فصل‌نامه مطالعاتی در مدیریت بانکی و بانکداری اسلامی

مؤسسه عالی آموزش بانکداری ایران
تابستان ۱۴۰۴، دوره ۱۱، شماره ۳۱، ۷۸-۳۱.

اعتبارسنجی هوشمند و پیش بینی رفتار آینده مشتری بانک: رویکرد مبتنی بر داده با مدل RFM و خوشه بندی

سهیل زارعی[۱] علی مهدیان[۲] رضا صحبت‌زاده[۳] و زهرا مصلاح[۴]

چکیده

با افزایش حجم داده‌ها در دنیای امروز، بهره‌گیری از علم داده برای یافتن راه‌حل‌های بهینه امری ضروری است. یکی از چالش‌های اساسی در صنعت بانکداری، اطمینان از بازگشت تسهیلات اعطایی و تخصیص بهینه منابع مالی است. بنابراین، سنجش سطح اعتباری مشتری و پیش‌بینی آن در آینده می‌تواند شاخصی سودآور و قابل‌اعتماد برای بانک‌ها باشد. در مطالعه پیش رو، با استفاده از تراکنش‌های مالی مشتریان مؤسسه اعتباری ملل، سطح اعتباری حال حاضر ایشان به وسیله تجزیه و تحلیل مدل RFM و استفاده از تکنیک‌های خوشه‌بندی انجام می‌پذیرد. در نوآوری انجام‌شده، سطح اعتباری آینده مشتری نیز با استفاده از الگوریتم یادگیری ماشین جنگل تصادفی و بر مبنای خروجی خوشه‌بندی مبتنی بر مدل پیش‌بینی می‌گردد. این رویکرد در مقایسه با روش‌های سنتی مبتنی بر امتیازدهی ایستا، دقت پیش‌بینی را افزایش داده و به بانک امکان می‌دهد تصمیم‌گیری‌های اعتباری را با اطمینان بیشتری انجام دهد. در این مطالعه با استفاده از شاخص ارزیابی معیار اطلاع بیزی پنج خوشه بهینه با استفاده از مدل RFM شناسایی شد. همچنین یک مدل جنگل تصادفی با شاخص‌های دقت میانگین ۷۲ درصد، بازیابی ۷۵ درصد و امتیاز میانگین متوازن دقت و بازخوانی ۹۱ درصد، جهت پیش‌بینی سطح اعتباری آینده مشتری طراحی گردید. این پژوهش با تکیه بر فناوری‌های نوین مالی، چارچوبی داده‌محور برای اعتبارسنجی هوشمند در نظام بانکداری اسلامی ارائه می‌دهد و به کاهش ریسک اعتباری در فرایند تخصیص تسهیلات کمک می‌کند.

واژه‌های کلیدی: اعتبار سنجی، یادگیری ماشین، خوشه بندی، RFM، فناوری‌های نوین مالی.

طبقه‌بندی JEL: *G21, G28, C38, C55, G32*

---

۱. معاون اداره مهندسی داده شرکت فناوری اطلاعات و تجارت الکترونیک ملل (فام)، تهران، ایران(نویسنده مسئول). Soheil.zarei@gmail.com

۲. کارشناس اداره مهندسی داده شرکت فناوری اطلاعات و تجارت الکترونیک ملل (فام)، تهران، ایران. Ali.mahdian72@gmail.com

۳. کارشناس اداره مهندسی داده شرکت فناوری اطلاعات و تجارت الکترونیک ملل (فام)، تهران، ایران. rezasohbatzadeh@gmail.com

۴ . کارشناس اداره مهندسی داده شرکت فناوری اطلاعات و تجارت الکترونیک ملل (فام)، تهران، ایران. zmosallah@yahoo.com

۱. مقدمه

با رشد تصاعدی تعداد و حجم تراکنش‌های مالی، صنعت بانکداری برای حفظ جایگاه خود در بازار و ارائه خدمات با کیفیت، نیازمند بهره‌گیری از فناوری‌های نوین مالی و حرکت به سمت بانکداری هوشمند است. در این پژوهش، منظور از هوشمندی در اعتبارسنجی، استفاده از مدل‌های داده‌محور و الگوریتم‌های یادگیری ماشین برای شناسایی الگوهای رفتاری و پیش‌بینی سطح اعتباری آینده مشتریان است. این مفهوم شامل سه بُعد اصلی است: تحلیل خودکار داده‌های تراکنشی مشتریان با استفاده از مدل‌های آماری و یادگیری ماشین، یادگیری تطبیقی از تغییرات رفتاری مشتریان برای به‌روزرسانی مدل اعتبارسنجی، و پشتیبانی از تصمیم‌گیری مدیریتی در تخصیص منابع اعتباری به‌صورت داده‌محور.

در چارچوب بانکداری اسلامی، که بر اصول عدالت و اجتناب از ربا تأکید دارد، این فناوری‌های نوین می‌توانند تنظیم‌گری را تقویت کنند و از هزینه‌های اضافی ناشی از تسهیلات پرریسک جلوگیری نمایند. بررسی فهرست‌های منتشرشده از تسهیلات کلان بانکی (بانک مرکزی جمهوری اسامی ایران، ۱۴۰۳) نیز نشان می‌دهد تمرکز بیش از حد تسهیلات در گروه محدودی از مشتریان، چالشی جدی در عدالت توزیعی نظام بانکی است که نیازمند به‌کارگیری ابزارهای داده‌محور برای شفاف‌سازی هوشمند می‌باشد. در چنین محیط داده‌محوری، چالش اصلی نه تنها حجم بالای داده‌ها بلکه تعدد متغیرهای تصمیم‌گیرنده و روابط پیچیده میان آن‌ها است که ارزیابی دستی یا تک‌متغیره را غیرممکن می‌سازد. از این‌رو، در مدل پیشنهادی، خوشه‌بندی مبتنی بر مدل برای کاهش ابعاد و خوشه‌بندی نمونه‌های مشابه به‌کار گرفته شده تا ناهمگنی داده‌ها کاهش یابد و الگوهای رفتاری اصلی استخراج شود. سپس، الگوریتم جنگل تصادفی با بهره‌گیری از ترکیب‌های مختلف متغیرهای رفتاری، امکان تحلیل هم‌زمان اثر چندین شاخص را فراهم کرده و به شناسایی دقیق‌تر مشتریان در معرض ریسک کمک می‌کند. این ترکیب روش‌ها موجب می‌شود فرآیند اعتبارسنجی از سطح توصیف داده‌ها به سطح تحلیل تصمیم‌یار ارتقا یابد. از همین رو، هوشمندسازی در حوزه‌هایی چون مدیریت ریسک، شخصی‌سازی خدمات، خودکارسازی فرآیندها و ارزیابی اعتباری مشتریان، می‌تواند موجب افزایش دقت تصمیم‌گیری، کاهش هزینه‌ها و ارتقای بهره‌وری شود. در چارچوب بانکداری اسلامی، اعتبارسنجی صرفاً ارزیابی توان مالی

نیست، بلکه ابزاری برای تحقق عدالت در دسترسی به منابع مالی و پیشگیری از تمرکز ریسک در گروه‌های خاص مشتریان است. از این‌رو، مدل پیشنهادی پژوهش حاضر علاوه بر جنبه تحلیلی، با هدف حمایت از تصمیم‌گیری‌های هوشمند در تخصیص منابع مالی طراحی شده است. به طور خلاصه، بانک‌ها همواره به دنبال ارتقای شیوه‌های بازاریابی و بهبود عملکرد مشتریان فعلی هستند، چراکه هزینه جذب مشتری جدید حدود پنج برابر هزینه حفظ مشتریان موجود است (Kumar, 2022). در بانکداری سنتی، اعتبارسنجی مبتنی بر داده‌های محدود و سوابق پرونده مشتری صورت می‌گرفت (Chiorazzo et al, 2018)؛ اما در بانکداری هوشمند، این فرآیند بایستی به کمک تحلیل داده‌های کلان و ابزارهای یادگیری ماشین ارتقا یابد تا ریسک‌های مالی کاهش یافته و نظام بانکی از هزینه‌های اضافی مصون بماند. بر اساس گزارش رسمی بانک مرکزی، نسبت مطالبات غیرجاری شبکه بانکی کشور در پایان سال ۱۴۰۲ معادل ۷٫۵ درصد اعلام شده است (بانک مرکزی جمهوری اسلامی ایران، ۱۴۰۱) که این امر ضرورت استفاده از مدل‌های داده‌محور در ارزیابی ریسک اعتباری را دوچندان می‌کند. در این میان، هوشمندسازی فرآیند اعتبارسنجی، علاوه بر مدیریت ریسک اعتباری، امکان شناسایی مشتریان وفادار، طراحی بسته‌های پیشنهادی متناسب و افزایش بازدهی سرمایه را فراهم می‌کند. از طرف دیگر، هوشمندسازی از طریق بهینه‌سازی بازار هدف مانند ارائه بسته‌های پیشنهادی و تشویق افراد به فعالیت بیشتر و شناسایی افراد با توانایی بازپرداخت تسهیلات اعطایی، در رشد سیستم بانکی کمک خواهد نمود. در این پژوهش با ارائه مدلی داده محور تلاش خواهیم کرد تا ضمن دسته‌بندی مشتریان و در نهایت پیش‌بینی وضعیت هر مشتری در آینده، گامی در جهت هوشمندسازی اعتباری برداریم.

۱-۱. معرفی مسئله

افزایش جمعیت و گسترش مبادلات مالی در بسترهای دیجیتال، نظام بانکی را با چالش‌های جدی در مدیریت منابع و اعطای تسهیلات مواجه کرده است. طبق گزارش رسمی بانک مرکزی جمهوری اسلامی ایران، نسبت مطالبات غیرجاری به کل تسهیلات در نظام بانکی کشور حدود ۵ تا ۷ درصد برآورد می‌شود و میزان بدهی غیرجاری بانک‌ها تا پایان سال ۱۴۰۳ به بیش از ۷۷۹ هزار میلیارد تومان

رسیده است (بانک مرکزی جمهوری اسلامی ایران، ۱۴۰۴). این آمارها نشان می‌دهد که ارزیابی دقیق سطح اعتباری مشتریان نه‌تنها از منظر سودآوری بانک‌ها بلکه به‌عنوان ابزاری برای کاهش ریسک سیستماتیک و ارتقای سلامت مالی کشور ضرورت دارد. از آنجا که مشتریان از نظر اعتبار و توان بازپرداخت در سطح یکسانی قرار ندارند، شناسایی وضعیت اعتباری آنان به یکی از الزامات تحلیل رفتار مشتریان تبدیل شده است (Greenberg, 2001). مطابق با گزارش سیاست پولی مرکز پژوهش‌های مجلس، حجم مطالبات معوق شبکه بانکی کشور به بیش از ۵۸۵ هزار میلیارد تومان رسیده است (مرکز پژوهش‌های مجلس شورای اسلامی, ۱۴۰۳)؛ آماری که بیانگر ضعف در نظام ارزیابی اعتباری سنتی و لزوم بهره‌گیری از مدل‌های داده‌محور برای پیش‌بینی رفتار بازپرداخت مشتریان است. این موضوع نه تنها برای تخصیص بهینه منابع بانکی ضروری است، بلکه به سیاست‌گذاران و تنظیم‌گران مالی نیز در جهت کنترل ریسک‌های سیستمی و ارتقای سلامت بازار پول کمک می‌کند. بنابراین پرسش اصلی پژوهش حاضر آن است که چگونه می‌توان به کمک مدل‌های داده‌محور، دسته‌بندی بهینه‌ای از مشتریان بانکی به دست آورد و آیا امکان پیش‌بینی سطح اعتباری آینده مشتریان بر اساس داده‌های تراکنشی گذشته وجود دارد یا خیر. با توجه به حجم بالای داده‌ها و پیچیدگی روابط میان شاخص‌های رفتاری مشتریان، انتخاب روش‌های تحلیل در این پژوهش مبتنی بر کاهش ابعاد، کشف الگوهای پنهان و افزایش تفسیرپذیری بوده است. در این راستا، از مدل خوشه‌بندی مبتنی بر مدل آماری برای خوشه‌بندی مشتریان با ویژگی‌های تراکنشی مشابه استفاده شد تا ناهمگنی داده‌ها کاهش یابد و ساختارهای پنهان میان متغیرها آشکار گردد. سپس الگوریتم جنگل تصادفی به‌عنوان ابزار پیش‌بینی انتخاب گردید تا ترکیب‌های گوناگون متغیرهای رفتاری را بدون فرض خطی‌بودن روابط بررسی کند و اهمیت نسبی هر متغیر را در پیش‌بینی سطح اعتباری مشتری مشخص سازد. متغیرهای ورودی مدل شامل شاخص‌های سه‌گانه‌ی RFM در بازه‌های ماهانه، روند تغییرات تراکنش‌ها و میانگین نوسانات ماهانه است که نمایانگر پویایی رفتار مالی مشتریان می‌باشند.

۱-۲. هدف و نوآوری پژوهش

پژوهش حاضر با بهره‌گیری از مدل RFM برای استخراج شاخص‌های رفتاری مشتریان، استفاده از الگوریتم‌های خوشه‌بندی برای بخش‌بندی داده‌ها و به‌کارگیری روش‌های یادگیری ماشین در پیش‌بینی سطح اعتباری آینده، چارچوبی داده‌محور برای اعتبارسنجی هوشمند ارائه می‌دهد. نوآوری اصلی این مطالعه در ترکیب سه رویکرد مکمل است:

- تحلیل RFM به عنوان ابزاری ساده و قابل‌فهم برای مدیران بانکی (Ho et al., 2023).

