فصل‌نامه مطالعاتی در مدیریت بانکی و بانکداری اسلامی

مؤسسه عالی آموزش بانکداری ایران
بهار ۱۴۰۲، دوره ۹، شماره ۲۲، صفحه ۱-۲۲

طراحی مدل اعتبارسنجی مشتریان فعال در صنعت گردشگری با رویکرد ترکیبی مبتنی بر آنتروپی

سید فرید موسوی*[۱]
فریده زنگنه[۲]
سید امیر رضا ابطحی[۳]
آرزو گازری نیشابوری[۴]

چکیده

در سال‌های اخیر، اعتبارسنجی یکی از روش‌های اصلی مؤسسه‌های مالی برای ارزیابی ریسک اعتباری بوده است. در میان روش‌های اعتبارسنجی، مشکل اصلی توزیع نامتوازن داده‌هاست که کارایی روش‌های اعتبارسنجی را محدود می‌کند. دلیل ایجاد مشکل یادشده این است که در مجموعه داده مشتریان، موارد بدحساب (نکول) موجود برای آموزش مدل ارزیابی، کمتر از موارد خوش‌حساب (عدم نکول) است و عملکرد رویکردهای اعمال‌شده در اعتبارسنجی را مختل می‌کند. پژوهش حاضر با استفاده از الگوریتم مبتنی بر آنتروپی و با هدف غلبه بر مشکل نامتوازن بودن داده‌ها، به اعتبارسنجی مشتریان فعال در صنعت گردشگری پرداخته است. در این پژوهش داده‌ها بر حسب آنتروپی شاخص‌های اعتبارسنجی مشتریان ارزیابی شده و معیاری تعریف خواهد شد که می‌تواند خوش‌حسابی یا بدحسابی مشتریان را تنها با در نظر گرفتن موارد خوش‌حساب مجموعه داده و نمونه متقاضی تسهیلات، اندازه‌گیری کند. در این پژوهش، ۲۰۴ مشتری فعال صنعت گردشگری بانک ملی ایران، به‌عنوان مجموعه داده انتخاب شده است. بر اساس نتایج پژوهش، مدل آنتروپی با قدرت پیش‌بینی خوب خود، برای اعتبارسنجی مشتریان کارایی مناسبی دارد.

واژه‌های کلیدی: اعتبارسنجی، مشتریان بانکی، صنعت گردشگری، رویکرد ترکیبی، آنتروپی.

طبقه‌بندی JEL: G۲ و G۳.

---

۱. استادیار، گروه مدیریت عملیات و فناوری اطلاعات، دانشکده مدیریت، دانشگاه خوارزمی، تهران، ایران (نویسنده مسئول)؛ mousavifarid@khu.ac.ir
۲. کارشناسی ارشد، گروه مدیریت صنعتی، دانشگاه خوارزمی، تهران، ایران؛ fzangenehh@gmail.com
۳. استادیار، گروه مدیریت عملیات و فناوری اطلاعات، دانشکده مدیریت، دانشگاه خوارزمی، تهران، ایران؛ abtahi@khu.ac.ir
۴. دکتری، گروه مهندسی صنایع، واحد تهران جنوب، دانشگاه آزاد اسلامی، تهران، ایران؛ arezoogazori.66@gmail.com

مقدمه

امروزه، با جهانی شدن اقتصاد و تشدید رقابت بین بانک‌ها، حاشیه سود فعالیت‌های سنتی بانکداری کاهش یافته و موجب افزایش ریسک در بانک‌ها شده است (فلاح‌پور، راعی و هندیجانی زاده، ۱۳۹۳). بانک‌ها در معرض چند ریسک مالی قرار دارند. در این خطرهای مالی ممکن است سپرده‌گذاران ناگهان سپرده‌های خود را پس بگیرند (ریسک نقدینگی)[۱]، وام‌گیرندگان وام خود را به‌موقع بازپرداخت نکنند (ریسک اعتباری)[۲]، نرخ بهره تغییر کند (ریسک نرخ بهره)[۳] و سیستم‌های رایانه‌ای بانک خراب شود (ریسک عملیاتی)[۴]. با وجود این، در میان این ریسک‌ها، ریسک‌های اعتباری و نقدینگی نه‌فقط ریسک‌های مهمی هستند که بانک‌ها با آن‌ها روبه‌رو هستند، بلکه به‌طور مستقیم با عملکرد و علت شکست بانک‌ها نیز ارتباط دارند (چسکتی و شونولز[۵]، ۲۰۱۱).

