیم. همانگونه که پیشتر نیز توضیح داده شد، تعداد تراکنش های انجام شده توسط مشتری j ام دارای یک توزیع پواسن با نرخ است. همچنین فرض کنید احتمال ریزش پس از هر بار خرید برابر در نظر گرفته شود.

ایده کلی مدل مرتبه ای بیز برای پیدا کردن مشتری های وفادار در شکل زیر نشان داده شده است.



مدلسازی فرکانس تراکنش ها

در شکل فوق، ​​ زمان وقوع اولین تراکنش، ​​ زمان وقوع آخرین تراکنش و x تعداد تراکنش های انجام شده است. گفتنی است که این سه داده می توانند برای نمونه گیری از توزیع های شرطی در زنجیره مارکوف - مونت کارلو کافی باشند. بدیهی است که هر یک از داده های اخیر به راحتی از داده های مربوط به پیشینه تراکنش های مشتریان قابل استخراج می باشد. اکنون با میانگیری روی مقادیر احتمال پسین یا ثانویه (که پیشتر بدست آمد) و نمونه گیری به کمک زنجیره مارکوف - مونت کارلو، هر یک از مقادیر ​​ و ​​ برآورد می گردند. بنابراین احتمال آنکه مشتری j ام در k بازه زمانی بعدی خریدی انجام دهد برابر است با:

شایان ذکر است که مدل مرتبه ای بیز می تواند با دقتی خوب مشتریانی که مجددا برای خرید باز می گردند را شناسایی کند. با این حال هنگامی که تعداد مشتریان بسیار زیاد شود، این روش در عمل کارایی خود را از دست می دهد زیرا برای هر مشتری، زنجیره مارکوف - مونت کارلو باید صد ها بار تکرار شود. به همین دلیل معمولا از این روش هنگامی استفاده می شود که تعداد مشتریان چندان زیاد نباشد.

اما در صورتی که تعداد مشتریان بسیار زیاد باشد، روش مرتبه ای بیز با استفاده از تقریب مبتنی بر رگرسیون [[51]](#footnote-51) توسعه داده می شود. در مدل توسعه یافته دو چند جمله ای با استفاده از اصول رگرسیون ساخته می شوند و از آنها برای تخمین پارامتر های و استفاده می گردد. بدین منظور، ابتدا پارمتر های و ​​ با استفاده از مدل مرتبه ای بیز برای تعدادی معین و محدود از مشتری ها تخمین زده می شوند و سپس لگاریتم برآورد این پارامتر ها به عنوان ورودی در چند جمله های یاد شده قرار می گیرد. هنگامی که مقادیر پارامتر های ​​ و برای هر مشتری به صورت منحصر به فرد از مدل مرتبه ای بیز توسعه یافته با تقریب رگرسیون بدست آمد، به طور مشابه در رابطه (B.H.M) قرار می گیرند تا احتمال آنکه هر مشتری در k بازه زمانی بعدی حداقل یک خرید انجام دهد بدست آید. لازم به ذکر است که استفاده از تقریب مبتنی بر رگرسیون در حالتی که جامعه آماری مورد مطالعه بسیار بزرگ باشد، می تواند حدود 180 برابر سریع تر از مدل مرتبه ای بیز ابتدایی عمل کند .

**آزمایش‌ها**

تا به این جا سعی شد مسئله و راهکار های معمول حل آن به خوبی تشریح گردد. همچنین چالش های موجود در هر روش ( از جمله تخمین پارامتر ها ) بیان شد. اکنون به ارائه روش پیشنهادی خود و مراحل انجام آن می پردازیم و در این راستا کارکرد هر یک از روش های پیشنهادی با پیاده سازی آن روش ارزیابی می گردد. کد پیاده‌ سازی و جزئیات اجرایی مربوط به آن از طریق این لینک در گیت‌هاب قابل مشاهده است.

**مجموعه داده‌های مورد استفاده شده در آزمایش**

**Data Files**

|  |  |  |
| --- | --- | --- |
| نام فایل | فرمت فایل | توضیح مجموعه داده |
| offers.csv | [.gz (431 b)](https://www.kaggle.com/c/acquire-valued-shoppers-challenge/download/offers.csv.gz) | شامل اطلاعات پیشنهادات تشویقی |
| sampleSubmission.csv | [.gz (444.55 kb)](https://www.kaggle.com/c/acquire-valued-shoppers-challenge/download/sampleSubmission.csv.gz) | نمونه ارسالی |
| testHistory.csv | [.gz (908.63 kb)](https://www.kaggle.com/c/acquire-valued-shoppers-challenge/download/testHistory.csv.gz) | مجموعه داده آزمایشی شامل سابقه پیشنهادات داده شده به هرمشتری بدون احتساب طرز برخورد مشتریان به پیشنهاد (جهت پیش بینی ستونهای تکرار شده برای هر کدشناسایی در این فایل ) |
| trainHistory.csv | [.gz (990.87 kb)](https://www.kaggle.com/c/acquire-valued-shoppers-challenge/download/trainHistory.csv.gz) | مجموعه داده آموزشی شامل سابقه پیشنهادات داده شده به هرمشتری و اطلاعات طرز برخورد مشتری به پیشنهاد |
| transactions.csv | [.gz (2.86 gb)](https://www.kaggle.com/c/acquire-valued-shoppers-challenge/download/transactions.csv.gz) | شامل سابقه تراکنش‌های کلیه مشتریان در طی یکسال و پیش از ارایه پیشنهاد تشویقی |

مجموعه‌داده‌ی تراکنش‌ها به صورت فشرده شده درآمده است که پس از اینکه از حالت فشرده خارج شود به فضایی در حدود 22گیگا بایت نیاز دارد .

این مجموعه داده برگرفته از فرآیند پیشنهاد طرح‌های تشویقی ( مانند کوپن‌های تشویقی ) به تعداد زیادی از مشتریان و پیش‌بینی آن‌هایی که نسبت به محصول وفادار خواهند بود است . فرض شده است به 100 نفر از مشتریان تخفیفی برای خرید دو بطری آب پیشنهاد گردیده است . از 100 مشتری انتخابی 60 نفر تصمیم به استفاده از طرح تشویقی گرفته اند . و این 60 نفر به عنوان هدف این مقاله درنظر گرفته شده اند که پیش‌بینی شود آیا برای خرید مجدد همان محصول(در زمان اجرای طرح تشویقی یا پس از اتمام آن ) اقدام خواهند کرد یا خیر .

برای انجام این پیش‌بینی مجموعه داده‌ایی از حداقل خرید های تشویقی یکسال اخیر هر مشتری ، به علاوه سابقه خرید‌های مشتریان دیگر ( برخی از آنهایی که پیشنهاد مشابهی را دریافت کرده‌اند) داده شده است . سابقه تراکنش‌ها شامل کلیه اقلام خریداری شده بوده و فقط اقلام مرتبط به پیشنهاد داده شده نمی‌باشد . فقط به ازای هر مشتری یک پیشنهاد داخل این مجموعه داده می‌باشد .