- خوشه‌بندی مبتنی بر مدل‌های آماری به منظور شناسایی الگوهای پنهان در داده‌های بانکی (Feczko & Fair, 2020; Han & Kamber, 2001; Ngai et al., 2009).

– مدل‌های یادگیری ماشین (جنگل تصادفی) برای پیش‌بینی تغییرات آتی در سطح اعتباری مشتریان (Breiman, 2001).

نتایج این پژوهش می‌تواند علاوه بر کمک به بانک‌ها در تصمیم‌گیری‌های اعتباری، برای نهادهای تنظیم‌گر نیز ارزشمند باشد، زیرا امکان طراحی سیاست‌های نظارتی هوشمند مبتنی بر داده را فراهم می‌آورد و به بهبود پایداری نظام مالی کمک می‌کند. از دیگر جنبه‌های نوآورانه این پژوهش، بهره‌گیری هم‌زمان از دو شاخص تحلیلی مکمل است. نخست، استفاده از معیار اطلاع بیزی برای تعیین تعداد بهینه خوشه‌ها که امکان ارزیابی عینی کیفیت مدل‌های خوشه‌بندی را فراهم می‌سازد و در مطالعات بانکی داخلی کمتر مورد استفاده قرار گرفته است. دوم، به‌کارگیری شاخص ارزش طول عمر مشتری به‌عنوان معیاری اقتصادی برای سنجش ارزش اعتباری هر خوشه که ضمن تفسیر دقیق‌تر نتایج، به پیوند میان تحلیل داده و تصمیم‌سازی اعتباری کمک می‌کند. هرچند در برخی مطالعات، ترکیب الگوریتم‌های خوشه‌بندی و جنگل تصادفی در زمینه‌ی تحلیل رفتار مشتریان یا طبقه‌بندی ریسک اعتباری به‌کار رفته است، تمایز این پژوهش در نحوه‌ی تلفیق و سطح تحلیل آن است. در این مدل، خروجی خوشه‌بندی مبتنی بر مدل نه‌تنها برای تقسیم‌بندی آماری مشتریان بلکه به‌عنوان متغیر ورودی ساختاری در مدل یادگیری ماشین استفاده شده است. بدین ترتیب، مدل از الگوهای آماری استخراج‌شده در خوشه‌ها برای بهبود دقت پیش‌بینی استفاده می‌کند. افزون بر آن، محاسبه‌ی شاخص ارزش طول

عمر مشتری در کنار معیار اطلاع بیزی به مدل امکان می‌دهد تا از هر دو منظر آماری و اقتصادی رفتار مشتریان را تحلیل نماید. این رویکرد چندمرحله‌ای، تمایز کلیدی پژوهش حاضر نسبت به مطالعات مشابه محسوب می‌شود.

۲. مبانی نظری پژوهش

موضوع اعتبارسنجی مشتریان بانکی و استفاده از روش‌های هوش مصنوعی و یادگیری ماشین برای این منظور یکی از زمینه‌های تحقیقاتی پرکاربرد در صنعت بانکداری است. برهمین اساس، آشنایی با مفاهیم و روش‌های دسته‌بندی مشتریان می‌تواند ما را به سمت انتخاب مناسب‌ترین و بهترین الگوریتم دسته‌بندی متناسب با داده‌های موجود رهنمون سازد.

۲-۱. روش‌های دسته‌بندی مشتریان

دسته‌بندی رفتاری مشتریان یکی از ابزارهای کلیدی در مدیریت ارتباط با مشتری و طراحی استراتژی‌های بازاریابی و اعتباری است. بر اساس مرور مطالعات، چهار رویکرد اصلی در دسته‌بندی مشتریان معرفی شده است: رفتاری، جمعیت‌شناختی، جغرافیایی و روان‌شناختی. دسته‌بندی رفتاری بر اساس تحلیل عملکرد و الگوهای رفتاری مشتریان شکل می‌گیرد. در این روش، مشتریانی که الگوهای تراکنشی مشابه دارند در یک گروه و مشتریانی با رفتارهای متفاوت در گروه‌های مجزا قرار می‌گیرند (Lancaster & Williams, 2002). دسته‌بندی جمعیت‌شناختی مبتنی بر ویژگی‌های فردی همچون سن، جنسیت، درآمد و ملیت است و اغلب در تحلیل‌های کلان بازار مورد استفاده قرار می‌گیرد (Beheshtian-Ardakani et al., 2018). دسته‌بندی جغرافیایی بر مبنای منطقه جغرافیایی مشتریان عمل می‌کند و در مواردی که محل سکونت یا فعالیت اقتصادی نقش مؤثری در رفتار مالی دارد، کاربرد فراوانی دارد (Tsiptsis & Chorianopoulos, 2011). دسته‌بندی روان‌شناختی بر اساس ارزش‌ها، باورها، نگرش‌ها، علایق و سبک زندگی مشتریان صورت می‌گیرد و می‌تواند در طراحی محصولات و خدمات شخصی‌سازی‌شده بسیار مؤثر باشد (Liu et al., 2009). به‌طور کلی

انتخاب هر یک از این روش‌ها وابسته به هدف کسب‌وکار، نوع داده‌های در دسترس و ماهیت تصمیمات مدیریتی است. در پژوهش حاضر تمرکز بر دسته‌بندی رفتاری مشتریان با استفاده از داده‌های تراکنش مالی قرار گرفته است، زیرا این داده‌ها بیشترین ارتباط را با ارزیابی و پیش‌بینی سطح اعتباری مشتریان بانکی دارند. انتخاب مدل RFM در این پژوهش بر مبنای ویژگی‌های خاص داده‌های بانکی و هدف پژوهش انجام شده است. این مدل به دلیل سادگی در محاسبه، تفسیرپذیری بالا و قابلیت اجرا در محیط‌های عملیاتی بانک، یکی از پرکاربردترین چارچوب‌ها در تحلیل رفتار مشتریان است. از آنجا که در نظام بانکی ایران، معمولاً داده‌های دموگرافیک و سوابق جامع مالی مشتریان به‌صورت کامل در دسترس نیست، مدل RFM امکان تحلیل رفتار تراکنشی بر اساس داده‌های واقعی درون‌بانکی را فراهم می‌سازد. مزیت اصلی RFM نسبت به مدل‌های دیگر، انعطاف‌پذیری آن در شرایط کم‌داده و قابلیت تلفیق با روش‌های یادگیری ماشین برای پیش‌بینی رفتار آینده مشتری است. در ادامه، برای شناسایی گروه‌های رفتاری متمایز در میان مشتریان، از خوشه‌بندی مبتنی بر مدل استفاده شده است تا الگوهای پنهان میان شاخص‌های RFM آشکار شود و مشتریان با رفتارهای مشابه در یک گروه قرار گیرند. این ترکیب روش‌ها به‌طور هم‌زمان دقت تحلیلی و تفسیرپذیری مدیریتی را افزایش می‌دهد و با نیاز بانک‌ها در تصمیم‌گیری‌های اعتباری سازگار است.

۲-۲. شاخص‌های کلیدی در دسته بندی مشتریان

یک دسته بندی زمانی اثربخش و قابل استفاده است که ویژگی‌های مشخصی داشته باشد. ادبیات پژوهش نشان می‌دهد که دسته بندی مطلوب مشتریان باید قابل شناسایی، در دسترس، قابل توجه، پایدار، متمایز و قابل اجرا باشد (Gavett, 2014). قابل شناسایی بودن به معنای امکان تشخیص ویژگی‌های متمایز هر بخش از مشتریان است. در دسترس بودن بیانگر قابلیت دسترسی به مشتریان هر بخش از طریق کانال‌های ارتباطی و خدماتی است. قابل توجه بودن نشان می‌دهد که هر بخش باید از نظر حجم یا ارزش اقتصادی آن‌قدر مهم باشد که استراتژی جداگانه‌ای برای آن تدوین شود. پایداری به دوام نسبی ویژگی‌ها و رفتار مشتریان در یک بازه زمانی اشاره دارد. تمایز بیان می‌کند که هر بخش باید رفتارها و ویژگی‌هایی متفاوت از سایر بخش‌ها داشته باشد. در نهایت، قابلیت اجرا به

امکان طراحی و پیاده‌سازی راهبردهای بازاریابی یا اعتباری متناسب با هر بخش اشاره دارد. رعایت این شاخص‌ها موجب می‌شود که بانک‌ها و نهادهای مالی بتوانند دسته بندی معناداری انجام دهند که برای تصمیم‌گیری تجاری سودمند باشد. در پژوهش حاضر، این اصول به‌عنوان مبنای انتخاب روش RFM برای دسته‌بندی مشتریان مورد توجه قرار گرفته‌اند.

۲-۳. مدل RFM

یکی از پرکاربردترین و در عین حال ساده‌ترین مدل‌ها در تحلیل مشتریان و دسته‌بندی رفتاری مشتریان، مدل RFM است. این مدل بر سه شاخص اصلی بنا شده است:

۱. تازگی[۱]: فاصله زمانی از آخرین تراکنش مالی موفق مشتری.

۲. تکرار یا فراوانی[۲]: تعداد تراکنش‌های مالی موفق در یک بازه زمانی مشخص.

۳. مبلغ تراکنش مالی[۳]: مجموع مبالغ تراکنش‌های مالی موفق مشتری.

در این پژوهش، فرآیند RFM به‌عنوان گام نخست دسته‌بندی مشتریان بر اساس شاخص‌های تازگی، فراوانی و مبلغ تراکنش مالی در نظر گرفته شده است، در حالی‌که مرحله خوشه‌بندی در گام بعد برای شناسایی الگوهای پنهان میان مشتریانِ دارای ویژگی‌های رفتاری مشابه به کار می‌رود. به این ترتیب، دسته‌بندی مفهومی در سطح شاخص‌ها و خوشه‌بندی آماری در سطح گروه‌های مشتریان تعریف شده است. مدل RFM به‌دلیل سادگی، قابلیت درک آسان برای مدیران بانکی و امکان اجرای سریع، به طور گسترده در اعتبارسنجی و تحلیل ارزش مشتریان مورد استفاده قرار می‌گیرد (Ho et al., 2023). در این روش، به هر مشتری براساس سه شاخص فوق امتیازی از یک تا پنج اختصاص داده می‌شود و بدین ترتیب امکان دسته بندی مشتریان به گروه‌های مختلف فراهم می‌شود. با وجود کارایی

---

[۱] Recency
[۲] Frequency
[۳] Monetary

بالا، مدل سنتی RFM محدودیت‌هایی دارد. از جمله می‌توان به تعداد کم شاخص‌ها و نادیده گرفتن برخی ابعاد رفتاری مشتریان اشاره کرد. به همین دلیل، در مطالعات اخیر تغییراتی در این مدل اعمال شده است؛ به‌طور مثال، برخی پژوهش‌ها بیشترین وزن را به فراوانی تراکنش‌ها اختصاص داده و کمترین وزن را به مبلغ تراکنش‌ها داده‌اند، در حالی‌که در پژوهش‌های دیگر ترتیب اولویت برعکس بوده است (Breiman, 2001; Rizkyanto & Ford, 2023). در پژوهش حاضر، از مدل RFM به‌عنوان یک چارچوب داده‌محور برای استخراج ویژگی‌های رفتاری مشتریان و مبنای اجرای الگوریتم‌های خوشه‌بندی استفاده شده است.

۲-٤. خوشه‌بندی در تحلیل مشتریان

خوشه‌بندی یکی از مهم‌ترین تکنیک‌های یادگیری ماشین بدون ناظر است که هدف آن تقسیم یک مجموعه داده ناهمگن به چندین گروه همگن می‌باشد. در این فرآیند، اعضای هر خوشه بیشترین شباهت را با یکدیگر و بیشترین تفاوت را با اعضای سایر خوشه‌ها دارند (Feczko & Fair, 2020). این روش به‌طور گسترده در حوزه‌هایی نظیر تحلیل مشتریان، تشخیص الگو، بازاریابی و حتی امنیت داده مورد استفاده قرار گرفته است (Han & Kamber, 2001; Ngai et al., 2009). در زمینه بانکداری و مدیریت مشتریان، خوشه‌بندی به بانک‌ها کمک می‌کند تا مشتریان خود را در گروه‌های رفتاری متمایز دسته‌بندی کرده و برای هر گروه راهبردهای متفاوتی همچون طراحی محصولات اعتباری خاص یا برنامه‌های وفاداری هدفمند تدوین کنند.روش‌های خوشه‌بندی متنوعی در ادبیات معرفی شده‌اند که مهم‌ترین آن‌ها عبارتند از:

خوشه‌بندی تفکیکی[۱] الگوریتم‌هایی مانند کا- میانگین [۲]که براساس میانگین فاصله داده‌ها عمل می‌کنند. این روش‌ها کارآمد و سریع هستند اما نیازمند تعیین تعداد خوشه از پیش بوده و به انتخاب نقاط اولیه حساس می‌باشند (Sander, 2017; Zhu & Ma, 2018).

خوشه‌بندی سلسله‌مراتبی[۳] از کوچک‌ترین خوشه‌ها آغاز می‌شود و به‌صورت تدریجی داده‌ها را ادغام می‌کند. این روش امکان محاسبه خودکار تعداد خوشه‌ها را فراهم می‌آورد (Nielsen, 2016).

خوشه‌بندی مبتنی بر چگالی[۴] الگوریتم‌هایی نظیر DBSCAN قادرند خوشه‌های غیرکروی و با اشکال پیچیده را شناسایی کنند و برای داده‌های با توزیع نامتقارن مناسب‌اند (Ester et al., 1996).