در سال‌های اخیر، اعتبارسنجی[۶] به یکی از روش‌های اصلی مؤسسه‌های مالی برای ارزیابی ریسک اعتباری، بهبود جریان نقدینگی، کاهش خطرهای احتمالی و اتخاذ تصمیم‌های مدیریتی تبدیل شده است (هوانگ، چن و وانگ[۷]، ۲۰۰۷). دقت اعتبارسنجی برای سودآوری مؤسسه‌های مالی بسیار مهم است. حتی یک درصد بهبود دقت در تشخیص متقاضیان با اعتبار نامناسب، ضرر بزرگی را برای مؤسسه‌های مالی کاهش می‌دهد (هند و هنلی[۸]، ۱۹۹۷). از این رو، یک مدل اعتبارسنجی خوب قادر خواهد بود مشتریان را به‌طور مؤثر به دو گروه دارای ریسک نکول[۹] یا عدم نکول[۱۰] طبقه‌بندی کند. هرچه مدل کارآمدتر باشد، هزینه بیشتری برای یک مؤسسه مالی صرفه‌جویی می‌شود (گو و لی[۱۱]، ۲۰۱۹). مشکل اصلی‌ای که کارایی روش‌های طبقه‌بندی اعتبارسنجی را محدود می‌کند، توزیع نامتوازن داده‌ها[۱۲] است (باتیستا، پرتی و مونارد[۱۳]، ۲۰۰۴). داده‌های نامتوازن به وضعیتی اطلاق می‌شود که یک کلاس که به‌عنوان اقلیت یا کلاس مثبت نامیده می‌شود، از کلاس دیگری که

---

1. Liquidity risk
2. Credit Risk
3. Interest Rate Risk
4. Operational Risk
5. Ceschetti & Schoenholtz
6. Credit Scoring
7. Huang, Chen &Wang
8. Hand & Henley
9. Default
10. Non-default
11. Goh & Lee
12. Unbalanced distribution of data
13. Batista, Prati & Monard

به‌عنوان کلاس اکثریت یا منفی نامیده می‌شود، بیشتر شده و در نتیجه، توزیع نابرابر نمونه‌ها انجام شود. داده‌های نامتوازن یک مسئله رایج در اعتبارسنجی است، به این معنا که در مجموعه داده‌های مشتریان بانکی تعداد نمونه مشتریان خوب به‌مراتب بیشتر از تعداد نمونه‌های بد است. این اهم باعث مشکلی می‌شود که نتایج طبقه‌بندی را می‌توان با سوگیری داده‌ها نسبت به طبقه اکثریت منحرف کرد که به زیان‌های مالی شایان توجهی ناشی از طبقه‌بندی نادرست نمونه بد به‌عنوان نمونه خوب منجر خواهد شد (شن، ژاو، لی، منگ و لی[1]، ۲۰۱۹).

برای غلبه بر چنین موضوعی، در این پژوهش مشتریان بانک را از نظر ویژگی‌های آنتروپی شاخص‌های آن‌ها ارزیابی کرده و روشی اجرا خواهیم کرد که می‌تواند سطح اطمینان آن‌ها را فقط با توجه به موارد خوش‌حساب و نمونه بررسی‌شده (شخص متقاضی وام) اندازه‌گیری کند و در عین حال، مشکل شروع سرد را نیز کاهش دهد (شارما، ویشراج، میتال، احلوت و میتال[2]، ۲۰۲۰). با توجه به اینکه، عمده پژوهش‌های انجام‌شده در زمینه اعتبارسنجی، پژوهش‌هایی هستند که مدل‌های استفاده‌شده در اعتبارسنجی را تأیید می‌کنند و به رتبه‌بندی و تفکیک مشتریان می‌پردازند، در این پژوهش علاوه بر استفاده از مدل اعتبارسنجی آنتروپی که دارای مزیت‌های متفاوتی (امکان غلبه بر عدم تعادل داده‌ها و مشکل شروع سرد) نسبت به بقیه مدل‌های اعتبارسنجی است، سعی بر آن بود که با بررسی پژوهش‌های پیشین معیارهایی برای ارزیابی انتخاب شوند که با ویژگی‌های صنعت بانکداری ایران متناسب باشند و با استفاده از آن‌ها بتوان با دقت بیشتری متقاضیان بخش حقوقی و حقیقی تسهیلات سرمایه در گردش مشتریان بانکی فعال در صنعت گردشگردی را رتبه‌بندی و تفکیک کرد.

مبانی نظری و پیشینه پژوهش

یکی از وظایف اصلی در بانک‌ها و مؤسسه‌های مالی، ارزیابی ریسک اعتباری است. زیرا از این راه می‌توان از خسارت‌های کلان ناشی از تصمیم‌های نادرست اعتباردهی به متقاضیان، تا اندازه زیادی جلوگیری کرد. اعتبارسنجی، به‌معنای ارزیابی توان بازپرداخت وام و تسهیلات مالی توسط متقاضیان و همچنین سنجش احتمال عدم بازپرداخت اعتبارات دریافتی از سوی آن‌هاست (کیس[3]، ۲۰۰۳). بنا بر تعریف کمیته بال به احتمال عدم بازپرداخت تسهیلات اعطایی به مشتریان و به‌عبارتی، ضرر