مجموعه داده‌آموزشی شامل پیشنهادات ثبت شده پیش از 01-05-2013 می‌باشد . و مجموعه داده آزمایشی شامل پیشنهادات ثبت شده بعد از آن تاریخ می‌باشد .

**فیلدهای مجموعه داده ها**

کلیه فیلدها برای محافظت از اطلاعات مشتریان و فروش‌ها صورت گرفته به صورت ناشناس و طبقه بندی شده گردیده‌اند. فیلدها دربردارنده معنای خاصی نیستند .

|  |  |  |
| --- | --- | --- |
| تشریح فیلد | نام فیلد | نام مجموعه داده |
| کدشناسایی منحصری به نمایندگی از یک مشتری | **id** | history |
| مقداری صحیح به نمایندگی از یک فروشگاه زنجیره‌ای | **chain** | history |
| یک کد شناسایی به نمایندگی از یک پیشنهاد خاص | **offer** | history |
| یک کد شناسایی به نمایندگی از منطقه جغرافیایی | **market** | history |
| تعداد دفعاتی که یک مشتری خریدهای تکراری داشته است | **repeattrips** | history |
| مقداری بولین برابر با یک اگر مشتری خرید تکراری داشته باشد | **repeater** | history |
| تاریخ دریافت پیشنهاد توسط یک مشتری | **offerdate** | history |
| کدشناسایی تراکنش‌های صورت یافته | **id** | transactions |
| مقداری صحیح به نمایندگی از یک فروشگاه زنجیره‌ای | **chain** | transactions |
| گروهی تجمیع شده از یک دسته (مثل آب معدنی) | **dept** | transactions |
| دسته محصول | **category** | transactions |
| کدشناسایی مربوط به شرکتی که اقلام را می‌فروشد | **company** | transactions |
| کدشناسایی نام تجاری ایی که اقلام به آنها تعلق دارند | **brand** | transactions |
| تاریخ خرید | **date** | transactions |
| مقدار خرد محصول ( مثلا 12 عدد آب معدنی ) | **productsize** | transactions |
| واحد محصول خریداری شده | **productmeasure** | transactions |
| تعداد واحدهای خریداری شده محصول | **purchasequantity** | transactions |
| مبلغ کل خرید شده به دلار | **purchaseamount** | transactions |
| یک کد شناسایی به نمایندگی از یک پیشنهاد خاص | **offer** | **offers** |
| دسته محصول | **category** | **offers** |
| تعداد واحدهایی که یک شخص برای دریافت تخفیف باید خرید کند | **quantity** | **offers** |
| کدشناسایی مربوط به شرکتی که اقلام را می‌فروشد | **company** | **offers** |
| مبلغ ارزش پیشنهاد به دلار | **offervalue** | **offers** |
| کدشناسایی نام تجاری ایی که اقلام به آنها تعلق دارند | **brand** | **offers** |

فایل تراکنش ها (transactions) می‌تواند به وسیله (id,chain) به فایل تاریخچه (history) بپیوندد . فایل تاریخچه می‌تواند به فایل پیشنهادات (offers) توسط (offer) متصل شود . فایل تراکنش ها می تواند به وسیله (category, brand, company) به فایل پیشنهادات متصل گردد. مقداری منفی در تعداد محصول [[52]](#footnote-52) و مقدار خرید [[53]](#footnote-53) نشان دهنده بازگشت است .

**مدل**

همانطور که در قسمت ابتدایی توضیح داده شد، فرکانس خرید ها در یک بازه زمانی معین و نیز زمانی که خریدار آخرین تراکنش خود را انجام می دهد، دو داده اساسی برای حل این مسئله است. در این جا ما علاوه بر این دو، معیار هزینه[[54]](#footnote-54) را نیز در نظر می گیریم. معیار هزینه در واقع مشخص کننده ارزش مشتری است و منظور از آن میزان پولی است که مشتری در یک بازه زمانی مشخص خرج می کند. چنین مدلی که سه معیار تاخر خرید، فرکانس خرید و هزینه مربوط به خرید در آن در نظر گرفته می‌شود را در اصطلاح مدل RFM [[55]](#footnote-55) می گویند و ما برای پیاده سازی از چنین مدلی استفاده خواهیم کرد. سپس در کنار استفاده از این سه معیار، با بکارگیری برخی مشخصه ها و معیار هایی خاص به صورت هیوریستیک، کارایی این مدل بهبود داده می شود و چنین مدلی را به عنوان یک راه حل قابل قبول از نظر کارایی مطرح خواهیم کرد.

**شاخص های مدل**

شاخص های مدل RFM و نحوه تاثیر هر شاخص در مدل به طور خلاصه به صورت زیر می باشد:

ارزش پولی تراکنش: این ﺷﺎﺧﺺ ﻧﺸﺎن دﻫﻨﺪه مقدار ﭘﻮلی است ﻛﻪ ﻳﻚ ﻣﺸﺘﺮی در ﻳﻚ دوره زﻣﺎنی ﺧﺎص ﺟﻬﺖ ﻣﺒﺎدﻻت، ﺻﺮف نموده اﺳﺖ . بیشتر ﺑﻮدن ﻣﻘﺪار ﭘﻮل خرج ﺷﺪه، بیانگر ﺑﺎﻻ ﺑﻮدن ارزش اﻳﻦ ﺷﺎﺧﺺ در ﻣﺪل می ﺑﺎﺷﺪ.

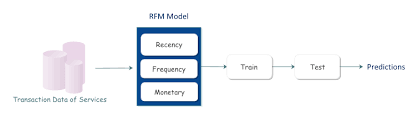
به طور کلی، برای انجام پیش بینی در مورد وفاداری مشتریان دانستن دو داده اصلی [[56]](#footnote-56) در مورد آنها پر اهمیت است:

1. تازگی خرید[[57]](#footnote-57): آخرین تراکنشی که خریدار انجام داده است چه زمانی بوده است. ﻛﻤﺘﺮ ﺑﻮدن اﻳﻦ ﻓﺎﺻﻠﻪ ﻧﺸﺎﻧﮕﺮ ﺑﺎﻻ ﺑﻮدن ارزش اﻳﻦ ﺷﺎﺧﺺ در ﻣﺪل میﺑﺎﺷﺪ.
2. دفعات یا فرکانس خرید[[58]](#footnote-58): در یک بازه زمانی مشخص، چه تعداد تراکنشی را خریدار انجام داده است. بیشتر ﺑﻮدن ﺗﻌﺪاد ﻣﺒﺎدﻻت، ﻧﺸﺎﻧﮕﺮ ﺑﺎﻻ ﺑﻮدن ارزش اﻳﻦ ﺷﺎﺧﺺ در ﻣﺪل میﺑﺎﺷﺪ.
3. ارزش پولی تراکنش: این ﺷﺎﺧﺺ ﻧﺸﺎن دﻫﻨﺪه مقدار ﭘﻮلی است ﻛﻪ ﻳﻚ ﻣﺸﺘﺮی در ﻳﻚ دوره زﻣﺎنی ﺧﺎص ﺟﻬﺖ ﻣﺒﺎدﻻت، ﺻﺮف نموده اﺳﺖ . بیشتر ﺑﻮدن ﻣﻘﺪار ﭘﻮل خرج ﺷﺪه، بیانگر ﺑﺎﻻ ﺑﻮدن ارزش اﻳﻦ ﺷﺎﺧﺺ در ﻣﺪل می ﺑﺎﺷﺪ.