خوشه‌بندی مبتنی بر مدل[۵] این رویکرد با بهره‌گیری از مدل‌های آماری آمیخته، احتمال تعلق هر داده به یک خوشه را برآورد می‌کند. مدل آمیخته نرمال[۶] از جمله روش‌های پرکاربرد در این دسته است و به دلیل انعطاف‌پذیری بالا در داده‌های پیچیده مورد استفاده قرار می‌گیرد (Dempster et al., 1977; McLachlan & Basford, 1988). در پژوهش حاضر، از خوشه‌بندی مبتنی بر مدل به‌عنوان روش اصلی استفاده شده است، زیرا این الگوریتم محدودیت شکل خوشه‌ها را برطرف کرده و با معیارهای آماری همچون اطلاع بیزی قادر به تعیین تعداد بهینه خوشه‌ها می‌باشد.

۲-۵. ارزش طول عمر مشتری[۷]

یکی از شاخص‌های پرکاربرد در تحلیل و دسته بندی مشتریان، ارزش طول عمر مشتری است. ارزش طول عمر مشتری، بیانگر ارزش فعلی تمامی سودهای آینده‌ای است که یک مشتری در طول رابطه

---

[۱] Partitioning clustering
[۲] K-means
[۳] Hierarchical clustering
[۴] Density-Based clustering
[۵] Model-Based clustering
[۶] Gaussian mixture model
[۷] Customer Lifetime Value (CLV)

خود با سازمان ایجاد خواهد کرد. این شاخص علاوه بر سنجش سودآوری مشتریان، به تصمیم‌گیرندگان کمک می‌کند تا در مورد سرمایه‌گذاری بر روی جذب یا نگهداری مشتریان تصمیمات دقیق‌تری اتخاذ کنند (Saedi et al., 2025). مطالعات نشان می‌دهد که استفاده از ارزش طول عمر مشتری می‌تواند به افزایش وفاداری مشتریان، کاهش نرخ ریزش و ارتقای سودآوری بانک‌ها منجر شود. پژوهشگران متعددی مدل‌های گوناگونی برای محاسبه ارزش طول عمر مشتری ارائه کرده‌اند. به‌عنوان نمونه، توماس و رینارتز چارچوب‌هایی معرفی کرده‌اند که برای جذب و نگهداری مشتریان و سنجش ارزش آن‌ها به‌کار می‌رود (Reinartz et al., 2005; Taherkhani et al., 2025). همچنین فیدر و همکارانش مدلی مبتنی بر سه شاخص تازگی تراکنش، تعداد تراکنش و مبلغ تراکنش پیشنهاد داده‌اند که از نظر ساختار داده‌ای با مدل RFM هم‌پوشانی دارد (Fader et al., 2004)، هرچند از نظر هدف تحلیلی متفاوت است. مدل RFM عمدتاً برای دسته‌بندی مشتریان بر اساس رفتار گذشته به‌کار می‌رود، در حالی‌که مدل ارزش طول عمر مشتری با استفاده از همین شاخص‌ها تلاش می‌کند ارزش اقتصادی آینده مشتری را برآورد کند. به همین دلیل، در پژوهش حاضر از ارزش طول عمر مشتری به‌عنوان شاخصی مکمل برای مدل RFM استفاده شده است تا تحلیل رفتاری و مالی مشتریان به‌صورت همزمان مورد ارزیابی قرار گیرد. با وجود تفاوت در روش‌های محاسبه، هدف مشترک مدل‌های ارزش طول عمر مشتری کمک به سازمان‌ها و نهادهای مالی برای تخصیص بهینه منابع و تمرکز بر مشتریان با ارزش بالاتر است. در این پژوهش، ارزش طول عمر مشتری به‌عنوان معیاری مکمل برای تحلیل نتایج خوشه‌بندی و پیش‌بینی اعتبار مشتریان مورد استفاده قرار گرفته است.

۲-۶. مدل جنگل تصادفی[1] در پیش‌بینی اعتبار مشتریان

جنگل تصادفی یکی از الگوریتم‌های پرکاربرد یادگیری ماشین است که بر پایه ترکیب چندین درخت تصمیم بنا شده است. در این روش، مجموعه داده به‌طور تصادفی به نمونه‌های کوچک‌تر تقسیم می‌شود و برای هر نمونه یک درخت تصمیم ساخته می‌شود. سپس با میانگین‌گیری یا رأی‌گیری میان نتایج

[1] Random forest

درخت‌ها، پیش‌بینی نهایی به دست می‌آید. این رویکرد موجب افزایش دقت مدل و کاهش خطر بیش‌برازش می‌گردد (Breiman, 2001). از مهم‌ترین مزایای جنگل تصادفی می‌توان به توانایی مدیریت داده‌های بزرگ، پایداری در برابر داده‌های پرت و قابلیت پردازش متغیرهای متعدد اشاره کرد. در عین حال، پارامترهایی مانند تعداد درخت‌ها، عمق هر درخت و روش نمونه‌برداری نقش مهمی در عملکرد مدل ایفا می‌کنند. افزایش تعداد درخت‌ها دقت مدل را بالا می‌برد اما به هزینه زمانی و محاسباتی بیشتری نیاز دارد، در حالی که عمق بیش از حد درخت‌ها می‌تواند منجر به بیش‌برازش شود و عمق بسیار کم نیز به کم‌برازش منجر خواهد شد. با توجه به این ویژگی‌ها، جنگل تصادفی به‌طور گسترده در حوزه‌های مختلف از جمله اعتبارسنجی و پیش‌بینی رفتار مشتریان بانکی مورد استفاده قرار می‌گیرد(Lessmann et al., 2015; Verbraken et al., 2014). در پژوهش حاضر، این الگوریتم به‌عنوان مدل اصلی برای پیش‌بینی سطح اعتباری آینده مشتریان به‌کار گرفته شده است.

۲-۷. مراحل انجام پژوهش

فرآیند اعتبارسنجی هوشمند و پیش‌بینی سطح اعتباری آینده مشتریان در این پژوهش مطابق با شکل شماره ۱ در پنج گام اصلی طراحی و اجرا شده است:

درک مسئله و جمع‌آوری داده‌ها: در این مرحله، اهداف پژوهش شامل اعتبارسنجی فعلی مشتریان و پیش‌بینی آینده آن‌ها مشخص می‌شود. سپس داده‌های تراکنشی واقعی مشتریان که شامل تاریخ، تعداد و مبلغ تراکنش‌ها است گردآوری می‌گردد.

پیش‌پردازش داده‌ها: داده‌های گردآوری‌شده مورد پاک‌سازی قرار می‌گیرند تا مقادیر ناقص یا نادرست حذف شده و داده‌ها به فرمت استاندارد و قابل استفاده برای تحلیل تبدیل شوند.

استخراج شاخص‌های RFM: در این گام با بهره‌گیری از مدل RFM سه شاخص رفتاری اصلی مشتریان یعنی تازگی، فراوانی و مبلغ تراکنش مالی تراکنش‌ها محاسبه و برای هر مشتری امتیازدهی می‌شود.

خوشه‌بندی و تحلیل سطح اعتباری مشتریان: مشتریان بر اساس شاخص‌های RFM در خوشه‌های رفتاری مختلف دسته‌بندی می‌شوند. در ادامه، ارزش طول عمر مشتری به‌عنوان شاخص تکمیلی محاسبه و نتایج خوشه‌بندی تفسیر می‌گردد.

پیش‌بینی سطح اعتباری آینده مشتریان: در گام پایانی، از الگوریتم جنگل تصادفی به‌عنوان یک روش یادگیری ماشین استفاده می‌شود تا سطح اعتباری آینده مشتریان پیش‌بینی گردد. این مدل با بهره‌گیری از داده‌های تاریخی و الگوهای رفتاری استخراج‌شده، توانایی بالایی در کاهش ریسک و بهبود فرآیند تصمیم‌گیری اعتباری دارد.

ترکیب سه‌مرحله‌ای روش‌های به‌کاررفته در این پژوهش، به‌عنوان یک زنجیره تحلیلی تصمیم‌یار طراحی شده است؛ به‌طوری‌که ابتدا مدل RFM شاخص‌های رفتاری مشتریان را استخراج می‌کند، سپس خوشه‌بندی مبتنی بر مدل با استفاده از این شاخص‌ها الگوهای پنهان رفتاری را شناسایی کرده و نهایتاً خروجی حاصل از خوشه‌بندی به‌عنوان ورودی مدل جنگل تصادفی به کار می‌رود تا احتمال تغییر یا پایداری خوشه‌های اعتباری در آینده پیش‌بینی شود. این ترکیب گام‌به‌گام، تمایز اصلی پژوهش از رویکردهای صرفاً توصیفی RFM یا مدل‌های مجزای پیش‌بینی است و چارچوبی هوشمند برای تصمیم‌گیری اعتباری فراهم می‌آورد. همچنین در این پژوهش، خوشه‌بندی و مدل جنگل تصادفی دو مرحله‌ی مکمل از یک زنجیره تحلیلی هستند. خروجی خوشه‌بندی (برچسب خوشه‌ها) به‌عنوان یکی از ورودی‌های کلیدی در مدل جنگل تصادفی به‌کار رفته است تا الگوهای رفتاری شناسایی‌شده توسط مدل RFM در فرآیند پیش‌بینی سط ح اعتباری آینده مشتریان مورد استفاده قرار گیرد. به عبارت دیگر، جنگل تصادفی بر اساس شاخص‌های رفتاری استخراج‌شده از RFM و نتایج خوشه‌بندی آموزش داده شده است تا بتواند اهمیت هر شاخص را در تعیین سطح اعتبار استخراج نماید. بنابراین، شاخص‌های اعتبارسنجی به‌صورت داده‌محور و بر پایه‌ی نتایج مراحل قبلی شناسایی شده‌اند، نه از پیش تعیین‌شده. این ساختار موجب انسجام میان مراحل تحلیل و افزایش دقت مدل در پیش‌بینی سطح اعتباری مشتریان گردیده است.

شکل ۱. مراحل انجام پژوهش

۳. روش ها و یافته ها

در این بخش به تشریح روش‌ها و اقدامات صورت گرفته برروی داده‌های جمع آوری شده خواهیم پرداخت و روال طی شده در هر گام به تفصیل بیان خواهد شد.

۳-۱. جمع‌آوری داده‌های خام

در این پژوهش، به‌منظور توسعه یک مدل داده‌محور مبتنی بر RFM برای دسته‌بندی مشتریان بانکی و پیش‌بینی سطح اعتباری آینده آن‌ها، داده‌های مربوط به تراکنش کارت‌های بانکی مشتریان مؤسسه اعتباری ملل در بازه زمانی هشت ماهه گردآوری گردید. بازه زمانی داده‌ها از فروردین تا آبان ۱۴۰۳ را پوشش می‌دهد. در مجموع، داده‌ها شامل ۲۴۷۳۸۵ رکورد تراکنش مربوط به ۵۲۶۸ مشتری حقیقی فعال بوده است. در این مطالعه صرفاً داده‌های مربوط به مشتریان حقیقی مورد استفاده قرار گرفته است، زیرا رفتار مالی اشخاص حقیقی از نظر تعداد، مبلغ و الگوی زمانی تراکنش‌ها تفاوت معناداری با مشتریان حقوقی دارد و هدف پژوهش، اعتبارسنجی مبتنی بر رفتار فردی مشتریان است. اطلاعات خام شامل نوع تراکنش،

مبلغ تراکنش و زمان دقیق انجام تراکنش بوده و مبنای استخراج شاخص‌های رفتاری RFM قرار گرفته است. بخشی از داده‌های اولیه در جدول شماره ۱ نمایش داده شده است.

جدول ۱. اطلاعات اولیه کارت‌های بانکی

کارت	نوع تراکنش	مبلغ تراکنش (ریال)	زمان تراکنش
کارت A	واریز	۱,۵۵۰,۰۰۰	۱۴:۲۵:۳۲ - ۱۳۹۹/۰۶/۰۵
کارت B	برداشت	۲۷۵,۰۰۰	۱۹:۴۵:۱۹ - ۱۳۹۹/۰۱/۱۷
کارت C	برداشت	۳۶۰,۰۰۰	۱۲:۳۷:۱۲ - ۱۳۹۹/۰۹/۲۵
کارت D	برداشت	۲۵,۰۰۰,۰۰۰	۰۹:۱۱:۲۵ - ۱۳۹۹/۰۲/۰۴

همان‌گونه که در جدول شماره ۱ نشان داده شده است، هر رکورد تراکنش شامل سه بعد اصلی بوده و این ویژگی‌ها مبنای محاسبات بعدی در فرآیند پیش‌پردازش و تحلیل قرار گرفته‌اند.

۳-۲. پیش‌پردازش داده‌ها

داده‌های خام گردآوری‌شده پیش از ورود به مرحله تحلیل نیازمند پالایش و استانداردسازی بودند تا نتایج حاصل از مدل‌سازی معتبر و قابل اتکا باشند. در این مرحله اقدامات زیر صورت گرفت:

۱) پاک سازی داده‌ها: کارت‌هایی که در طول بازه زمانی هشت‌ماهه حضور نداشتند (مانند

کارت‌های صادرشده یا منقضی‌شده در همان دوره)، حذف شدند. به این ترتیب تنها کارت‌هایی وارد فرآیند پژوهش شدند که در کل بازه زمانی حضور داشتند.

۲) تجمیع داده‌ها: به منظور کاهش حجم داده‌ها و تسهیل در تحلیل، اطلاعات تراکنش مشتریان به‌صورت ماهانه تجمیع گردید. شاخص‌های استخراج‌شده در این مرحله شامل:

۲-۱) مجموع مبلغ تراکنش‌های واریزی هر کارت در هر ماه

۲-۲) مجموع مبلغ تراکنش‌های برداشت

۲-۳) تعداد تراکنش‌های واریزی و برداشت

۲-۴) زمان آخرین تراکنش ماهانه هر کارت

بخشی از داده‌های تجمیع‌شده در جدول شماره ۲ نمایش داده شده است.