1. Shen & Zhao & Li & Li & Meng
2. Sharma, Vishraj, Ahlawat, Mittal & Mittal
3. Kiss

حاصل از عدم ایفای تعهد به‌وسیله بدهکار یا طرف مقابل بانک را ریسک اعتباری گویند. در واقع، می‌توان گفت که ارزیابی و سنجش توان بازپرداخت متقاضیان اعتبار تسهیلات مالی و از طرفی، احتمال عدم بازپرداخت اعتبارات دریافتی توسط آنان، اعتبارسنجی نامیده می‌شود. همچنین، رتبه‌سنجی (مشخص کردن رتبه اعتباری) در واقع بینش ضروری در خصوص شناخت ریسک اعتباری مشتریان را برای بنگاه‌های مالی و اعتباری فراهم می‌کند.

از لحاظ نظری، بسیاری از پژوهشگران در این زمینه، تعاریف زیادی از اعتبارسنجی ارائه داده‌اند که برخی از آن‌ها شرح داده شده است:

به‌گفته توماس، ادلمن و کروک[1] (۲۰۰۲) اعتبارسنجی به‌عنوان «مجموعه‌ای از مدل‌های تصمیم‌گیری و تکنیک‌های اساسی آن‌ها که به وام‌دهندگان در اعطای اعتبار مشتری کمک می‌کند»، تعریف می‌شود (هند و هنلی، ۱۹۹۷). اعتبارسنجی را به‌عنوان اصطلاحی که برای توصیف روش‌های آماری رسمی استفاده‌شده برای طبقه‌بندی متقاضیان اعتبار در رده‌های «خوب» و «بد» استفاده می‌شود، تعریف کردند.

تعریف دیگر مبتنی بر «اختصاص یک اندازه‌گیری یا امتیاز واحد به یک وام‌گیرنده بالقوه است که برآورد عملکرد وام بعدی وام‌گیرنده را نشان می‌دهد» (فریم، سرینیواسان و ووسلی[2]، ۲۰۰۱).

پیشینه پژوهش در زمینه مدل‌های اعتبارسنجی

هدف اصلی، هنگام ایجاد یک مدل اعتبارسنجی، ایجاد بهترین تکنیک‌های طبقه‌بندی است که می‌تواند بین اعتبار خوب و بد تمایز قائل شود و بر این اساس، رفتار متقاضیان جدید وام را پیش‌بینی کند. مدل‌های اعتبارسنجی به‌طور گسترده‌ای در حوزه مالی و به‌ویژه در بانک‌ها استفاده می‌شود. برای نخستین بار فیشر (۱۹۳۶)، استفاده از تکنیک‌های آماری را برای حل یک مسئله طبقه‌بندی معرفی کرد و فیر آیزاک در اواخر دهه ۱۹۶۰ ایجاد شرکت اعتبارسنجی را پیشنهاد داد (توماس[3]، ۲۰۰۰). از آن زمان، تکنیک‌های آماری در توسعه روش‌های اعتبارسنجی تا ظهور تکنیک‌های یادگیری ماشین یا هوش مصنوعی (AI) که با تکامل فناوری‌های کامپیوتری ظهور کرد، به‌کار گرفته

1. Thomas, Edelman & Crook
2. Frame, Srinivasan & Woosley
3. Thomas

شد. با این حال، اعتقاد بر این است که تکنیک‌های یادگیری ماشین در مقایسه با تکنیک‌های آماری، عملکرد بهتری دارند (هوانگ، چن، هسو، چن و وو[1]، ۲۰۰۴).

در پژوهش (لی، مونکدالایی و ریو[2]، ۲۰۲۰) مدل نمره‌گذاری اعتباری ترکیبی[3] با استفاده از شبکه‌های عصبی عمیق[4] و رگرسیون لجستیک برای بهبود دقت پیش‌بینی آن پیشنهاد می‌شود. مدل نمره‌گذاری اعتبار ترکیبی پیشنهادی شامل دو مرحله است. در مرحله نخست، چند مدل شبکه عصبی را آموزش دادند و در مرحله دوم، این مدل‌ها با رگرسیون لجستیک ادغام می شوند. نتایج نشان داد که مدل پیشنهادی از نظر H-measure، مساحت زیر منحنی (AUC) و دقت، از مدل‌های پایه در بیش از سه مجموعه داده معیار بهتر عمل کرد.

رحمان، اینگنادیس، حاتمی ماربینی، دامودارن و خوشنویس[5] (۲۰۱۸) در پژوهش خود ابزار پشتیبانی تصمیم، در خصوص مدل امتیاز اعتباری براساس اصول تصمیم‌گیری چندمعیاره[6] توسعه می‌دهند. در روش پیشنهادی، وزن معیارها توسط AHP فازی شد. نظریه زبانی فازی[7] در AHP برای توصیف عدم قطعیت‌ها و ابهامات ناشی از ذهنیت افراد در تصمیم‌گیری‌ها استفاده شده و در نهایت، با استفاده از تابع فاصله ریسک[8]، TOPSIS برای رتبه‌بندی گزینه‌ها براساس کمترین ریسک استفاده شد. تجزیه و تحلیل حساسیت نیز با استفاده از روش فازی AHP-TOPSIS نشان داده شد.