تجربه نشان داده است بررسی رفتارهای مصرف کننده در گذشته بهترین راه پیش بینی رفتار آینده وی است. خریداری که برای چندمین بار از یک فروشگاه خرید می کند به احتمال زیادی مجددا رفتار خود را در مقایسه با فردی که تنها یک بار خرید داشته تکرار خواهد نمود. در واقع فرکانس یا تعداد دفعات خرید می تواند عاملی مهم در تعیین احتمال مربوط به خرید مجدد باشد.اما اگر بخواهیم اقدامات آینده مشتری و ارزش وی را با دقت بیشتری پیش بینی نماییم، لازم است زمان انجام آخرین تراکنش ( خرید ) را نیز در نظر بگیریم. در نظر گرفتن عامل تازگی همیشه به ما این قابلیت را خواهد داد تا بفهمیم اکنون چه می گذرد چرا که آن چه که از همه چیز مهم تر است وضعیت فعلی است. در واقع می توان این گونه گفت که با ارزش ترین مشتری ها کسانی هستند که اخیرا خرید کرده اند و نیز رفتار تکرار خرید را دارا بوده اند. خریدارانی که پیشتر رفتار خرید تکراری خود را داشته اند اما اخیرا خرید نکرده اند، از نظر ارزش در رده بعدی قرار می گیرند.

**شرح مدل**

پیشتر گفتیم که در مدل RFM، حرف R نشان دهنده تازگی خرید ، حرف F معرف فرکانس خرید و حرف M نمایانگر ارزش پولی خرید است. باید توجه داشت که هر چه F و R بیشتر باشند احتمال آنکه تراکنش جدیدی با مشتری صورت بگیرد بیشتر است و همچنین اگرM نیز بزرگتر باشد احتمال بازگشت مشتری برای خرید بیشتر است. در مدل RFM فرض بر این است که مشتریانی که در هر یک از متغیرهای مدل دارای ارزش بالاتری هستند جز بهترین مشتریان هستند، البته تا زمانی که فرض شود آنها در آینده نیز همانند گذشته رفتار نمایند که در این صورت اعتقاد بر این است که این مشتریان نسبت به دیگران برای شرکت سودآوری بالاتری دارند. لذا فرض اساسی مدل این است که الگوهای آینده مبادله و خرید مشتری، همانند الگوهای گذشته و حال هستند. راحتی محاسبه، قابل درک بودن و توانایی RFM در پیش بینی رفتار آینده مشتری برخی از مزایای این مدل است به طوری که آن را به عنوان راهی برای اندازه گیری ارزش طول عمر مشتری مطرح کرده است.



روند کلی کار تا رسیدن به خروجی با استفاده از یک مدل RFM

**بهبود مدل با بکارگیری LRFM [[59]](#footnote-59)**

مدل LRFM روشی است که برای خوشه‌بندی مشتریان در مدیریت ارتباط با مشتری (CRM) استفاده می‌شود. در این مدل مشتریان براساس چهار ویژگی طول ارتباط مشتری ، تازگی خرید ، تناوب خرید و ارزش مالی خرید دسته‌بندی می‌شوند. براساس حرف اول هریک از این چهار ویژگی واژه LRFM نوآوری شده است. این مدل برای خوشه‌بندی بسیار مورد استفاده قرار می‌گیرد.

بر طبق نظررینارتزو کومار (2000)، چانگ و تسای (2004) مدل RFM نمی تواند مشتریان دارای ارتباط بلند مدت و مشتریان دارای ارتباط کوتاه مدت با سازمان را مشخص نماید. آنها در تحقیق خود ایده طول ارتباط مشتری را پیشنهاد می دهند و به بررسی تاثیر آن بر وفاداری و سود آوری مشتری می پردازند. آنها بیان می کنند که افزایش طول ارتباط با مشتری، وفاداری مشتری را بهبود خواهد بخشید. واین متغیر را که نشان دهنده فاصله زمانی بین اولین و آخرین خرید مشتری در بازه مورد مشاهده است تعریف کرده اند. مدل RFM مشتریانی را که به تازگی ارزش مالی بالایی برای شرکت ایجاد کرده و در کوتاه مدت دارای تناوب خرید بیش از متوسط تناوب خرید در بین مشتریانی که تکرارخرید داشته اند به عنوان مشتریان با ارزش انتخاب کرده درحالی که عامل طول ارتباط با شرکت نادیده گرفته شده است.

بنابراین بعد طول ارتباط مشتری Length(L) به مدل RFM اضافه می شود.

طول ارتباط مشتری با سازمان، نشان دهنده مدت زمانی است که یک مشتری ارتباط خود را با سازمان آغاز کرده است. نتایج مطالعه بهاتاچاریا (1998) نشان می دهد طول دوره هایی که طی آن می توان یک فرد را به عنوان مشتری سازمان معرفی کرد، در ارتباط مثبت با دوره هایی خواهد بود که مشتری متمایل است طی آن ارتباط خود را با سازمان ادامه دهد . هم چنین در این مطالعه بیان شده که طول ارتباط مشتری با سازمان با احتمال پایداری رابطه او در آینده، دارای ارتباط مثبت است. از طرفی مبنای دیگر برای دنبال کردن مشتریان در نظریه مزیت منابع پیدا شده است. برطبق بررسی های هانت و مورگان ( 1995 ) منابع شرکت، مالی، فیزیکی، قانونی، انسانی، سازمانی ، اطلاعاتی ، و رابطه ای هستند . ازبررسی های آنان ، نشان داده شده که روابط با مشتریان می تواند بعنوان یک منبع شرکت بهره گرفته شده که می تواند بعنوان مزیت رقابتی بکاربرده شود.

در نظر گرفتن برخی معیار های خاص دیگر در کنار استفاده از مدل RFM می تواند عملکرد این مدل را بهبود دهد. این معیار ها در واقع نقش هیوریستیکی دارند و در صورتی که خوب در نظر گرفته شوند، می توانند کارایی و دقت پیش بینی را بالا برند. پس می توان گفت بخش مهم این مسئله مهندسی معیار ها یا همان مهندسی خصوصیات [[60]](#footnote-60)است که هر چه دقیق تر باشند خروجی از دقت بهتری برخوردار خواهد بود.