جدول۲. اطلاعات تجمیع شده کارت‌های بانکی

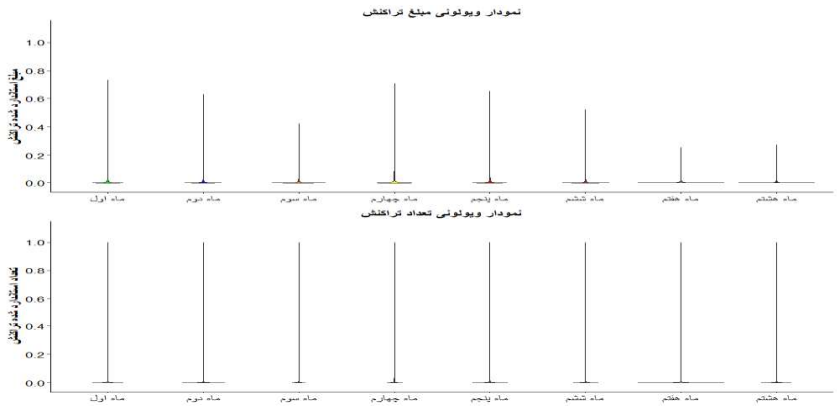
کارت	ماه اول				ماه دوم	ماه سوم	ماه چهارم	ماه پنجم	ماه ششم	ماه هفتم	ماه هشتم
کارت A	مجموع مبلغ تراکنش‌های واریزی	مجموع مبلغ تراکنش‌های برداشت شده	تعداد تراکنش‌های واریزی	تعداد تراکنش‌های برداشت شده	زمان آخرین تراکنش در ماه	...	...	...	...	...	...

نرمال‌سازی داده‌ها: برای اطمینان از مقیاس‌پذیری شاخص‌ها، داده‌ها با استفاده از روش Min–Max در بازه [۰،۱] نرمال شدند، نرمال‌سازی داده‌ها به روش Min–Max با هدف

یکسان‌سازی مقیاس متغیرها انجام شد، زیرا شاخص‌های تراکنشی دارای دامنه‌های عددی متفاوت بودند و بدون نرمال‌سازی انجام دسته بندی مبتنی بر RFM تحت تأثیر متغیرهایی با مقادیر بزرگ‌تر قرار می‌گرفتند:

$$X_{norm} = \frac{X - Min(x)}{\mathbf{Max}(x) - Min(x)}$$

نمایش توزیع داده‌ها: به منظور بررسی بهتر پراکندگی داده‌ها، از نمودار ویولونی استفاده شد که ترکیبی از نمودار چگالی و نمودار جعبه‌ای است. در شکل شماره ۲ نمودار ویولونی تعداد و مبلغ تراکنش‌ها به تفکیک ماه نشان داده شده است. مقدار شاخص چولگی برای توزیع مبالغ تراکنش‌ها برابر ۲,۶ محاسبه شد که بیانگر وجود چولگی شدید در سمت راست داده‌ها میباشد. در چنین شرایطی، بخش عمده‌ای از داده‌ها در مقادیر پایین متمرکز شده و تعداد محدودی از مشتریان دارای تراکنش‌های با مبلغ بسیار بالا هستند. برای کنترل اثر این چولگی بر نتایج تحلیل، در مرحله‌ی بعد از روش چندک‌دهی و حذف داده‌های پرت مبتنی بر دامنه‌ی میان‌چارکی استفاده شد تا توزیع داده‌ها متعادل‌تر گردد.



شکل ۲. نمودار ویولونی مبلغ و تعداد تراکنش کارت‌ها به تفکیک هر ماه

انتخاب متغیرهای رفتاری در این پژوهش بر مبنای چارچوب RFM و منطق داده‌محوری صورت گرفته است. در میان متغیرهای متعدد تراکنشی، چهار شاخص مجموع مبالغ واریز، مجموع مبالغ برداشت، تعداد تراکنش‌ها و زمان آخرین تراکنش به‌عنوان بازتاب‌دهنده‌ی مستقیم سه بُعد اصلی مدل RFM یعنی تازگی، فراوانی و ارزش مالی در نظر گرفته شده‌اند. این متغیرها به‌طور مستقیم از داده‌های تراکنشی استخراج می‌شوند و برخلاف متغیرهای دموگرافیک (نظیر سن یا جنسیت) در تمامی مشتریان در دسترس و قابل اندازه‌گیری هستند. علاوه بر این، آزمایش اولیه با متغیرهای مکمل مانند میانگین مانده حساب یا نسبت برداشت به واریز انجام شد، اما این متغیرها همبستگی بالایی با شاخص‌های RFM داشتند و اطلاعات جدید معناداری به مدل اضافه نمی‌کردند. بنابراین، تمرکز بر متغیرهای رفتاری منتخب موجب سادگی، تعمیم‌پذیری و پایداری بیشتر مدل در محیط‌های بانکی مختلف گردید.

۳-۳. محاسبه و استخراج مقادیر RFM

به‌منظور اعتبارسنجی مشتریان و سنجش وضعیت اعتباری آنان، سه شاخص رفتاری تازگی تراکنش، فراوانی تراکنش و مبلغ تراکنش مالی تراکنش برای هر مشتری محاسبه گردید. با توجه به سیاست پژوهش در تحلیل ماهانه رفتار مشتریان، این سه شاخص برای هر ماه به‌صورت مجزا استخراج شد.

۱. تازگی تراکنش: فاصله زمانی (بر حسب روز) از آخرین تراکنش مشتری تا پایان ماه محاسبه شد (حداکثر ۳۱ روز در هر ماه).

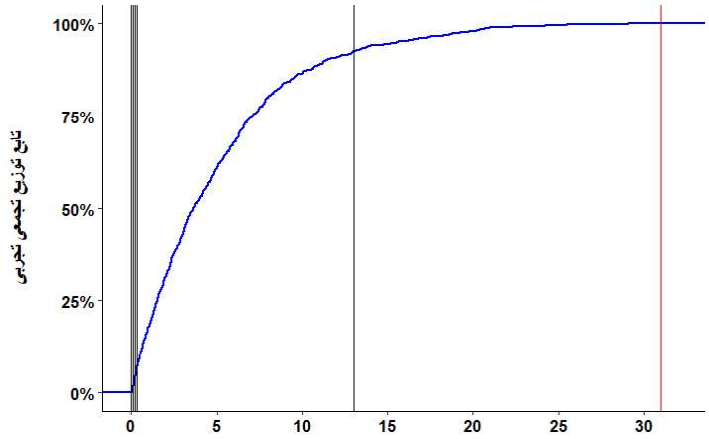
۲. فراوانی تراکنش: مجموع تعداد تراکنش‌های واریز و برداشت هر کارت در ماه به‌عنوان شاخص فعالیت مالی مشتری در نظر گرفته شد.

۳. مبلغ مالی تراکنش: مجموع مبالغ واریزی ماهانه هر مشتری محاسبه شد، زیرا این شاخص بیشترین ارتباط را با سطح اعتبار مشتری دارد.

در مجموع، برای هر مشتری طی هشت ماه، ۸ مقدار تازگی، ۸ مقدار فراوانی و ۸ مقدار مبلغ تراکنش محاسبه گردید. هر یک از این مقادیر بیانگر رفتار مشتری در یک ماه مجزا است و در مراحل بعدی، از میانگین یا روند این مقادیر برای تحلیل نهایی رفتار مالی و تعیین سطح اعتباری مشتری استفاده شد.

۳-۴. امتیازدهی مقادیر RFM

به‌منظور مقایسه‌پذیر کردن شاخص‌ها، مقادیر RFM در مقیاس ۱ تا ۵ امتیازدهی شدند. در روش متداول، این کار از طریق چندک‌های آماری انجام می‌شود. برای این منظور چندک‌های ٪۲۰، ٪۴۰، ٪۶۰، ٪۸۰ و ٪۱۰۰ محاسبه شده و به‌عنوان مرزهای امتیازگذاری در نظر گرفته شدند. در شکل شماره ۳ نمونه‌ای از فرایند امتیازدهی برروی مولفه تازگی تراکنش مشتریان با این روش نشان داده شده است.



شکل ۳. نمودار امتیازبندی مولفه تازگی تراکنش مشتریان در ماه اول با استفاده روش چندک‌دهی

همانطور که در شکل شماره ۳ نیز مشاهده می‌شود، با استفاده از روش چندک دهی، شش عدد مرز برای داده های حاصل از تازگی تراکنش مشتریان محاسبه شده است، به گونه‌ای که مرزهای اول تا ششم برپایه چندک‌های۰٪ تا ۱۰۰٪ محاسبه شده اند. بخوبی ملاحظه می شود که به دلیل یکسان بودن چندک‌های ۰٪ تا ۶۰٪، چهار مرز ابتدایی با یکدیگر برابر بوده و از این رو هیچ دسته‌ای دربین این مرزها تولید نمی شود. علت این امر، وجود چولگی سنگین در سمت راست داده‌ها و عدم توزیع متقارن آن‌ها است. برهمین اساس، روش چندک دهی به این شکل قابل استفاده نبوده و نیازمند تغییراتی است که درادامه بیان خواهد شد.

۳-۵. مدیریت داده‌های پرت

از آنجا که داده‌های پرت می‌توانند بر بازه‌های چندک اثرگذار باشند، ابتدا این داده‌ها شناسایی و موقتاً حذف شدند. معیار شناسایی داده‌های پرت بر اساس دامنه میان‌چارکی (IQR) تعریف شد:

$$\boldsymbol{IQR = Q_3 - Q_1}$$

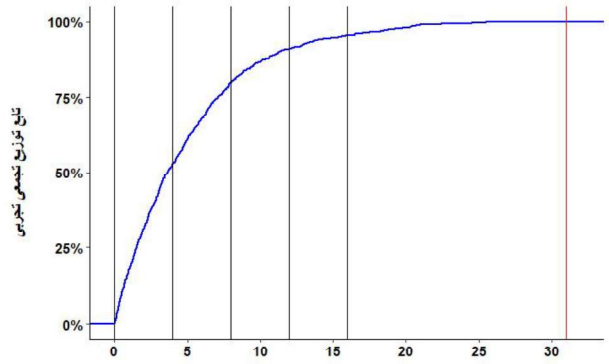
$$\boldsymbol{Outlier = \begin{cases} x < Q_1 - (1.5 \times IQR) \\ x > Q_3 + (1.5 \times IQR) \end{cases}}$$

۳-۶. بازه‌بندی نهایی

پس از حذف داده‌های پرت، دامنه مطلق داده‌های نرمال‌شده محاسبه و به چهار بازه مساوی تقسیم شد:

$$\boldsymbol{Range = Max\,(x_{norm}) - Min\,(x_{norm})}$$

سپس داده‌های پرت مجدداً به مجموعه اضافه شدند و در قالب دسته پنجم قرار گرفتند. این روش باعث شد امتیازدهی مشتریان به‌صورت عادلانه‌تر و متناسب با رفتار واقعی آن‌ها انجام شود. نمونه‌ای از فرایند امتیازبندی برای مولفه تازگی تراکنش مشتریان ماه اول در شکل شماره ۴ نمایش داده شده است.



شکل ۴. نمودار امتیازبندی تراکنش مشتریان در ماه اول با استفاده روش بهبود یافته

همانطور که در شکل شماره ۴ ملاحظه می‌شود، دسته‌بندی مشتریان و امتیازگذاری مقادیر تازگی تراکنش مشتریان در ماه اول به شیوه عادلانه‌تری صورت گرفته است. در این روش، ابتدا مقادیر پرت از داده‌ها کنار گذاشته شده و دامنه‌ی مطلق داده‌های باقی‌مانده به چهار بازه‌ی مساوی تقسیم گردید و کلیه‌ی مشتریانی که نسبت به سایر مشتریان رفتار غیرمعقول‌تری داشتند در یک دسته جداگانه قرار گرفتند.

۳-۷. اجرای خوشه‌بندی مشتریان

پس از محاسبه امتیازهای RFM، مرحله بعدی شناسایی الگوهای رفتاری مشتریان از طریق خوشه‌بندی انجام شد. با توجه به اینکه وضعیت اعتبار مشتریان از پیش مشخص نبود، از

روش‌های یادگیری بدون ناظر استفاده گردید. خوشه‌بندی به‌عنوان یکی از رایج‌ترین رویکردها در این زمینه شناخته می‌شود. برای نمونه، ریزکیانتو و همکاران از الگوریتم کا- میانگین برای خوشه‌بندی مشتریان بر اساس RFM استفاده کردند (Rizkyanto & Ford, 2023) و شیرول و همکاران نیز رویکرد تلفیقی RFM-خوشه‌بندی را در تحلیل اعتبار مشتریان یک شرکت تجارت الکترونیک به کار گرفتند (Shirole et al., 2021).

اگرچه الگوریتم کا- میانگین به دلیل سادگی و سرعت بالا کاربرد گسترده‌ای دارد، اما محدودیت‌هایی همچون وابستگی به انتخاب تعداد خوشه و شکل کروی داده‌ها موجب کاهش دقت آن می‌شود. به همین دلیل در این پژوهش از خوشه‌بندی مبتنی بر مدل استفاده گردید که بر اساس توابع احتمال آمیخته عمل میکند.