هریس[9] (۲۰۱۵) در پژوهشی، عملکرد امتیازدهی اعتباری و معرفی استفاده از ماشین بردار پشتیبان خوشه‌بندی[10] شده (CSVM) را بررسی می‌کند. بر این اساس، این پژوهش CSVM را با سایر تکنیک‌های مبتنی بر SVM غیرخطی مقایسه کرده و نشان داد که CSVM می‌تواند به سطوح قابل مقایسه عملکرد طبقه‌بندی دست یابد، در حالی که از نظر محاسباتی به‌نسبت ارزان باقی می‌ماند.

حسین‌زاده کاشان و گروسی (۱۳۹۹) در پژوهشی، روش ماشین بردار پشتیبان (SVM) را به‌عنوان طبقه‌بندی‌کننده اصلی با یک روش انتخاب ویژگی به نام الگوریتم مورچگان باینری

---

1. Huang, Chen, Hsu, Chen & Wu
2. Lee, Munkhdalai & Ryu
3. Hybrid credit scoring model
4. Deep neural networks
5. Rahman, Ignatius, Hatami-Marbini, Dhamotharan & Khoshnevis
6. Multi-criteria decision-making
7. Fuzzy linguistic theory
8. Risk distance function
9. Harris
10. Clustered support vector machine

(BACO-SVM)[1] ترکیب کرده و از داده‌های مربوط به ۸۵ شرکت از تسهیلات گیرنـدگان حقـوقی یک بانک ایرانی در یک بازه پنج‌ساله (۱۳۸۹ تا ۱۳۹۳) به‌همراه ۱۶ ویژگی مربوط به هر یک از آن‌ها استفاده کردنـد. نتـایج روش BACO-SVM بـا روش PSO-SVM[2]، GA-SVM و روش SVM به‌تنهایی مقایسه شده است. یافته‌های پژوهش بر این دلالت داشت که در ارزیابی ریسـک اعتبـاری، مدل BACO-SVM در مقایسه با روش‌های دیگر از عملکرد خوبی برخـوردار اسـت. در نتیجـه، بـا استفاده از روش BACO-SVM به طبقه‌بنـدی مشـتریان بـه دو گـروه مشـتریان خـوش‌حسـاب و بدحساب پرداختند.

روش‌شناسی پژوهش

از آنجا که هدف این پژوهش، شناسایی مشتریانی است که احتمال نکول آن‌ها بالاست، این پژوهش از نظر هدف و نتیجه، کاربردی است. از نظر نوع شاخص داده‌های مشتریان هر دو طیف شاخص‌های کمی و کیفی در این پژوهش استفاده خواهند شد که برای اعتبارسنجی شاخص‌های کیفی به مقـادیر کمی تبدیل می‌شوند. جامعه آماری در این پژوهش شامل ۲۰۴ نفر از مشتریان فعال در حوزه صـنعت گردشگری طی سال های ۱۳۹۵ تا۱۴۰۰ است که از کلیه شعب منتخب بانـک ملـی ایـران در نظـر گرفته شده‌اند. دلیل انتخاب مشتریان، در دسترس بودن داده‌های مالی موثق و حسابرس شده آن‌هـا است. این جامعه آماری متشکل از مشتریان خوش‌حساب (وضعیت بازپرداخـت مطلـوب) و مشـتریان بدحساب (وضعیت پرداخت نامطلوب) است.

در گام نخست، شاخص‌های اصلی تأثیرگذار بر رتبه اعتباری مشتریان فعال در صنعت گردشگری شناسایی خواهند شد. سپس، داده‌های مشتریان جمـع‌آوری شـده و پـس از نرمـالایز کـردن توسـط الگوریتم مبتنی بر آنتروپی، اعتبارسنجی خواهند شد.

شناسایی شاخص‌های مؤثر بر اعتبارسنجی مشـتریان فعـال در صـنعت گردشـگری بـا اسـتفاده از نظرهای خبرگان بانکی انجام شده است. در نهایت، شاخص‌های مدت زمان رابطه مشتری بـا بانـک، سرمایه مشتری، درآمد حاصل از فروش و ارائه خدمات، نسبت حاشیه سود، نسبت بازده دارایی، نسبت بدهی، نسبت بدهی جاری، نسبت جاری، نسبت آنی، نسبت دارایی‌های جاری، سابقه چـک برگشـتی (دارد یا ندارد)، تحصیلات مدیرعامل یا مالک، نوع وثیقه (سند رهنی یا سایر وثیقه‌هـا)، نـوع مالکیـت

1. Binary Ant Colony Optimization Algorithm
2. Particle Swarm Optimization

محل فعالیت، سابقه بدهی سایر بانک‌ها (دارد/ ندارد)، سابقه بـدهی معـوق (دارد/ نـدارد)، داشـتن یـا نداشتن گزارش حسابرسی، دارا بودن بدهی مالیاتی (دارد/ ندارد) و روند سودآوری شرکت در سه سال گذشته (افزایشی/ نوسانی/ کاهشی)، به‌عنوان شاخص‌هایی که شامل اطلاعات مالی و همچنین مؤثر در تحلیل رفتار بازپرداختی مشتری هستند، استخراج شده است.