**اعمال معیار ها**

با استفاده از معیار ها و مشخصه های در نظر گرفته شده، با به کار گیری یک اسکریپ مجموعه های تست و یادگیری به گونه ای بدست می آیند که این معیار ها نیز در آنها اعمال شده است، سپس برای یادگیری و تست مدل با استفاده از مجموعه های بدست آمده از کتابخانه VW یا همان Vowpal Wabbit استفاده می کنیم. VW در واقع یک کتابخانه متن باز یادگیری است که با سرعتی قابل قبول می تواند مدل را یادبگیرد [[61]](#footnote-61) و یا آنکه آن را تست نماید. یکی از مزایای مهم آن این است که دارای خطای محاسباتی کمی است به طوری که در این مسئله میزان داده های از دست رفته در مرحله یادگیری تنها در حدود 0.156 است. همچنین قابلیت دیگر مهم این کتابخانه آن است که می تواند میزان ارتباط معیار ها یا مشخصه های در نظر گرفته شده را با یک مجموعه داده ورودی بسنجد. در واقع دستور vw-varinfo از کتابخانه VW، اطلاعاتی را در مورد میزان ارتباط معیار ها با مجموعه دادگان ورودی در اختیار می گذارد.

**تشریح معیار ها**

فرض کنیم اینکه یک مشتری از یک کمپانی تا به حال خرید کرده است یاخیر، را یک معیار در نظر بگیریم. در این صورت این معیار به وضوح یک مقدار دودویی خواهد داشت. اما می توان این معیار را از حالت دودویی خارج کرد به گونه ای که حاوی اطلاعات بیشتری باشد. برای مثال آنکه اگر مشتری از کمپانی خرید کرده است، در کل چند قلم کالا خریده است و یا چه میزان پول خرج کرده است، معیاری کامل تر نسبت به حالت قبل است که اطلاعات بیشتری را در خود نهفته است. از طرفی پارامترهای زمانی نیز می تواند به این معیار اضافه شود. برای مثال آنکه در مدت 3 ماه، 6 ماه یا 9 ماه، مشتری چند قلم کالا از یک کمپانی خریده است و یا آنکه در همین مدت چه میزان هزینه کرده است، به‌وضوح یک معیار کامل تر نسبت به هر دو حالت قبل می باشد. حال میزان هزینه خرج شده را با حرف a و تعداد اقلام خرید شده را با حرف q نشان دهیم و پارامتر زمان را نیز بر حسب روز در نظر می گیریم.

**بررسی کارایی معیارها**

در این قسمت کارایی برخی از معیارهای در نظر گرفته شده با یکدیگر مقایسه می شوند. اکنون برای مثال یک معیار مقدار هزینه ای است که یک مشتری در مدت یک ماه برای خرید یک برند خاص صرف کرده است و یا معیار دیگر مقدار هزینه ای است که همان مشتری این بار در مدت سه ماه برای خرید همان برند صرف کرده است. طبق آنچه گفتیم معیار اول را با has-bought-brand-a-30 نشان می دهیم و معیار دوم با has-bought-brand-a-90 مشخص می شود. در مقایسه کارایی این دو معیار، به وضوح معیار اول کارایی بهتری نسبت به معیار دوم خواهد داشت. یکی از دلایل این امر آن است که معیار اول عامل تاخر یا تازگی تراکنش ها را به نسبت معیار دوم بهتر در نظر می گیرد. به عبارت دیگر، اگر چه تمام تراکنش هایی که در معیار اول در نظر گرفته می شوند، در معیار دوم نیز در نظر گرفته می شوند اما میانگین زمانی تمام تراکنش های انجام شده در معیار اول در زمان نزدیک تری نسبت به میانگین زمانی تمام تراکنش های معیار دوم می باشد و بنابراین تراکنش های معیار اول تازگی بیشتری دارند. در واقع اگر مشتری در یک ماه گذشته هیچ تراکنشی انجام نداده باشد، در معیار اول مشخص می شود ولی در معیار دوم مشخص نمی گردد چراکه ممکن است مشتری در دو ماه قبل تراکنشی انجام داده باشد و لذا مقدار این معیار نمی تواند عدم وجود خرید یا تراکنش را در یک ماه قبل نشان دهد. اکنون به وضوح کارایی سایر معیار های موجود نیز با استدلال هایی مشابه قابل بررسی و توجیه است.

**مراحل**

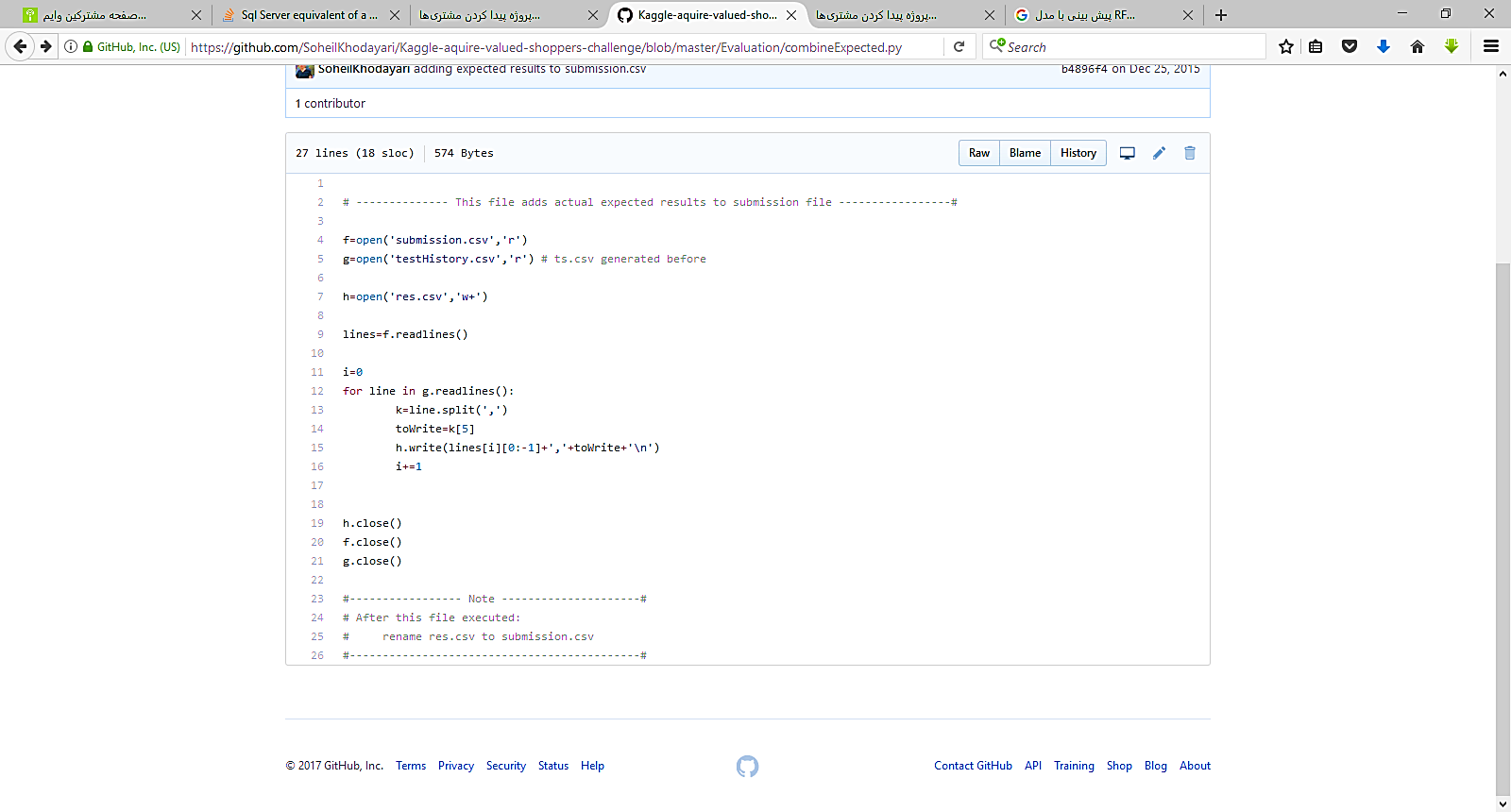
در این قسمت روند کلی پیاده سازی مسئله به اختصار توضیح داده می شود.