الگوریتم خوشه‌بندی مبتنی بر مدل نسبت به روش‌های کلاسیک مانند کا-میانگین مزایای مهمی دارد. نخست، در این رویکرد فرض می‌شود داده‌ها از ترکیب چند توزیع آماری مانند توزیع‌های گاوسی تولید شده‌اند، بنابراین خوشه‌ها می‌توانند اشکال، اندازه‌ها و جهت‌های متفاوتی داشته باشند (Grün, 2019)، در حالی که کا-میانگین خوشه‌ها را به‌صورت کره‌های هم‌گن با واریانس برابر فرض می‌کند (BHARDWAJ & RAO, 2025; Xiao, 2024). دوم، این روش امکان تخصیص نرم[1] را فراهم می‌سازد؛ هر نقطه با احتمال مشخصی به خوشه‌ها تعلق دارد، که در داده‌های دارای هم‌پوشانی یا مرزهای مبهم بسیار مفید است (Gormley et al., 2023). در مقابل، کا-میانگین تنها از تخصیص سخت استفاده می‌کند و هر نقطه را دقیقاً در یک خوشه قرار می‌دهد. سوم، در خوشه‌بندی مبتنی بر مدل با استفاده از معیار اطلاع بیزی می‌توان تعداد بهینه خوشه‌ها و مدل مناسب را تعیین کرد، در حالی که کا-میانگین نیازمند تعیین از پیش تعداد خوشه بوده و ابزار آماری مشخصی برای آن ندارد (Gormley et al., 2023; Kessler, 2019).

[1] Soft Assignment

جزئیات بکارگیری این الگوریتم کا میانگین و در نهایت ترجیح الگوریتم مبتنی بر مدل در بخش ضمیمه بیان شده است.

در ادامه سه گام متوالی با هدف بررسی پایداری ساختار خوشه‌ها در بازه‌های زمانی متفاوت طراحی گردید. این رویکرد، با الهام از روش تحلیل پنجره‌ای[۱] در مطالعات داده‌های زمانی پیشنهاد شده است که امکان ارزیابی ثبات خوشه‌ها را در طول زمان فراهم می‌سازد. به همین دلیل، بازه‌های زمانی با یک ماه جابه‌جایی انتخاب شدند (۱ تا ۶، ۲ تا ۷ و ۳ تا ۸ ماه). در انتخاب ساختار مدل آماری، از چارچوب خانواده‌ی توزیع‌های گاوسی آمیخته[۲] استفاده گردید. در این میان، مدل[۳] EEE به‌عنوان ساختاری با فرض برابری واریانس‌ها و کوواریانس‌ها در میان خوشه‌ها شناخته می‌شود که بر اساس معیار اطلاع بیزی بهترین برازش را با داده‌های پژوهش داشت.

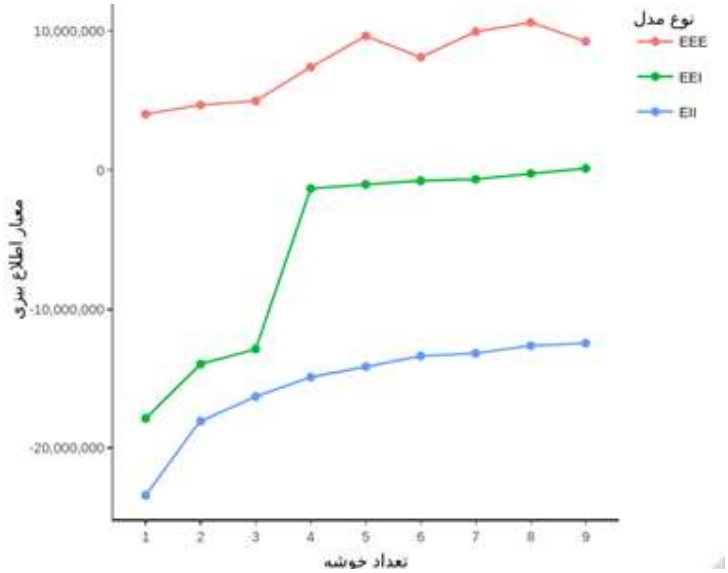
در گام نخست، داده‌های RFM در بازه زمانی یک تا شش ماهه مورد تحلیل قرار گرفتند. نتایج معیار اطلاع بیزی برای مدل‌های مختلف در شکل شماره ۵ نشان داده شده است. مطابق نتایج، مدل EEE با پنج خوشه بالاترین مقدار اطلاع بیزی را داشته و به‌عنوان مدل بهینه انتخاب گردید.

---

۱ Rolling Window Analysis

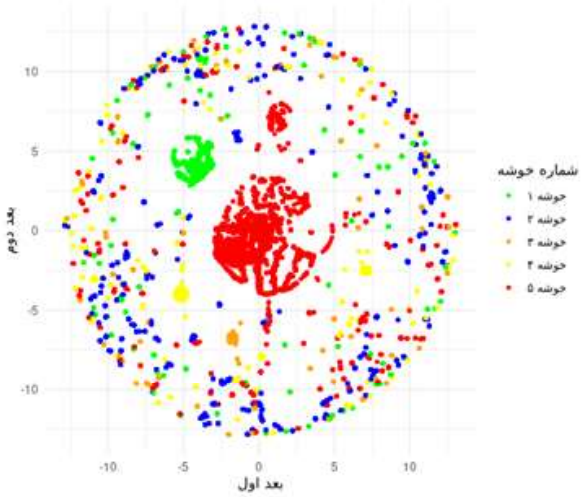
۲ Gaussian Mixture Models

3. Ellipsoidal ،equal volume, equal shape, equal orientation.



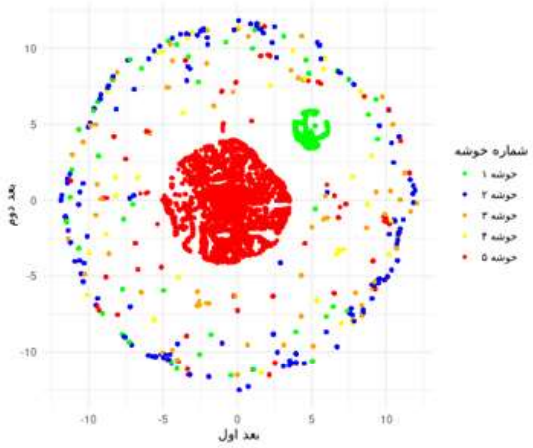
شکل ۵. نمودار ارزیابی بهترین مدل خوشه‌بندی و تعداد خوشه بهینه بر حسب معیار اطلاع بیزی در بازه‌ی زمانی ۱ تا ۶ ماه

توزیع فراوانی مشتریان در هر خوشه بررسی گردید؛ خوشه شماره ۵ با ۴۶,۷۲٪ بیشترین سهم را داشته و خوشه شماره ۳ با ۱۰,۱۹٪ کمترین سهم را به خود اختصاص داده است. برای نمایش گرافیکی پراکنش خوشه‌ها، از روش کاهش ابعاد t-SNE استفاده شد که نتایج آن در شکل شماره ۶ ارائه شده است. بر اساس این شکل، خوشه‌های شماره ۱ و ۵ تفکیک‌پذیری بالاتری نسبت به سایر خوشه‌ها داشته‌اند.



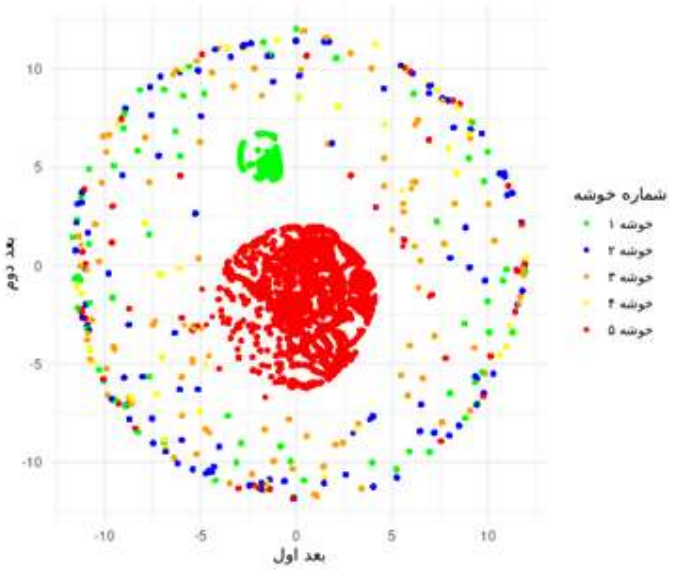
شکل ۶. نمودار پراکنش خوشه‌های مختلف برحسب t-SNE در بازه‌ی زمانی ۱ تا ۶ ماه

در گام دوم، داده‌های RFM در بازه زمانی دو تا هفت ماهه بررسی شدند. نتایج پراکنش آن‌ها بر اساس t-SNE در شکل شماره ۷ آمده است. در این بازه نیز خوشه شماره ۵ بیشترین جمعیت مشتریان (۵۴٫۹۷٪) را شامل شد و خوشه شماره ۲ کمترین سهم (۵٫۸۸٪) را داشت.



شکل ۷. نمودار پراکنش خوشه‌های مختلف برحسب t-SNE در بازه‌ی زمانی ۲ تا ۷ ماه

در گام سوم، داده‌های مربوط به بازه زمانی سه تا هشت ماهه تحلیل شدند. نتایج پراکنش آن‌ها در شکل شماره ۸ ارائه شده است. مشابه نتایج قبلی، خوشه شماره ۵ با سهم ۴۹,۹٪ بزرگ‌ترین خوشه و خوشه شماره ۲ با ۱۱,۲۳٪ کوچک‌ترین خوشه بودند. به‌طور کلی، اجرای خوشه‌بندی مبتنی بر مدل نشان داد که مشتریان در پنج گروه رفتاری متمایز قرار می‌گیرند و بخش عمده‌ای از آن‌ها در خوشه شماره ۵ متمرکز هستند. نتایج همچنین تأیید می‌کند که این روش قادر به تفکیک مناسب مشتریان براساس شاخص‌های RFM در بازه‌های زمانی مختلف بوده است.



شکل ۸. نمودار پراکنش خوشه‌های مختلف برحسب t-SNE در بازه‌ی زمانی ۳ تا ۸ ماه

۳-۸. تعیین ارزش طول عمر مشتری

به‌منظور سنجش میزان ارزشمندی مشتریان در هر خوشه و ارزیابی سطح اعتباری آن‌ها، شاخص ارزش طول عمر مشتری محاسبه گردید. این شاخص نشان‌دهنده ارزش فعلی تعامل مشتری با سازمان در طول یک دوره مشخص است و می‌تواند معیار دقیقی برای تصمیم‌گیری‌های اعتباری و بازاریابی باشد. در نسخه به‌روزشده این مدل، برای غنی‌تر کردن محاسبه‌ی ارزش طول عمر مشتری و افزایش دقت، از میانگین هندسی وزنی نرمال‌شده به‌جای ترکیب خطی ساده شاخص‌های RFM استفاده شد. این رویکرد امکان در نظر گرفتن اثرات هم‌زمان و تعاملی میان شاخص‌های تازگی، فراوانی و مبلغ تراکنش را فراهم می‌کند و از نظر آماری پایداری بیشتری نسبت به روش‌های خطی دارد.

فرمول نهایی محاسبه‌ی ارزش طول عمر مشتری با توجه به تجربه و نظر خبرگان به صورت زیر تعریف می‌شود:

$$\boldsymbol{CLV} = \left(\boldsymbol{R}^{(W_{n_R})} \times \boldsymbol{F}^{(W_{n_F})} \times \boldsymbol{M}^{(W_{n_M})}\right)^{\left(\frac{1}{(W_{N_R}+W_{N_F}+W_{N_M})}\right)}$$

در این معادله:

$\boldsymbol{W_{n_R}}$ وزن نرمال شاخص تازگی تراکنش،

$\boldsymbol{W_{n_F}}$ وزن نرمال شاخص فراوانی تراکنش،

$\boldsymbol{W_{n_M}}$ وزن نرمال شاخص مبلغ تراکنش مالی،

$\boldsymbol{R, F, M}$ به ترتیب مقادیر سه شاخص RFM برای هر مشتری هستند.

از آنجا که اهمیت نسبی این شاخص‌ها ممکن است در هر کسب‌وکار متفاوت باشد، تعیین وزن‌ها با بهره‌گیری از فرآیند تحلیل سلسله‌مراتبی[۱] بر اساس نظر خبرگان حوزه تحلیل مشتریان در مؤسسه مالی و اعتباری ملل انجام شد. وزن‌های به‌دست‌آمده در جدول شماره ۳ ارائه شده‌اند.

جدول۳. وزن‌های تعیین شده برای RFM

وزن	شاخص تازگی تراکنش	شاخص فراوانی تراکنش	شاخص مبلغ تراکنش مالی
وزن نرمال شده	۰,۲۱۷	۰,۳۴۸	۰,۴۳۵

بر اساس این وزن‌ها، شاخص ارزش طول عمر مشتری برای هر مشتری و در هر ماه محاسبه شد. این محاسبات مبنای تحلیل خوشه‌ها و مقایسه سطح اعتباری آن‌ها قرار گرفت.