جدول ۱. شاخص‌های اعتبارسنجی در حوزه صنعت گردشگری

شاخص‌های کمی		
ردیف	**عنوان شاخص**	**تعریف شاخص**
۱	مدت زمان رابطه مشتری با بانک	استمرار، تمرکز و تداوم فعالیت یـک مـشتری و سابقه آن نـــزد بانـک، باعـث آگـاهی بیشتر از میزان و نحوه فعالیت‌ها وتوانایی‌های وی از یک سو وکاهش ریسک‌های آتـی از سوی دیگر می‌شود.
۲	سرمایه مشتری	ارزش و اهمیت کسب‌وکار مشتری را می‌توان از میزان سرمایه آن‌ها درک کرد. هر قدر سرمایه ثبت شده بیـشتر باشـد، برای افتتـاح حـساب و دریافـت تـــسهیلات از اعتبـار بالاتری برخـوردار خواهند بود.
۳	درآمد حاصل از فروش و ارائه خدمات	در شرکت‌های خدماتی درآمد آن‌ها با عنوان حساب درآمد حاصل از خدمات ارائه شده و در شرکت‌های پیمانکاری حساب درآمد ناخالص پیمانکاری ثبت می‌شود.
۴	نسبت حاشیه سود	تقسیم سود (خالص) بر درآمد یا فروش به‌دست‌آمـده و میـزان سـودآوری را مشـخص می‌کند.
۵	نسبت بازده دارایی	از تقسیم کل دارایی ها بر سود به دست می‌آید.
۶	نسبت بدهی	(دارایی کل/ بدهی کل) میزان وجوهی را نشان می‌دهد کـه بـه‌وسـیله بـدهی تـأمین شده است.
۷	نسبت بدهی جاری	(بدهی جاری/ دارایی جاری). نسبت جاری نـشان مـی‌دهد کـه دارایـی جـاری تـا چـه اندازه بدهی جـاری را می‌پوشاند. این نـسبت را مقیاس نقدینگی در کوتاه‌مدت می‌دانند.
۸	نسبت جاری	(بدهی جاری/ موجودی نقد+ مطالبات) ایـن نـسبت بـه وضوح نشان می‌دهـد کـه آن قسمت از دارایی جاری که از لحـاظ ارزش، ثبات بیشتری دارد و احتمال کـاهش در آن کمتر است تـا چه میزان می‌تواند پشتوانه طلبکاران کوتاه‌مدت قرار گیرد.
۹	نسبت آنی	این نسبت نشان‌دهنده دارایی جاری به دارایی کل است.
۱۰	نسبت دارایی‌های جاری	این نسبت نشان‌دهنده دارایی جاری به دارایی کل است.

ادامهٔ جدول ۱

شاخص‌های کیفی		
ردیف	**عنوان شاخص**	**تعریف شاخص**
۱۱	سابقه چک برگشتی (دارد یا ندارد)	بررسی وضعیت چک‌های برگشتی مشتری از گذشته تا امروز است که در این پژوهش به‌صورت کیفی (دارد/ ندارد) بیان می‌شود.
۱۲	تحصیلات مدیرعامل یا مالک	–
۱۳	نوع وثیقه (سند رهنی یا سایر وثیقه‌ها)	یکی از عواملی که حسن نیت مشتری را در پای‌بندی به ایفای به‌موقع تعهدات خود نشان می‌دهد، ارائه وثایق قابل ترهین و تضمینات متقاضی نزد مؤسسه است.
۱۴	نوع مالکیت محل فعالیت	منظور مالکیت دفتر یا مرکز تجاری (استیجاری، رهنی، ملکی و شهرک‌های صنعتی) معتبری است که مشتری طی سالیان متمادی در آن مشغول فعالیت هستند.
۱۵	سابقه بدهی سایر بانک‌ها (دارد/ندارد)	منظور داشتن پیشینه عدم پرداخت تسهیلات در زمان مقررشده، توسط مشتری در سایر بانک‌ها به‌ازای دریافت تسهیلات است.
۱۶	سابقه بدهی معوق (دارد/ ندارد)	منظور داشتن پیشینه عدم پرداخت تسهیلات در زمان مقرر در بانک ارائه‌کننده تسهیلات است.
۱۷	داشتن یا نداشتن گزارش حسابرسی	گزارشی است که حسابرس پس از بررسی بی‌طرفانه صورت‌های مالی یک کسب‌وکار آماده می‌کند.
۱۸	دارا بودن بدهی مالیاتی (دارد/ ندارد)	این شاخص وجود و نبود بدهی مالیاتی مشتری را نمایش می‌دهد که در این پژوهش به‌طور کیفی (دارد/ ندارد) بیان می‌شود.
۱۹	روند سودآوری شرکت در سه سال گذشته (افزایشی/ نوسانی/ کاهشی)	میزان سود خالص سه سال گذشته شرکت را محاسبه و به‌تبع آن روند سود را از بابت صعودی یا نزولی بودن آن نمایش می‌دهد.