* **بارگذاری داده ها:** در ابتدا مجموعه داده های موجود در برنامه بارگذاری می شود.
* **استخراج معیار ها:** در حالت ابتدایی، استخراج معیار ها تنها مبتنی بر مدل RFM صورت می گیرد که این حالت را آزمایش اول در نظر می گیریم. اما همانطور که پیشتر گفته شد، با اعمال برخی از معیار های دیگر می توان کارایی این مدل را افزایش داد. چنین حالتی به عنوان آزمایش دوم بررسی می شود.
* **بررسی معیار ها:** معیار های در نظر گرفته شده به کمک vw-varinfo قابل بررسی است.
* **یادگیری مدل:** در این قسمت مدل با استفاده از VW یادگیری می شود.
* **تست مدل:** اکنون به کمک مدل یادگیری شده و با استفاده از VW، مدل تست شده و خروجی تولید می گردد.
* **تغییر فرمت خروجی:** با بکارگیری یک اسکریپت خروجی بدست آمده به فرمت نمونه قرار داده شده در مجموعه دادگان تبدیل می شود. این فرمت اینگونه است که در هر سطر دو ستون قرار می گیرد به طوری که در ستون اول شناسه خریدار و در ستون دوم احتمال انجام تراکنش ( خرید ) توسط خریدار قرار می گیرد.

**ارزیابی**

در این قسمت، ضمن معرفی بهترین روش ارزیابی مسئله با توجه به چالش های موجود، نتایج حاصل از هر یک از آزمایش های صورت گرفته با روش‌هایی که در ادامه برای ارزیابی مطرح می کنیم نشان داده می شوند. لازم به ذکر است که برای ارزیابی، داده های مربوط به قسمت یادگیری به دو قسمت تقسیم می شود و از نیمی از آن برای تست و از نیم دیگر آن برای یادگیری استفاده می کنیم. علت اینکار آن است که داده های مربوط به تست در مسابقه دارای خروجی صحیح نیستند. بنابراین برای ارزیابی نتایج، مراحل قبل با دادگان جدید تکرار خواهند شد. کدهای مربوط به ارزیابی نتایج در زیر آمده است .

فایل combineExpected.py جهت ادغام فایل های submission.csv و testHistory.cv و ساخت فایل موقتی به نام res.csv که در نهایت باید به submission.csv تغیر نام بدهیم .



**چالش های ارزیابی**

یکی از چالش های ارزیابی نتایج حاصل شده از مراحل قبل آن است که این نتایج به صورت احتمال هستند و مقادیری پیوسته دارند در حالی که خروجی در واقعیت حالتی گسسته دارد. به عبارت دیگر، فرض کنید خروجی بدست آمده برای یک مشتری با شماره شناسه مفروض 25 برابر 0.43 باشد. این بدان معناست که مشتری شماره 25 به احتمال 0.43 مجددا برای خرید باز خواهد گشت. اما بدیهی است که آنچه که در واقعیت رخ می دهد دو حالت بیشتر ندارد و یا آن است که مشتری مجددا برای خرید باز می گردد و یا آنکه دیگر برای خرید بر نخواهد گشت یعنی در واقعیت خروجی یک مقدار دودویی دارد. حال سوال اینجاست که احتمال 0.43 برابر کدام حالت در نظر گرفته شود( یک یا صفر). یک روش ساده برای رفع این مشکل آن است که یک احتمال خاص را به عنوان یک برش [[62]](#footnote-62) و یا در اصطلاح آستانه تمایز [[63]](#footnote-63)در نظر بگیریم و تمام احتمالات بیشتر از آن یک و احتمالات کمتر از آن صفر فرض شوند. روش دیگر آن است که نتایج به ازای چندین برش خاص بدست آیند و از آنها میانگین گرفته شود. بدیهی است که دو روش ارزیابی اخیر دقیق نیستند و نمی توانند به خوبی میزان دقت نتایج بدست آمده را برآورد کنند. راهکاری که به وسیله آن می توان میزان دقت چنین داده هایی را بررسی نمود استفاده از منحنی ROC [[64]](#footnote-64) است که در ادامه تشریح می گردد.

**شاخص های ارزیابی**

شاخص های ارزیابی خروجی ها به همراه روابط محاسبه آنها به صورت زیر می باشد.

در روابط فوق F1 در واقع میانگین هارمونیک دقت و فراخوانی است که می تواند این دو شاخص را به طور هم زمان نشان دهد. در واقع می دانیم هر روش خوب باید هم دقت خوبی داشته باشد و هم فراخوانی آن بالا باشد و لذا می توان گفت در مجموع F1 هر دو شاخص اخیر را به طور متعادل در نظر می گیرد.

در روابط فوق منظور از tp تعداد حالاتی است که به درستی صحیح و منظور از fp تعداد حالاتی است که به غلط، صحیح تشخیص داده شده‌اند. همچنین منظور از fn تعداد حالاتی است که به غلط نادرست و tn تعداد حالاتی است که به درستی غلط تشخیص داده شده اند. جدول زیر به وضوح این چهار مفهوم را در خود خلاصه کرده است.