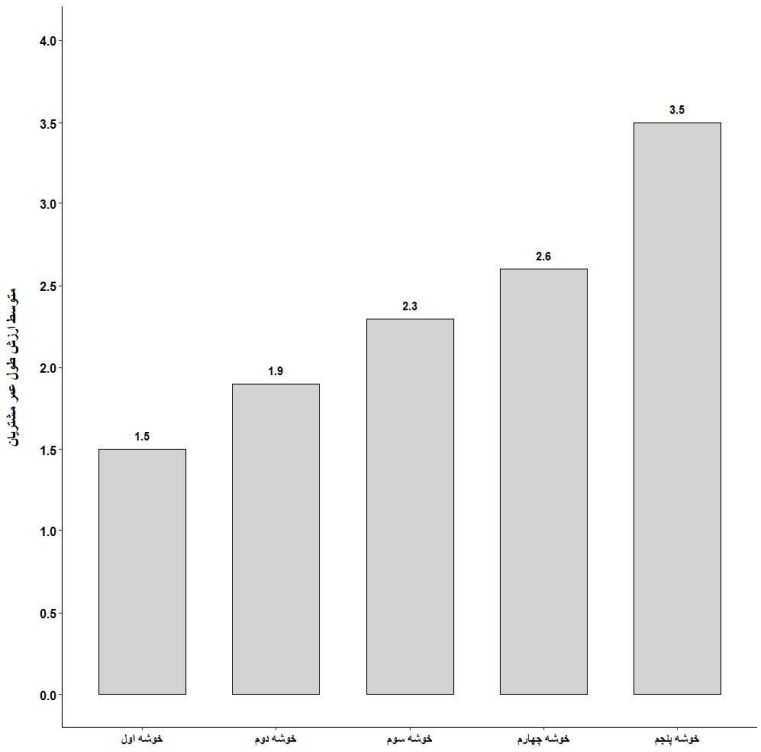
۳-۹. تفسیر نتایج خوشه‌ها پس از تعیین ارزش طول عمر مشتری

پس از محاسبه شاخص ارزش طول عمر مشتری برای مشتریان، میانگین این شاخص در سه بازه زمانی مختلف (۱ تا ۶ ماه، ۲ تا ۷ ماه و ۳ تا ۸ ماه) محاسبه شد تا تصویری کلی از وضعیت اعتباری هر خوشه به دست آید. نتایج در قالب نمودارهای ستونی (شکل‌های شماره ۹ تا ۱۱) ارائه شده‌اند. نتایج نشان داد که تطبیق شاخص معیار اطلاع بیزی و مقادیر ارزش طول عمر مشتری در هر خوشه، هم‌خوانی معناداری دارد و این موضوع نشان‌دهنده دقت بالای مدل در تفکیک رفتارهای اعتباری است. بازه ۱ تا ۶ ماه (شکل شماره ۹): خوشه پنجم

[۱] AHP

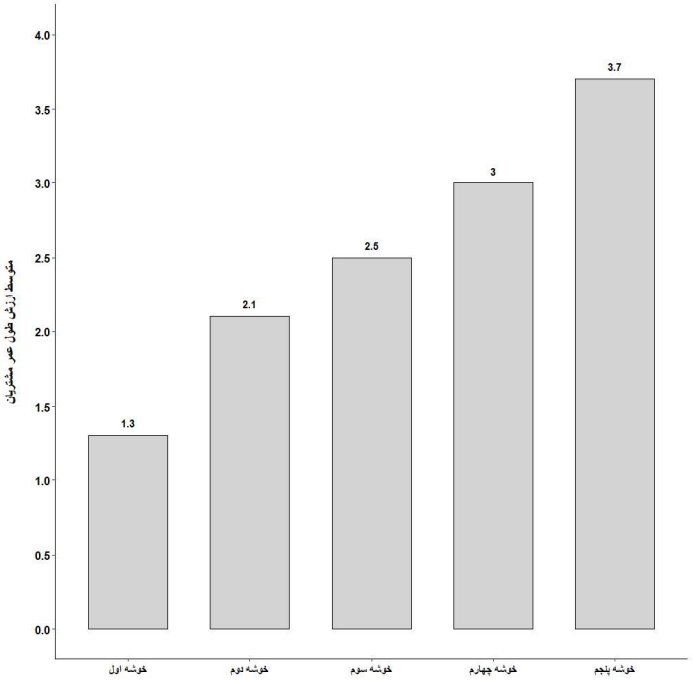
با میانگین ارزش طول عمر مشتری برابر ۳,۵ بالاترین ارزش را داشته و به‌عنوان مشتریان پلاتینیومی شناخته شد. پس از آن خوشه چهارم با میانگین ۲,۶ در رتبه دوم قرار گرفت و به‌عنوان مشتریان طلایی معرفی گردید. خوشه سوم (۲,۳) و خوشه دوم (۱,۹) به ترتیب در گروه‌های نقره‌ای و برنزی قرار گرفتند و در نهایت خوشه اول با میانگین ۱,۵ کمترین ارزش را داشت و به‌عنوان مشتریان مسی شناسایی شد. تحلیل نتایج نشان می‌دهد که مشتریان خوشه‌های پلاتینیومی و طلایی نه‌تنها بالاترین مقدار شاخص ارزش طول عمر مشتری را دارند، بلکه از نظر میانگین مبلغ تراکنش و مانده سپرده نیز در سطوح بالاتری قرار دارند.

این هم‌راستایی میان ارزش طول عمر مشتری و سطح فعالیت مالی، بیانگر آن است که شاخص ارزش طول عمر مشتری می‌تواند معیار مناسبی برای شناسایی مشتریان کلیدی بانک باشد. از منظر مدیریتی، توصیه می‌شود بانک در سیاست‌های نگهداشت مشتری، تمرکز بیشتری بر خوشه‌های پلاتینیومی و طلایی داشته و از طریق پیشنهاد تسهیلات ترجیحی و طرح‌های وفاداری، مانع از ریزش آن‌ها شود. در مقابل، خوشه‌های نقره‌ای و برنزی نیازمند طراحی برنامه‌های فعال‌سازی و افزایش تعامل هستند. همچنین توزیع سپرده‌ها و تراکنش‌ها در میان خوشه‌ها می‌تواند مبنایی برای تنظیم سقف اعتباری و مدیریت پرتفوی تسهیلاتی بانک باشد.



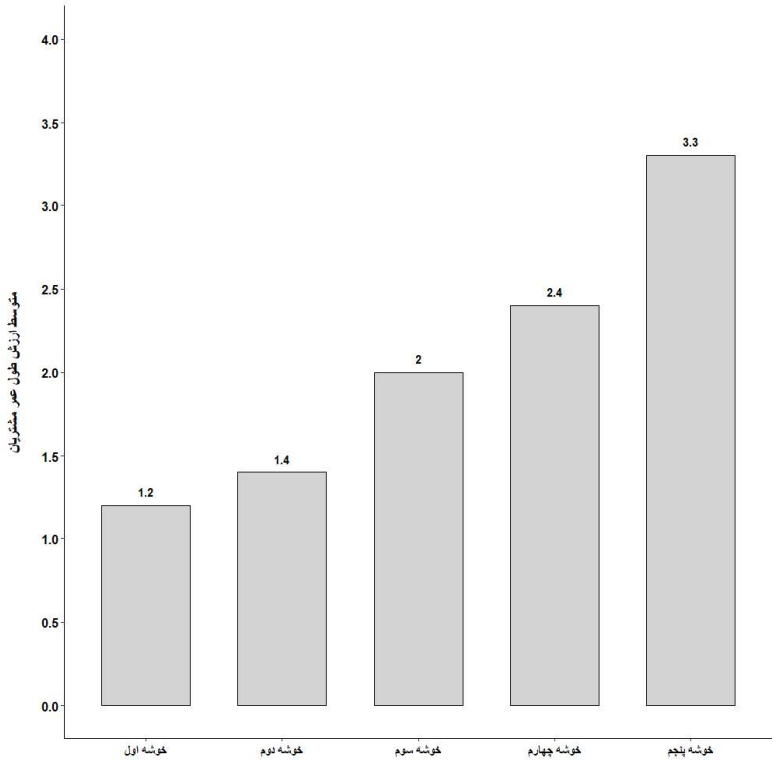
شکل ۹. نمودار ستونی میانگین شاخص ارزش طول عمر مشتریان در هریک از خوشه‌های ۱ تا ۶ ماه

بازه ۲ تا ۷ ماه (شکل شماره ۱۰): مشابه بازه قبل، خوشه پنجم (۳,۷) در جایگاه نخست باقی ماند. خوشه چهارم با میانگین ۳ و خوشه سوم با میانگین ۲,۵ عملکردی نزدیک به هم داشتند و به‌ترتیب به‌عنوان طلایی و نقره‌ای معرفی شدند. خوشه دوم با میانگین ۲,۱ در سطح برنزی و خوشه اول با ۱,۳ در سطح مسی قرار گرفت.



شکل ۱۰. نمودار ستونی میانگین شاخص ارزش طول عمر مشتریان در هریک از خوشه‌های ۲ تا ۷ ماه

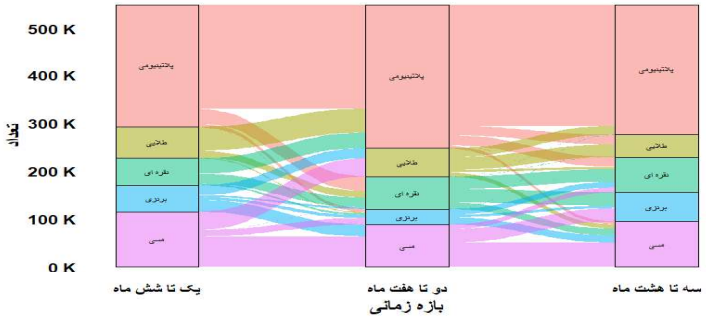
بازه ۳ تا ۸ ماه (شکل شماره ۱۱): خوشه پنجم همچنان بالاترین میانگین (۳,۳) را داشت و در گروه پلاتینیومی باقی ماند. خوشه چهارم با ۲,۴ در سطح طلایی، خوشه سوم با ۲ در سطح نقره‌ای و خوشه دوم با ۱,۴ در سطح برنزی قرار گرفتند. خوشه اول نیز با ۱,۲ کمترین ارزش را داشت و به‌عنوان مسی معرفی شد.



شکل۱۱. نمودار ستونی میانگین شاخص ارزش طول عمر مشتریان در هریک از خوشه‌های ۳ تا ۸ ماه

به‌منظور بررسی تغییرات سطح اعتباری مشتریان در طول زمان، روند سه بازه زمانی فوق در شکل شماره ۱۲ نمایش داده شده است. در این شکل، نمودار جریان جابه‌جایی مشتریان میان سطوح مختلف اعتباری در سه بازه‌ی زمانی تحلیلی نمایش داده شده است. هر ستون بیانگر توزیع مشتریان در یک بازه‌ی شش‌ماهه و هر نوار اتصال‌دهنده نشان‌دهنده انتقال مشتریان از یک خوشه به خوشه‌ی دیگر در بازه‌ی بعدی است. پهنای هر نوار متناسب با تعداد مشتریان

منتقل‌شده است. بنابراین، این نمودار روند پویایی خوشه‌های اعتباری را در بازه‌های متوالی نشان می‌دهد و امکان مشاهده‌ی تغییرات سطح وفاداری و رفتار اعتباری مشتریان را فراهم می‌سازد. نتایج نشان می‌دهد که اگرچه بخش عمده‌ای از مشتریان در گروه پلاتینیومی قرار دارند، بخشی از آن‌ها در بازه‌های زمانی بعدی به سطح نقره‌ای یا برنزی تنزل یافته‌اند. این تغییر جایگاه خوشه‌ای بیانگر کاهش نسبی ارزش اعتباری برخی مشتریان در مقایسه با سایرین است، نه ریزش واقعی در طول زمان. در مقابل، برخی مشتریان با ارزش پایین‌تر نیز ثبات نسبی در جایگاه خود نشان داده‌اند. به‌طور کلی می‌توان نتیجه گرفت که خوشه پنجم پایدارترین و ارزشمندترین گروه مشتریان است، در حالی که خوشه اول همواره ضعیف‌ترین عملکرد اعتباری را داشته است. تحلیل روند تغییرات سطح اعتباری می‌تواند مبنای طراحی سیاست‌های مدیریت وفاداری مبتنی بر داده قرار گیرد. علاوه بر این، بررسی همبستگی میان شاخص ارزش طول عمر مشتری و سطح پایداری خوشه‌ها نشان داد که مقدار ارزش طول عمر مشتری با ثبات رفتاری رابطه‌ای مستقیم دارد. خوشه‌های پلاتینیومی و طلایی نه‌تنها دارای بالاترین مقادیر ارزش طول عمر مشتری بودند، بلکه احتمال انتقال آنان به خوشه‌های پایین‌تر در مدل پیش‌بینی آینده نیز کمتر بود. این یافته نشان می‌دهد ارزش اقتصادی مشتریان با ثبات اعتباری آن‌ها هم‌راستا است.



شکل۱۲. نمودار روند تغییر سطح اعتباری مشتریان در سه بازه‌ی زمانی مختلف

۱۰,۳ توسعه مدل پیش‌بینی سطح اعتباری آینده مشتری

یکی از اهداف کلیدی این پژوهش، پیش‌بینی تغییرات سطح اعتباری مشتریان در آینده بر اساس تاریخچه رفتار مالی آن‌ها است. اهمیت این موضوع از آنجا ناشی می‌شود که شناسایی به‌موقع تغییرات در سطح اعتبار مشتریان می‌تواند به بانک‌ها و نهادهای مالی در تصمیم‌گیری درباره اعطای تسهیلات، مدیریت ریسک بازپرداخت و طراحی سیاست‌های وفاداری کمک کند. برای این منظور، داده‌های تراکنشی شش‌ماهه به‌عنوان ورودی مدل مورد استفاده قرار گرفتند و شاخص‌های RFM ماهانه به‌عنوان ویژگی‌های اصلی وارد الگوریتم شدند. عملکرد مدل در دو حالت مختلف مورد ارزیابی قرار گرفت و در نهایت، نتایج هر دو حالت جمع‌بندی و مقایسه شد تا عملکرد مدل در شرایط مختلف ارزیابی شود. این تحلیل امکان بررسی پایداری و تعمیم‌پذیری مدل پیش‌بینی سطح اعتباری را فراهم می‌سازد.