پیاده‌سازی رویکرد ترکیبی مبتنی بر آنتروپی

با توجه به مجموعه‌ای از داده‌های طبقه‌بندی‌شده شامل مشتریان خوش‌حساب و بدحساب $T = \{t_1, t_2, \dots, t_k\}$ و مجموعه‌ای از ویژگی‌های نمونه‌ها $F = \{f_1, f_2, \dots, f_M\}$ که هر $t \in T$، مجموعه $T_+ = \{t_1, t_2, \dots, t_N\}$ را به‌عنوان زیرمجموعه‌ای از نمونه‌هایی که تسهیلات را بازپرداخت کرده‌اند و خوش‌حساب هستند (پس خواهیم داشت $T_+ \subseteq T$) و مجموعه $T_- = \{t_1, t_2, \dots, t_J\}$ مجموعه نمونه‌هایی که تسهیلات را بازپرداخت نکرده‌اند و بدحساب هستند (پس خواهیم داشت $T_- \subseteq T$) تعریف می‌کنیم.

همچنین، مجموعه $\hat{T} = \{\hat{T}_1, \hat{T}_2, \dots, \hat{T}_U\}$ را به‌عنوان مجموعه نمونه‌های جدید طبقه‌بندی‌نشده که قرار است اعتبارسنجی شوند و متقاضی دریافت تسهیلات هستند و $E = \{e_1, e_2, \dots, e_U\}$ مجموعه این نمونه‌ها پس از طبقه‌بندی و تقسیم به خوش‌حساب یا بدحساب تعریف کرده‌ایم.

باید توجه داشت که یک نمونه فقط می‌تواند متعلق به یک کلاس خوش‌حساب یا بدحساب باشد ($c \in C$) که C={ قابل اعتماد، غیرقابل اعتماد}. با توجه به چنین مفاهیمی، اجرای رویکرد در چهار مرحله زیر خواهد بود:

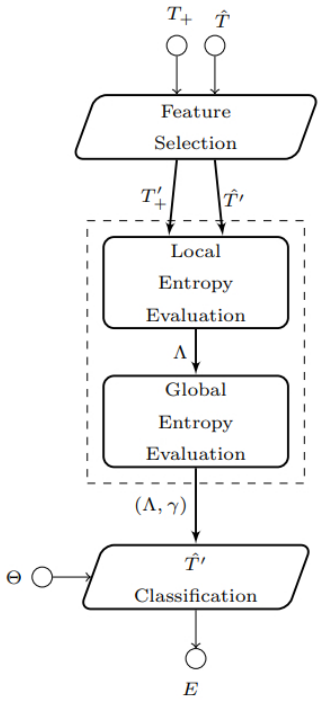
۱. فرایند انتخاب ویژگی: ارزیابی هر ویژگی (شاخص) نمونه به‌منظور ارزیابی سهم آن در چارچوب تعریف مدل ارزیابی.
۲. محاسبه آنتروپی محلی[1]: محاسبه آنتروپی محلی Λ که اطلاعاتی در خصوص سطح آنتروپی در نظر گرفته‌شده توسط هر یک از ویژگی‌های منفرد در مجموعه T_+ می‌دهد.
۳. محاسبه آنتروپی سراسری[2]: محاسبه آنتروپی سراسری γ، یک متا اطلاعات که با محاسبه انتگرال ناحیه زیر منحنی Λ تعریف می‌شود.
۴. رویکرد تفاوت آنتروپی[3]: تعریف رویکرد تفاوت آنتروپی (EDA) که می‌تواند نمونه‌های جدید را براساس اطلاعات Λ و γ طبقه‌بندی کند. نحوه اجرای رویکرد EDA پیشنهادی در شکل ۱ نشان داده شده است.