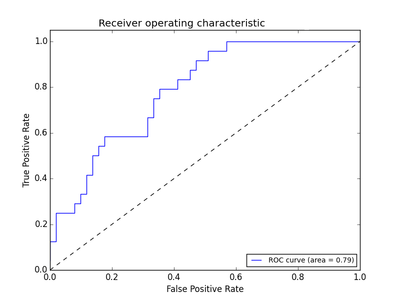
|  |  |
| --- | --- |
| مفهوم | نماد |
| تعداد حالاتی است که به درستی صحیح تشخیص داده شده شده‌اند. | tp |
| تعداد حالاتی است که به درستی غلط تشخیص داده شده اند. | tn |
| تعداد حالاتی است که به غلط، صحیح تشخیص داده شده‌اند. | fp |
| تعداد حالاتی است که به غلط نادرست تشخیص داده شده‌اند. | fn |

بدیهی است که هر روش خوب، باید هر یک از پنچ شاخص یاد شده را به طور همزمان افزایش دهد.

**روش های ارزیابی**

دو روش مناسب برای برآورد میزان دقت خروجی ها به اختصار به صورت زیر می باشد:

1. **میانگین گیری برش ها:** در این روش نتایج به ازای همه برش های بین 1 تا 99 حساب می شود و هر یک از معیار های مورد نیاز برای ارزیابی در هر حالت بدست آمده و سپس از آنها میانگین گرفته می شود. اگر چه این روش به نسبت آنکه تنها یک یا چندین برش خاص در نظر گرفته شود بهتر است اما باز هم میزان دقت روش بعد را ندارد.
2. **سطح زیر منحنی:** در این روش نموداری با استفاده از دو نرخ tp و fp و با تغییر آستانه تمایز( حد برش ) بدست خواهد آمد به طوری که نرخ tp در محور عمودی و نرخ fp در محور افقی قرار می گیرد. منحنی بدست آمده از این نمودار را ROC می نامند. اکنون سطح زیر این منحنی میزان دقت خروجی را مشخص خواهد کرد. از این رو این روش ارزیابی در اصطلاح AUC[[65]](#footnote-65) نامیده می شود. این روش دقیق ترین راهکار ارزیابی خروجی ها می باشد.



منحنی ROC و سطح زیر آن ( AUC )

**نتایج ارزیابی**

در این قسمت نتایج حاصل از بکارگیری هر یک از مدل هایی که در قسمت های قبلی بیان شد، در قالب جداولی نشان داده می شود. در هر قسمت هر دو روش ارزیابی ذکر شده و نیز تمام شاخص ها برای هر روش در نظر گرفته خواهد شد اما در نهایت، نتایج حاصل از AUC ملاک عمل خواهد بود چرا که می تواند دقیق تر از روش میانگین‌گیری میزات دقت خروجی ها را ارزیابی کند.

**آزمایش اول**

آزمایش اول مربوط به حالتی است که از مدل RFM ( غیر بهبود یافته ) به تنهایی برای تولید خروجی استفاده شده است. نتایج حاصل از این حالت پس از طی مراحلی که پیشتر بیان شد، در جدول زیر خلاصه شده است.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Accuracy | True Negative Rate | F-Score | Recall | Precision | Method |
| 86.6 | 85.826 | 86.519 | 87.398 | 85.657 | AUC |
| 78.9 | 78.388 | 78.708 | 79.429 | 78 | Average Of Cuts |

همانگونه که از جدول فوق مشخص است، مقدار F1 با بکارگیری این نوع مدل طبق AUC برابر 86.51 % است.

**بررسی و مقایسه نتایج**

با مقایسه نتایج بدست آمده از جداول فوق ، به وضوح مشخص است که شاخص F1 در روش دوم رشدی 19 درصدی نسبت به روش اول داشته است و این موضوع از دقت بیشتر و کارآمدی این روش نسبت به روش اول حکایت دارد. همچنین با دقت بر روی جداول می توان دریافت که در آزمایش دوم، مقادیر فراخوانی و دقت، هر دو نسبت به آزمایش اول بهبود داده شده اند. نکته ی جالب آنکه اگر حتی از نتایج ارزیابی دیگر ( یعنی میانگیری نتایج در برش ها ) استفاده کنیم، در این حالت برای آزمایش دوم نتایج بدتر (به نسبت نتایج AUC) را انتخاب کرده ایم و برای آزمایش اول نتایج بهتر (به نسبت نتایج AUC) انتخاب شده است و در این حال باز هم نتایج مربوط به آزمایش دوم بهتر از نتایج آزمایش اول می باشد. بنابراین روش مطرح شده در آزمایش دوم یعنی استفاده از یک مدل RFM توسعه یافته ، به عنوان یک راه حل کارا و با دقت مناسب برای حل مسئله مشتری وفادار مطرح می شود.

**بررسی بهبود مدل**

پیشتر مدل RFM توسعه یافته به عنوان مدل پیشنهادی برای حل این مسئله معرفی شد. گفتیم با در نظر گرفتن برخی معیار های خاص در کنار استفاده از مدل RFM ابتدایی می توان عملکرد این مدل را بهبود داد و چنین معیارهایی در واقع نقش هیوریستیکی داشتند. سپس برای حل این مسئله با این حجم از دادگان معیار هایی در نظر گرفته شد و به بررسی کارایی این معیارها پرداخته شد و نتایج حاصل از ارزیابی روش بهبود یافته به همراه نتایج روش قبلی ارائه شد. اکنون به بررسی کارایی مدل بهبود یافته می پردازیم.

**کارایی مدل**

نکته بسیار مهمی که در این میان نهفته است و بر آن تاکید می شود آن است که چالش برانگیز ترین قسمت حل این مسئله کشف و انتخاب معیارهاست که در اصطلاح آن را مهندسی معیارها می نامند. این که چه معیارهایی در نظر گرفته شود و دقت و ارتباط آنها با داده ها کدام است و ابزار سنجش این دقت یا ارتباط چیست در واقع همان مهندسی معیارهاست. در این جا از کتابخانه VW استفاده شد و نشان دادیم که چگونه می توان با استفاده از این کتابخانه ( و به طور مشخص با vw-varinfo ) میزان ارتباط معیارها را با مجموعه داده ورودی سنجید. در این جا ذکر چند نکته خالی از لطف نیست:

* در ابتدا متذکر می شویم که داده های مربوط به تراکنش ها که در صفحه اصلی مسابقه حدود 22 گیگابایت در حالت غیر فشرده بود، به حدود 1.6 گیگابایت کاهش داده شد.
* معیارهای انتخابی اکیدا به داده های ورودی بستگی دارند به گونه ای که ارتباط آنها با VW تخمین زده شد ( کما این که خود کتابخانه VW نیز دارای خطاست).
* از آنجا که معیارهای انتخابی بسته به دادگان ورودی متفاوت بوده و باید متناسب با آن انتخاب شوند، معیارهای انتخابی در این جا نیز تنها برای دادگان کاهش یافته این مسئله یعنی حجمی در حدود 1.6 گیگابایت و آن هم ضمن رعایت استراتژی کاهش داده انتخابی، معتبر است.
* بدیهی است که برای حجم بالای 22 گیگابایتی تراکنش ها، معیارهای انتخابی یا ممکن است کمی متفاوت باشند و یا آنکه میزان کارایی آنها متفاوت باشد. این کارایی با vw-varinfo در این جا بررسی شد.
* تخمین معیارهای مناسب برای دادگان با حجم بالاتر با توجه به دشواری آنالیز دادگان و نیز خطای قابل توجه vw-varinfo، پیچیده تر خواهد بود.
* هر چه حجم دادگان ورودی بالاتر می رود vw-varinfo خطای بیشتری از خود نشان می دهد که این خطا در انتهای عملیات انتخابی توسط خود این کتابخانه گزارش می شود. بنابراین با خطای بیشتری معیار ها انتخاب می شوند و این خطا حتما در کارایی تاثیر خواهد داشت.