۱۱,۳ حالت ۱: پیش‌بینی ماه آینده بر اساس بازه شش‌ماهه گذشته (تخصیص کل داده‌های گذشته تا ماه فعلی به مجموعه آموزش)

در این حالت، داده‌های مربوط به شش ماه نخست برای آموزش مدل به کار گرفته شدند و سطح اعتباری مشتری در ماه هفتم به‌عنوان برچسب هدف مورد ارزیابی قرار گرفت. برای آزمون پایداری مدل، فرآیند مشابهی بر روی بازه ماه دوم تا ماه هفتم انجام شد و سطح اعتباری مشتریان در ماه هشتم پیش‌بینی گردید. در این حالت، فرآیند خوشه‌بندی پیش از آموزش مدل انجام شد و نتایج خوشه‌بندی به‌عنوان متغیرهای ورودی در مدل جنگل تصادفی به کار گرفته شد. بدین ترتیب، مدل یادگیری ماشین از ساختار خوشه‌ای استخراج‌شده در مرحله‌ی بدون ناظر برای بهبود دقت پیش‌بینی استفاده می‌کند. داده‌ها پس از مرحله‌ی خوشه‌بندی، به دو بخش آموزشی (به عنوان مثال شش ماه نخست) و آزمون (به عنوان مثال

ماه هفتم) تقسیم شدند تا اعتبار مدل در داده‌های خارج از نمونه سنجیده شود. مدل طراحی‌شده در این حالت به دلیل استفاده از داده‌های خارج از نمونه در مرحله آزمون، اهمیت بیشتری نسبت به سایر حالت‌ها دارد و کارآمدی آن در کاربردهای واقعی نظام بانکی بیشتر است.

نتایج کمی مدل

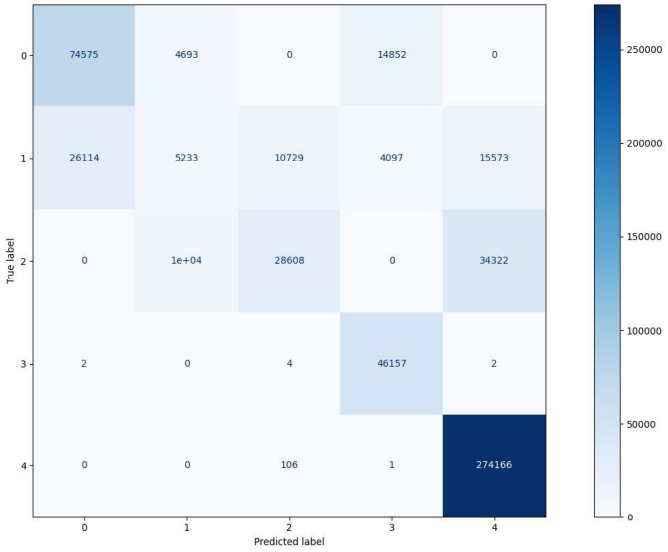
عملکرد مدل با سه شاخص دقت، بازیابی و میانگین متوازن ارزیابی شد. همچنین میانگین کل دقت مدل در سطح ۷۲ درصد و میانگین میانگین متوازن دقت و بازخوانی برابر با ۷۱ درصد برآورد گردید که بیانگر توان قابل‌قبول مدل در شناسایی الگوهای اعتباری است. بررسی جزئی‌تر جدول شماره ۴ نشان می‌دهد که مدل در پیش‌بینی خوشه‌های بزرگ‌تر (به‌ویژه خوشه ۵) عملکرد بهتری داشته است، زیرا حجم بالای داده‌ها در این گروه موجب یادگیری پایدارتر مدل شده است. در مقابل، کاهش مقادیر بازیابی در خوشه‌های ۲ و ۳ نشان می‌دهد که این گروه‌ها دارای رفتارهای پراکنده و ناهمگن‌تری بوده‌اند و الگوریتم در شناسایی دقیق اعضای آن‌ها با چالش مواجه شده است، علت این موضوع عدم وجود الگوی رفتاری منظم در این خوشه‌ها است؛ مسئله‌ای که در نتایج خوشه‌بندی پیشین نیز مشاهده شد. با این حال، میانگین متوازن دقت و بازیابی در سطح ۷۰ درصد به بالا بیانگر پایداری کلی مدل و امکان اتکای مدیریتی آن در محیط‌های واقعی است. از منظر کاربردی، تمرکز تصمیم‌گیران بر خوشه‌های دارای میانگین متوازن بالاتر (۴ و ۵) می‌تواند منجر به تخصیص بهینه‌تر منابع اعتباری شود.

جدول ۴. خروجی جنگل تصادفی برای بازه شش‌ماهه در پیش‌بینی ماه آینده

خوشه	precision	Recall	F1-score	support
۱	۰,۶۵	۰,۸۶	۰,۷۴	۹۴۱۲۰
۲	۰,۲۶	۰,۰۸	۰,۱۳	۶۱۷۴۶
۳	۰,۵۶	۰,۴۶	۰,۵۰	۷۳۳۰۵
۴	۰,۷۳	۰,۶۵	۰,۶۸	۴۶۱۶۵
۵	۰,۸۶	۰,۹۵	۰,۹۱	۲۷۴۲۷۳
میانگین وزنی	۰,۷۲	۰,۷۵	۰,۹۱	

تحلیل کیفی مدل

در شکل شماره ۱۳، ماتریس درهم‌ریختگی حاصل از اجرای الگوریتم نمایش داده شده است. نتایج نشان می‌دهد که مقادیر پیش‌بینی‌شده برای خوشه‌های شماره ۲ و ۳ بیشترین پراکندگی را دارند، در حالی‌که خوشه شماره ۵ به‌عنوان بزرگ‌ترین خوشه و خوشه شماره ۴ به‌عنوان کوچک‌ترین خوشه تفکیک‌پذیری بهتری داشته‌اند.

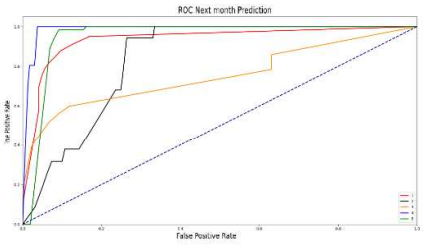


شکل ۱۳. ماتریس در هم ریختگی در پیش بینی ماه آینده

همچنین در شکل شماره ۱۴ مقادیر مساحت زیر منحنی[۱] منحی راک[۲] برای هر کلاس ترسیم شده است. در این نمودار، محور y بیانگر حساسیت[۳] و محور x بیانگر نرخ خطای نوع اول[۴] است. نتایج نشان می‌دهد که کلاس‌های شماره ۲ و ۳ دارای مساحت زیر منحنی کمتری نسبت به سایر کلاس‌ها هستند و بنابراین تفکیک‌پذیری آن‌ها کمتر است. در این پژوهش برای محاسبه مساحت‌ها از روش یک در مقابل همه[۵] استفاده شد که در آن هر کلاس با مجموع سایر کلاس‌ها مقایسه می‌شود.

---

۱ Area Under the ROC (AUC)
۲ Roc curve
۳ True Positive Rate (TPR)
۴ False Positive Rate (FPR)
۵ One-vs-Rest (OVR)

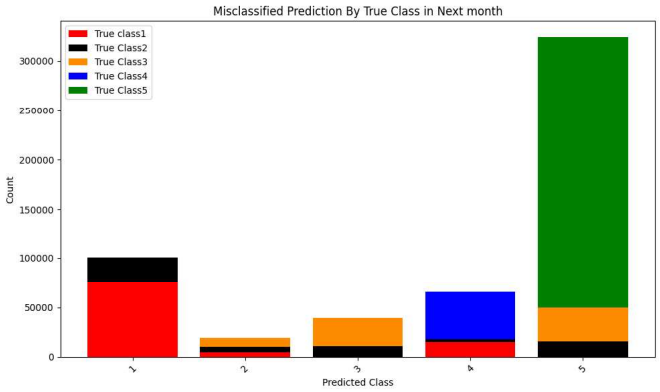


کلاس	AUC
۱	۰,۹۳
۲	۰,۸۳
۳	۰,۷۵
۴	۰,۹۸
۵	۰,۹۵

شکل ۱۴: مقادیر AUC و نمودار ROC برای هر کلاس در پیش بینی ماه آینده

تحلیل خطاها

در شکل شماره ۱۵، نمودار میله‌ای انباشته نتایج پیش‌بینی برای هر خوشه نشان داده شده است. مشاهده می‌شود که برای کلاس‌های شماره ۲ و ۳ بخش بزرگی از مشاهدات به اشتباه در سایر کلاس‌ها قرار گرفته‌اند. به‌طور مثال، بخش‌هایی از کلاس شماره ۳ به اشتباه در کلاس شماره ۲ و بخشی از کلاس شماره ۱ در کلاس شماره ۲ پیش‌بینی شده‌اند. این وضعیت ناشی از عدم کشف ارتباط منطقی پایدار توسط درخت‌های تصمیم در الگوریتم جنگل تصادفی برای این خوشه‌ها است. در مقابل، سایر خوشه‌ها به‌ویژه خوشه پنجم، عملکرد بهتری داشته‌اند. به‌طور کلی، نتایج حالت اول نشان می‌دهد که مدل جنگل تصادفی در پیش‌بینی سطح اعتباری مشتریان عملکرد مناسبی داشته، اما در مورد خوشه‌های با رفتار نامنظم (۲ و ۳) دقت پایین‌تری ارائه داده است. این مسئله نشان‌دهنده ضرورت استفاده از رویکردهای تکمیلی برای این دسته از مشتریان است.



شکل ۱۵. نمودارمیله ای انباشته در بازه زمانی ۶ ماهه با روش پیش بینی ماه آینده

۱۲,۳ حالت ۲: پیش‌بینی ماه فعلی برحسب بازه شش ماهه گذشته (تقسیم داده‌های گذشته تا ماه فعلی به مجموعه آموزش و آزمون)

در این حالت، داده‌های شش‌ماهه به دو بخش تقسیم شدند: ۸۰ درصد برای آموزش مدل و ۲۰ درصد برای آزمون مدل. نتایج این حالت در جدول شماره ۵ ارائه شده است. مقایسه با حالت اول نشان می‌دهد که اگرچه دقت مدل در این حالت بالاتر است، اما از منظر کاربرد عملی، حالت اول همچنان برای پیش‌بینی وضعیت واقعی آینده مشتریان مناسب‌تر به نظر می‌رسد. به‌منظور بررسی پایداری مدل، تحلیل حساسیت نسبت به تغییر پارامترهای کلیدی شامل تعداد درخت‌ها و عمق هر درخت انجام شد. نتایج نشان داد افزایش تعداد درخت‌ها از ۱۰۰ به ۳۰۰ تغییر محسوسی در شاخص‌های دقت و میانگین متوازن دقت و بازخوانی ایجاد نکرد، در حالی‌که عمق زیاد منجر به بیش‌برازش و کاهش عملکرد بر روی داده‌های آزمون گردید. بر این اساس، مقدار بهینه برای مدل، ۲۰۰ درخت با عمق حداکثر ۱۰ تعیین شد.

نتایج کمی مدل

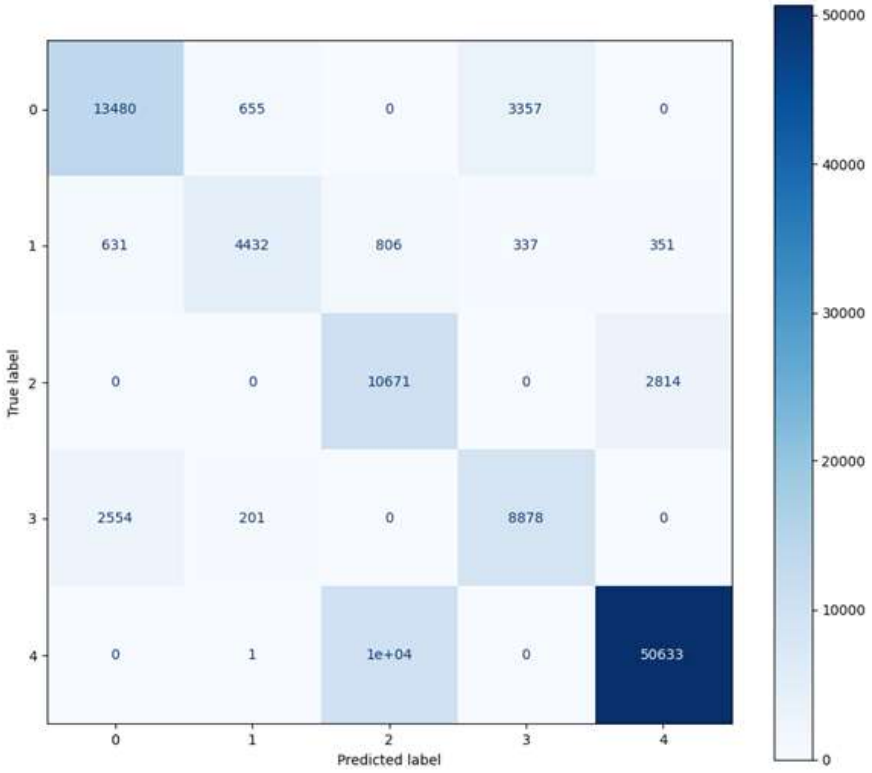
بر اساس جدول شماره ۵، شاخص‌های دقت، بازیابی و میانگین متوازن دقت و بازخوانی نسبت به حالت اول بهبود یافته‌اند. این موضوع نشان می‌دهد که تقسیم درون‌نمونه‌ای به مدل کمک کرده تا الگوهای رفتاری را بهتر شناسایی کند.