در مرحله نخست، مجموعه نمونه‌های قابل اعتماد قبلی T_+ (مشتریان خوش‌حساب) و مجموعه نمونه‌هایی که باید ارزیابی شوند که خوش‌حساب هستند یا بدحساب $\hat{T}$ ، به‌عنوان ورودی الگوریتم انتخاب ویژگی وارد می‌شوند که هدف آن، حذف شاخص‌هایی با سطح پایین ارتباط از فرایند ارزیابی است، به این معنا که شاخصی که با خروجی (خوش‌حسابی یا بدحسابی) ارتباط کمتری دارد، حذف شده و در فرایند اعتبارسنجی استفاده نمی‌شود. این مرحله، پیچیدگی محاسباتی را کاهش می‌دهد و مجموعه‌های با ویژگی‌های کاهش‌یافته $\hat{\hat{T}}$ و $\acute{T}_+$ را بر می‌گرداند (شاخص‌هایی که از اهمیت کمتری برخوردارند حذف خواهند شد).

1. Local Entropy Calculation
2. Global Entropy Calculation
3. Entropy Difference Approach

در مراحل بعدی، آنتروپی محلی برای هر یک از ویژگی‌های مجموعه $\acute{T}_+$ و همچنین آنتروپی سراسری تمام ویژگی‌های $\acute{T}_+$ محاسبه می‌شود. مرحله آخر، مقایسه بین آنتروپی محلی و سراسری است که پیش‌تر برای مجموعه $\acute{T}_+$ محاسبه شده است و همان اطلاعات بعد از اضافه کردن هر عنصر از مجموعه $\hat{T}$ به $\acute{T}_+$ محاسبه می‌شود (به‌عبارتی، بعد از هر نمونه جدید که برای طبقه‌بندی وارد الگوریتم می‌شود، آنتروپی محلی و سراسری مجدد محاسبه می‌شود) و نمونه‌های ارزیابی‌نشده را براساس حد آستانه Θ طبقه‌بندی می‌کند.

در روش پیشنهادی با عنوان رویکرد ترکیبی مبتنی بر آنتروپی، انتخاب ویژگی با بهره‌برداری از یک رویکرد مبتنی بر آنتروپی دوگانه انجام می‌شود که اهمیت ویژگی‌ها را به‌صورت جداگانه و متقابل ارزیابی می‌کند. برای این منظور، از دو معیار استفاده می‌کنیم که به‌صورت زیر تعریف شده است.



شکل ۱. مراحل اجرای رویکرد مبتنی بر آنتروپی (کارتا، فریرا، ریکوپرو، سایا و سایا[1]، ۲۰۲۰)

1. Carta, Ferreira, Recupero, Saia & Saia

آنتروپی شانون پایه: عدم قطعیت مربوط به یک متغیر تصادفی را با ارزیابی میانگین حداقل تعداد بیت‌های مورد نیاز برای رمزگذاری رشته‌ای از نمادها براساس فرکانس آن‌ها اندازه‌گیری می‌کند یا به بیان دیگر، آنتروپی معیاری است که میزان عدم قطعیت یک متغیر تصادفی را مشخص می‌کند. مقادیر بالای آنتروپی نشان‌دهنده سطح بالایی از عدم قطعیت در فرایند پیش‌بینی داده‌ها و برعکس، مقادیر پایین آنتروپی نشان‌دهنده درجه پایین‌تر عدم قطعیت در این فرایند است. به‌طور کلی، با توجه به مجموعه‌ای از مقادیر $f \in F$، آنتروپی H(F) همان‌طور که در رابطه ۱ نشان داده شده است، تعریف می‌شود، در این رابطه P(f) احتمال وجود عنصر f در مجموعه F است.

$$H(F) = -\sum_{f \in F} P(f) \log_2[P(f)] \qquad \text{رابطهٔ ۱)}$$

آنتروپی شانون اطلاعات متقابل: مقدار اطلاعاتی که یک متغیر تصادفی در خصوص متغیر دیگر می‌دهد را اندازه گیری می‌کند (به‌عبارتی، در این رویکرد ارتباط هر یک از ویژگی‌ها با خروجی محاسبه می‌شود و در نهایت، ویژگی مناسب، ویژگی است که ارتباط مستقیم با خروجی داشته باشد). به‌طور کلی با توجه به دو متغیر گسسته X و Y با توزیع احتمال مشترک $P_{XY}(x, y)$، که اطلاعات متقابل بین X و Y را به‌عنوان $\mu(X, Y)$ نشان می‌دهد، آنتروپی شانون متقابل همان‌طور که در رابطه ۲ نشان داده شده است، محاسبه می‌شود.

$$\mu(X, Y) = \sum_{x,y} P_{XY}(x, y) \log_2 \frac{P_{XY}(x, y)}{P_X(x) P_Y(y)} \qquad \text{رابطهٔ ۲)}$$

با در نظر گرفتن این دو معیار، انتخاب ویژگی از طریق مراحل زیر انجام می‌شود:

۱. آنتروپی اصلی هر یک از ویژگی‌ها اندازه‌گیری می‌شود و سهم آن در توصیف رفتار نمونه ارزیابی می‌شود.
۲. آنتروپی متقابل هر ویژگی با توجه به ویژگی‌های دیگر ارزیابی می‌شود.
۳. نتایج دو مرحله قبلی با هم ترکیب می‌شوند و ویژگی‌هایی را انتخاب می‌کنند که در فرایند تعریف مدل استفاده می‌شوند.