بدیهی است که موارد فوق تلویحا به این نکته اشاره دارد که ارزیابی صورت گرفته فقط و فقط برای این دادگان معتبر است و این اصولا بدان دلیل است که مهندسی معیار ها در روش RFM بهبود یافته پیشنهادی مبتنی برداده است. برای مثال هنگامی که حجم دادگان تراکنش ها به ترتیب به حدود 3 و 4 گیگابایت افزایش یافتند، دقت با استفاده از AUC به ترتیب به حدود 84.878 و 83.643 کاهش یافت و دلیل آن هم آن است که مهندسی معیار ها متناسب با داده مجددا برای دادگان جدید صورت نگرفته است و از همان معیار های روش قبل به صورت کلی استفاده شده است و نیز آنکه خطای حاصل از تست و یادگیری افزایش داشت.

یک نکته بسیار مهم دیگر نیز در اینجا وجود دارد که مغفول مانده است و آن این است که VW در مراحل تست و یادگیری همواره دارای خطاست به گونه ای که برای حجم بالای دادگان، این خطا بسیار چشمگیر، غیر قابل انکار و تاثیرگذار است. بدیهی است که میزان این خطا در انتهای مراحل تست و یادگیری نشان داده می شود وبه وضوح قابل مشاهده و بررسی است.

**نتیجه گیری**

در مقایسه استفاده از روش RFM توسعه یافته پیشنهادی با روش های مطرح شده که پیشتر بیان شد ، باید خاطر نشان کرد که آن روش ها اگر چه ممکن است دقت خوبی داشته باشند اما پیاده سازی آنها به ویژه به هنگام تخمین پارامتر ها بسیار دشوار و چالش برانگیز است تا آنجا که تا به امروز کمتر کسانی روی به پیاده سازی کامل آنها آورده اند. همچنین در مقایسه این روش با روش مطرح شده ، یعنی استفاده از مدل مرتبه ای بیزی باید گفت که در صورتی که تعداد مشتریان بسیار زیاد شود، مدل مرتبه ای بیزی در عمل کارایی خود را از دست خواهد داد چرا که برای هر مشتری، زنجیره مارکوف - مونت کارلو باید صد ها بار تکرار شود و این خود نیاز به زمان زیادی دارد. همچنین در صورتی که در این روش از تکنیک تقریب مبتنی بر رگرسیون برای کاهش زمان اجرا استفاده گردد، اگر چه ممکن است زمان اجرا کاهش یابد، اما چون از تقریب استفاده می شود در عمل میزان دقت و کارایی کاهش خواهد داشت. همه آنچه گفته شد نشان از برتری روش ارائه شده نسبت به روش های دیگر دارد.

**کارهای آینده**

در بازار رقابتی امروز نباید از ریزش مشتریان غافل شد اما واقعیت آن است که ریزش امری اجتناب ناپذیر است که نمی توان آنرا به صفر رساند ولی می توان آنرا مدیریت کرد و کاهش داد. برای نیل به چنین هدفی، گام اول شناسایی مشتریانی است که با احتمال بیشتری امکان ریزش دارند و با آشکار نمودن علت ریزش آنها می توان سعی در حفظ آنها داشت. بنابراین پیش بینی مشتریان وفادار برای یک سازمان نه تنها از جهت حفظ مشتریان وفادار حائز اهمیت است بلکه از جهت کاهش نرخ ریزش نیز اهمیت دارد. روشی که در این مقاله برای پیش بینی مشتریان وفادار مطرح شد، استفاده از یک مدل RFM توسعه یافته به کمک برخی معیارهایی بود که نشان داده شد. آنچه لازم است در ادامه بررسی شود آن است که آیا می توان معیار هایی بهتر و کارآمد تر برای این مدل پیدا کرد یا خیر؟ آیا می توان معیار هایی پیدا کرد که مستقل از اندازه داده های تراکنش ها همیشه به یک اندازه خوب عمل کنند؟ آیا استفاده از این مدل در کنار روش های دیگری از جمله تقریب مبتنی بر رگرسیون و سپس میانگین گیری از احتمالات دو روش بدست آمده و استفاده از مقدار میانگین به عنوان خروجی باعث بهتر شدن نتایج می گردد یا خیر؟ اگر در چنین روشی به جای میانگین گیری به صورت تصادفی یکی از احتمالات هر بار انتخاب شوند در دقت روش چه تاثیری خواهد گذاشت؟ به طور کلی ادغام مدل RFM مطرح شده با سایر روش ها، آیا می تواند میزان دقت خروجی را افزایش دهد و یا خیر؟ این ها همگی سوالات و ایده هایی هستند که در آینده نیاز به بررسی دارند چرا که هر یک ممکن است منجر به روشی کارآمد تر و با دقت بیشتر نسبت به روش کنونی مطرح شده گردند.

همچنین مورد مهم دیگری که می توان به آن اشاره کرد مهندسی معیارها در یک مدل RFM توسعه یافته است. در واقع پیدا کردن معیارهایی که مستقل از اندازه داده های تراکنش ها همیشه به یک اندازه خوب عمل کنند و این که اساسا چنین معیارهایی وجود دارد یا خیر، گامی اصلی و تاثیرگذار در حل مسئله مشتریان وفادار است.

شاید مهمترین نکته ای که در این پژوهش به صورت سطحی به آن پرداخته شد و به بررسی بیشتری در آینده نیاز دارد، روش تقریب مبتنی بر رگرسیون و یا به طور کلی هر روشی مبتنی بر آن باشد که انتظار می رود در کنار استفاده از یک مدل RFM، کارایی قابل قبولی از خود نشان دهد. روش تقریب مبتنی بر رگرسیون از این حیث حائز اهمیت است که زمان اجرا را کاهش می دهد و این کاهش زمان اجرا یا همان افزایش سرعت، خود را در داده های با حجم بالا نشان خواهد داد، به خصوص از آن جهت که مهندسی معیارها در حجم زیاد داده ها یکی از گلوگاه هاست و اگر بتوان معیارها را به صورت کارا و با دقت مناسب و در زمان کم بدست آورد، دقت نتایج خروجی نیزاز مطلوبیت بیشتری برخوردار خواهد شد.