جدول ۵. خروجی جنگل تصادفی بصورت ۸۰-۲۰

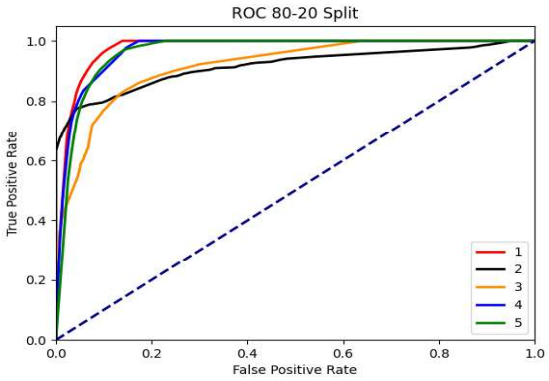
خوشه	precision	Recall	F1-score	support
۱	۰,۸۱	۰,۷۷	۰,۷۹	۱۷۴۹۲
۲	۰,۸۴	۰,۶۸	۰,۷۵	۶۵۵۷
۳	۰,۶۱	۰,۷۹	۰,۶۱	۱۳۴۸۵
۴	۰,۷۱	۰,۷۶	۰,۷۳	۱۱۶۳۳
۵	۰,۹۴	۰,۸۳	۰,۸۸	۶۰۷۵۵
میانگین وزنی	۰,۸۳	۰,۸۰	۰,۸۱	

تحلیل کیفی مدل

در شکل شماره ۱۶، ماتریس درهم‌ریختگی برای این حالت نمایش داده شده است. مقایسه با شکل ۱۳ نشان می‌دهد که در این حالت تعداد بیشتری از پیش‌بینی‌ها بر روی قطر اصلی قرار گرفته‌اند، که بیانگر تفکیک دقیق‌تر خوشه‌ها است. همچنین در شکل شماره ۱۷ نتایج مربوط به مساحت زیر منحنی و نمودار منحنی راک برای هر کلاس نشان داده شده است. مقادیر بالاتر مساحت زیر منحنی نسبت به حالت اول، گویای تفکیک‌پذیری بهتر میان TPR و FPR است.



شکل ۱۶. ماتریس در هم ریختگی بصورت ۸۰-۲۰

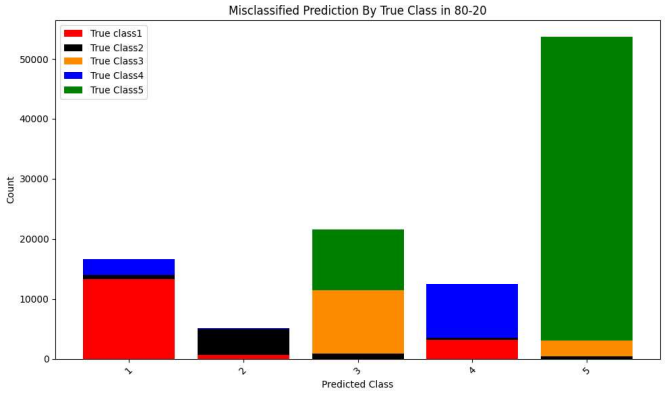


کلاس	AUC
۱	۰,۹۷
۲	۰,۹۱
۳	۰,۹۱
۴	۰,۹۶
۵	۰,۹۶

شکل ۱۷ . نتایج ROC و AUC برای هر کلاس بصورت ۸۰–۲۰

تحلیل خطاها

نمودار میله‌ای انباشته در شکل شماره ۱۸ وضعیت پیش‌بینی کلاس‌ها را نشان می‌دهد. مشاهده می‌شود که در کلاس شماره ۳ بخشی از داده‌ها به اشتباه در کلاس شماره ۵ قرار گرفته‌اند، موضوعی که در ماتریس درهم‌ریختگی نیز قابل مشاهده است. با این حال، در مقایسه باحالت اول، میزان خطا کاهش یافته و توزیع پیش‌بینی‌ها منظم‌تر شده است.



شکل ۱۸. نمودارمیله ای انباشته در بازه زمانی ۶ ماهه بصورت ۸۰-۲۰

به‌طور کلی، نتایج حالت دوم نشان می‌دهد که مدل جنگل تصادفی در تقسیم داده‌های درون‌نمونه‌ای عملکرد بهتری از خود نشان می‌دهد. با این حال، از آنجا که این روش پیش‌بینی وضعیت آینده واقعی مشتریان را منعکس نمی‌کند، کاربرد آن در سناریوهای عملی محدودتر است. بنابراین، حالت اول با وجود دقت کمتر، برای تصمیم‌گیری‌های اعتباری بانک‌ها مناسب‌تر خواهد بود.

۴. جمع‌بندی و نتیجه‌گیری

تعیین سطح اعتباری مشتری، به بانک کمک می‌کند تا ریسک اعطای اعتبار به مشتریان را کاهش دهد و در عین حال، خدمات مالی مناسب‌تری را به آن‌ها ارائه دهد. این امر موجب می‌شود بانک بتواند مشتریانی را که احتمال بازپرداخت بدهی‌های خود را دارند، شناسایی کند. این شناسایی همچنین منجر به کاهش ریسک اعطای اعتبار به مشتریان بدحساب شده و به بانک این امکان را می‌دهد تا خدمات مالی متناسب‌تری ارائه دهد. به عنوان مثال، بانک می‌تواند به مشتریانی که سطح اعتباری بالاتری دارند، تسهیلات بیشتری با نرخ سود پایین‌تر اعطا نماید. استفاده از علم داده، به بانک‌ها این امکان را می‌دهد تا فرآیند تعیین سطح اعتباری

مشتریان خود را با دقت و اطمینان بیشتری انجام دهند. در این مطالعه، با استفاده از داده‌های واقعی تراکنش‌های مالی مشتریان مؤسسه اعتباری ملل، پس از انجام مراحل پیش‌پردازش، شاخص‌های RFM برای هر مشتری محاسبه و سپس با استفاده از روش خوشه‌بندی مبتنی بر مدل، سطوح اعتباری مشتریان تعیین گردید. در ادامه، با به‌کارگیری شاخص ارزش طول عمر مشتری ، هر خوشه از نظر ارزش اقتصادی مورد ارزیابی قرار گرفت و نشان داده شد که خوشه‌های دارای ارزش طول عمر مشتری بالاتر، رفتار اعتباری باثبات‌تر و ریسک نکول پایین‌تری دارند.

در بخش پیش‌بینی، با استفاده از داده‌های RFM شش‌ماهه، سطح اعتباری فعلی مشتریان مدل‌سازی شد و سپس با به‌کارگیری الگوریتم جنگل تصادفی، سطح اعتباری مشتریان در ماه آینده پیش بینی گردید. این تحلیل به بانک کمک می‌کند تا مشتریان در معرض کاهش سطح اعتبار را زودتر شناسایی کرده و سیاست‌های پیشگیرانه‌ای مانند بازنگری در سقف تسهیلات یا ارائه بسته‌های وفاداری را اعمال کند. نتایج مدل پیشنهادی با شاخص‌های رسمی بانک مرکزی درباره مطالبات غیرجاری و عملکرد تسهیلات بانکی همخوانی دارد و نشان می‌دهد که استفاده از مدل‌های داده‌محور می‌تواند در چارچوب تصمیم‌گیری بانک‌های داخلی برای کاهش ریسک اعتباری و بهبود بازدهی منابع مؤثر باشد.

به‌طور کلی، یافته‌های این پژوهش نشان داد تمرکز بر مشتریان متعلق به خوشه‌های با ارزش طول عمر مشتری بالا، ضمن بهبود بازدهی تخصیص منابع، می‌تواند نرخ نکول را کاهش دهد و کارایی سیاست‌های اعطای تسهیلات را افزایش دهد. این نتایج می‌تواند به مدیران بانکی در طراحی سیاست‌های اعتباری هدفمند، توسعه نظام‌های وفاداری مشتری و مدیریت هوشمند ریسک در چارچوب بانکداری اسلامی کمک نماید.

۴-۱. پیشنهادهای کاربردی

راه حل ارائه شده به صورت کاربردی می تواند در جهت تعیین سطح اعتباری مشتریان بانک و اطمینان بیشتر از تغییر یا عدم تغییر سطح اعتباری آنها بکار گرفته شود. با این وجود رفع محدودیت های ذیل می تواند در بهبود راه حل ارائه شده کمک شایانی نماید.

- بررسی روش های ترکیبی از الگوریتم های مختلف

- استفاده از خوشه بندی مبتنی برمدل های چوله تی و نرمال به ویژه برای داده های با چولگی شدید

- با توجه به استفاده از داده های بانکی در این پژوهش محدودیت دسترسی به داده ها وجود دارد وچالشی جدی محسوب می شود.

- تعیین بازه های بزرگ تر از نظر زمانی و تعداد مشتریان و تحلیل تشابه و تفاوت نتایج آن با بازه در نظر گرفته شده.

منابع

بانک مرکزی. (۱۴۰۱). گزارش اقتصادی و ترازنامه سال ۱۴۰۱. بانک مرکزی جمهوری اسلامی ایران. https://cbi.ir/category/EconomicReport_fa.aspx

بانک مرکزی. (۱۴۰۳). آیین‌نامه وصول مطالبات غیرجاری مؤسسات اعتباری (ریالی و ارزی). بانک مرکزی جمهوری اسلامی ایران https://www.cbi.ir/page/13744.aspx .

مرکز پژوهش‌های مجلس شورای اسلامی. (۱۴۰۳). گزارش سیاست پولی، شهریور ۱۴۰۳. مرکز پژوهش‌های مجلس شورای اسلامی https://rc.majlis.ir/fa/report/show/1832439 .

Beheshtian-Ardakani, A., Fathian, M., & Gholamian, M. (2018). A novel model for product bundling and direct marketing in e-commerce based on market segmentation. Decision Science Letters, 7(1), 39–54.

BHARDWAJ, H., & RAO, D. D. S. (2025). A MODIFIED K-MEANS APPROACH FOR EFFECTIVE CLUSTERING USING WEIGHTED ADJACENT MATRIX. Journal of Theoretical and Applied Information Technology, 103(5).

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5–32.

Chiorazzo, V., D'Apice, V., DeYoung, R., & Morelli, P. (2018). Is the traditional banking model a survivor? Journal of Banking & Finance, 97, 238–256.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society: series B (methodological), 39(1), 1–22.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. kdd,

Fader, P., Hardie, B., & Berger, P. D. (2004). Customer-base analysis with discrete-time transaction data. Available at SSRN 596801.

Feczko, E., & Fair, D. A. (2020). Methods and challenges for assessing heterogeneity. Biological psychiatry, 88(1), 9–17.

Gavett, G. (2014). What you need to know about segmentation. Harvard business review, 9.

Gormley, I. C., Murphy, T. B., & Raftery, A. E. (2023). Model-based clustering. Annual Review of Statistics and Its Application, 10(1), 573–595.

Greenberg, P. (2001). CRM at the speed of light: Capturing and keeping customers in Internet real time. McGraw-Hill Professional.

Grün, B. (2019). Model-based clustering. In Handbook of mixture analysis (pp. 157–192). Chapman and Hall/CRC.

Han, J., & Kamber, M. (2001). Data mining concepts and techniques San Francisco Moraga Kaufman.

Ho, T., Nguyen, S., Nguyen, H., Nguyen, N., Man, D.-S., & Le, T.-G. (2023). An extended RFM model for customer behaviour and demographic analysis in retail industry. Business Systems Research: International journal of the Society for Advancing Innovation and Research in Economy, 14(1), 26–53.

Kessler, D. (2019). Introducing the MBC Procedure for Model-Based Clustering. SAS Proceeding, 1–21.

Kumar, S. (2022). Customer retention versus customer acquisition. Forbes Business Council (https://www. forbes. com/sites/forbesbusinesscouncil/2022/12/12/customer-retention-versus-customeracquisition.

Lancaster, G., & Williams, I. (2002). Consumer segmentation in the grey market relative to rehabilitation products. Management decision, 40(4), 393–410.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124–136.

Liu, D.-R., Lai, C.-H., & Lee, W.-J. (2009). A hybrid of sequential rules and collaborative filtering for product recommendation. Information Sciences, 179(20), 3505–3519.

McLachlan, G. J., & Basford, K. E. (1988). Mixture models. Inference and applications to clustering. Statistics: textbooks and monographs.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert systems with applications, 36(2), 2592–2602.

Nielsen, F. (2016). Hierarchical clustering. In Introduction to HPC with MPI for Data Science (pp. 195–211). Springer.

Reinartz, W., Thomas, J. S., & Kumar, V. (2005). Balancing acquisition and retention resources to maximize customer profitability. Journal of marketing, 69(1), 63–79.

Rizkyanto, H., & Ford, L. G. (2023). Customer segmentation of personal credit using recency, frequency, monetary (RFM) and K-means on financial industry. International Journal of Advanced Computer Science and Applications, 14(4).

Saedi, A., Abbasi, M., & Mehdi Nejad, A. (2025). Customer segmentation using the extended RFMP model based on Customer lifetime value and data mining. International Journal of New Political Economy, 6(1), 23–46.

Sander, J. (2017). Density-based clustering. In Encyclopedia of machine learning and data mining (pp. 349–353). Springer.

Shirole, R., Salokhe, L., & Jadhav, S. (2021). Customer segmentation using rfm model and k-means clustering. Int. J. Sci. Res. Sci. Technol, 8(3), 591–597.

Taherkhani, L., Daneshvar, A., Amoozad Khalili, H., & Sanaei, M. (2025). Analysis and Optimization of Customer Lifetime Value Prediction using Machine Learning and Deep Learning Models by RFM Techniques. International Journal of Web Research, 8(2), 79–92.

Tsiptsis, K. K., & Chorianopoulos, A. (2011). Data mining techniques in CRM: inside customer segmentation. John Wiley & Sons.

Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. European Journal of Operational Research, 238(2), 505–513.

Xiao, E. (2024). Comprehensive K-Means Clustering. Journal of Computer and Communications, 12(3), 146–159.

Zhu, E., & Ma, R. (2018). An effective partitional clustering algorithm based on new clustering validity index. Applied soft computing, 71, 608–621.