**مراجع**

1. Peter.S.Fader, Bruce Hardie , Ka Lok Lee. (Vol. 24:No. 2, Spring 2007, pp. 275-284). “Counting your customers” the Easy Way: An alternative to thePareto/NBD model. Marketing Science.
2. Peter S. Fader, Bruce G.S. Hardie. ( November 2005 ). A Note on Deriving the Pareto/NBD Model and Related Expressions
3. Babak Sohrabi, Amir Khanlari. ( Spring 2007,Vol. 14 No. 47, pp 7- 20 ) Customer lifetime value (CLV) measurement based on RFM model
4. Biswajit Pal , Ritwik Sinha , Abhishek Saha , Peter Jaumann and Subhasish Misra1.(2012 ). Customer Targeting Framework: Scalable Repeat Purchase Scoring Algorithm for Large Databases - Proceedings of 4th International Conference on Machine Learning and Computing
5. Galit Shmueli. (2010, Vol. 25, No. 3, 289–31). To Explain or to Predict?
6. Andrew S. C. Ehrenberg. (2000). Repeat Buying ( part two , chapter 3): "The repeat-buying structure of a market"
7. Jayanta Kumar Pal, Abhisek Saha, Subhasish Misra.(No 85, July 2010). Customer repeat purchase modeling- A Bayesian HierarchicalFramework. HP Labs technical report.
8. Ehrenberg, A (2000) "Repeat Buying", Journal of Empirical Generalisations in Marketing Science, Vol 5, No.2
9. Andrew S. C. Ehrenberg. (2000). Repeat Buying ( part four, chapter 7): "The NBD Theory"
10. Berger, P. D. and Nasr, N. I. (1998). "Customer lifetime value: marketing models and applications", Journal of Interactive Marketing, 12(1), pp.17- 30.
11. Pfeifer, P. E. and Carraway, R. L. (2000). "Modeling customer relationships as Markov Chains", Journal of Interactive Marketing, 14(2), pp.43- 55.
12. Gupta, S. and Lehmann, D. R. (2003). "Customers as assets", Journal of Interactive Marketing, 17(1), pp.9- 24.
13. Peter S. Fader Bruce G.S. Hardie Ka Lok Lee. (July 2004 ).RFM and CLV: Using Iso-value Curves for Customer Base Analysis
14. Wu, Couchen, Hsiu-Li Chen. (2000). Counting your customers: Compounding customer’s in-store decisions, interpurchase time, and repurchasing behavior. Eur. J. Oper. Res. 127(1) 109–119.

1. Insight [↑](#footnote-ref-1)
2. Pareto/NBD (Negative Bionomial Distribution ) [↑](#footnote-ref-2)
3. BG/NBD (Beta-Geometric / Negative Binomial Distribution) [↑](#footnote-ref-3)
4. Bayesian Hierarchical Model(B.H.M) [↑](#footnote-ref-4)
5. Kaggle-aquire-valued-shoppers-challenge [↑](#footnote-ref-5)
6. Philip Kotler [↑](#footnote-ref-6)
7. Gary Armstrong [↑](#footnote-ref-7)
8. Chaim Zins [↑](#footnote-ref-8)
9. Descriptive Model [↑](#footnote-ref-9)
10. Supervised Methods [↑](#footnote-ref-10)
11. Learning [↑](#footnote-ref-11)
12. Testing [↑](#footnote-ref-12)
13. Training Dataset [↑](#footnote-ref-13)
14. Test Dataset [↑](#footnote-ref-14)
15. Previously Unseen Data [↑](#footnote-ref-15)
16. Classification [↑](#footnote-ref-16)
17. Regression [↑](#footnote-ref-17)
18. Anomaly Detection [↑](#footnote-ref-18)
19. Unsupervised Methods [↑](#footnote-ref-19)
20. Clustering [↑](#footnote-ref-20)
21. Association Rules Mining [↑](#footnote-ref-21)
22. Sequential Pattern Discovery [↑](#footnote-ref-22)
23. Time Series Prediction [↑](#footnote-ref-23)
24. Credit Card Fraud Detection [↑](#footnote-ref-24)
25. Threshold [↑](#footnote-ref-25)
26. Intra-Cluster Density [↑](#footnote-ref-26)
27. Inter-Cluster Density [↑](#footnote-ref-27)
28. Item Sets [↑](#footnote-ref-28)
29. Classifier [↑](#footnote-ref-29)
30. Parametric [↑](#footnote-ref-30)
31. NonParametric [↑](#footnote-ref-31)
32. Semi Parametric [↑](#footnote-ref-32)
33. Multivariate adaptive regression splines [↑](#footnote-ref-33)
34. Pareto principle [↑](#footnote-ref-34)
35. law of the vital few [↑](#footnote-ref-35)
36. principle of factor sparsity [↑](#footnote-ref-36)
37. Joseph M. Juran [↑](#footnote-ref-37)
38. Vilfredo Pareto [↑](#footnote-ref-38)
39. Schmittlein [↑](#footnote-ref-39)
40. Heterogeneity [↑](#footnote-ref-40)
41. Shape Parameter [↑](#footnote-ref-41)
42. Scale Parameter [↑](#footnote-ref-42)
43. Gaussian Hyper-Geometric Function [↑](#footnote-ref-43)
44. Survival Function [↑](#footnote-ref-44)
45. Posterior Distribution [↑](#footnote-ref-45)
46. Updated Probability Estimate [↑](#footnote-ref-46)
47. Prior Distribution [↑](#footnote-ref-47)
48. Hyper-parameter [↑](#footnote-ref-48)
49. Hyper-prior Distribution [↑](#footnote-ref-49)
50. Markov chain Monte Carlo [↑](#footnote-ref-50)
51. Regression based Approximation [↑](#footnote-ref-51)
52. productquantity [↑](#footnote-ref-52)
53. purchaseamount [↑](#footnote-ref-53)
54. Monetary [↑](#footnote-ref-54)
55. Recency - Frequency - Monetary [↑](#footnote-ref-55)
56. Essential data [↑](#footnote-ref-56)
57. Purchase recency [↑](#footnote-ref-57)
58. Purchase frequency [↑](#footnote-ref-58)
59. Length Recency Frequnecy Monetary [↑](#footnote-ref-59)
60. Feature Engineering [↑](#footnote-ref-60)
61. Training Model [↑](#footnote-ref-61)
62. Cut-off [↑](#footnote-ref-62)
63. Discrimination Threshold [↑](#footnote-ref-63)
64. Receiver operating characteristic [↑](#footnote-ref-64)
65. Area under the curve [↑](#footnote-ref-65)