الگوریتم مجموعه T_+ از نمونه‌های قابل اعتماد، مجموعه $\hat{T}$ از نمونه‌های ارزیابی‌نشده و مقادیر $\min_1$ و $\min_2$ را به‌عنوان ورودی دریافت می‌کند که نشان‌دهنده آستانه‌های استفاده‌شده برای تعیین مقدار آنتروپی مناسب برای انتخاب ویژگی است.

سپس الگوریتم، دو مجموعه از نمونه‌های $\acute{T}_+$ و $\hat{\acute{T}}$ را برمی‌گرداند که فقط شامل ویژگی‌هایی هستند که توسط الگوریتم حذف شده‌اند تا از آن‌ها در فرایند تعریف مدل استفاده شود.

محاسبه حداکثر آنتروپی محلی: با نشان دادن $\mathrm{H}(\acute{f})$ آنتروپی اندازه‌گیری‌شده به‌ازای هر یک از ویژگی‌های $\acute{f} \in \acute{F}$ در مجموعه $\acute{T}_+$، مجموعه Λ را به‌عنوان آنتروپی به‌دست‌آمده با هر $\acute{f} \in \acute{F}$ تعریف می‌کنیم. بنابراین خواهیم داشت: $|\Lambda| = |\acute{F}|$

چنین محاسبه‌ای همان‌طور که در رابطه ۳ نشان داده شده است، انجام می‌شود:

رابطهٔ ۳)
$$\Lambda = \left\{\lambda_1 = \max\left(H(\acute{f}_1)\right), \lambda_2 = \max\left(H(\acute{f}_2)\right), \dots, \lambda_M = \max\left(H(\acute{f}_M)\right)\right\}$$

در رویکرد تفاوت آنتروپی پیشنهادی، چنین متریکی دو بار محاسبه می‌شود؛ قبل و بعد از اضافه کردن یک نمونه ارزیابی‌نشده $\hat{t} \epsilon \hat{T}$ به $\acute{T}_+$.

محاسبه حداکثر آنتروپی سراسری

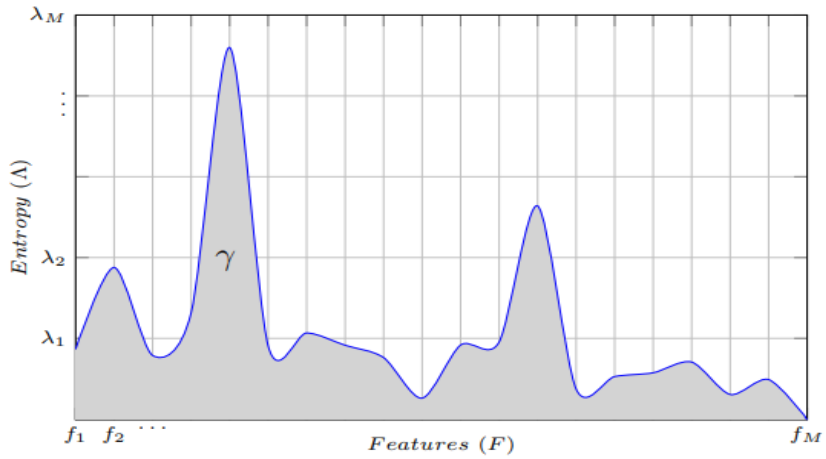
همان‌طور که در شکل ۲ نشان داده شده است، انتگرال ناحیه زیر منحنی آنتروپی محلی Λ را به‌عنوان حداکثر آنتروپی سراسری γ نشان می‌دهیم.

به‌طور کلی، مقدار γ با استفاده از قانون ذوزنقه، همان‌طور که در رابطه زیر نشان داده شده است، محاسبه می شود:

رابطهٔ ۴)
$$\gamma = \int_{\lambda_1}^{\lambda_M} f(x)dx \approx \frac{\Delta x}{2}\sum_{n=1}^{|\Lambda|}(f(x_{n+1}) + f(x_n))$$

$$\Delta x = \frac{(\lambda_M - \lambda_1)}{|\Lambda|}$$

$$\gamma = \int_{\lambda_1}^{\lambda_M} \mathrm{f(x)dx} \approx \frac{\Delta x}{2}\sum_{n=1}^{|\Lambda|}(\mathrm{f(x_{n+1}) + f(x_n)}) \quad (4\_۳)$$



شکل ۲. محاسبه آنتروپی سراسری

آنتروپی سراسری که از مساحت زیر نمودار به دست آمده است، یک متا ویژگی است که اطلاعاتی در خصوص آنتروپی به‌دست‌آمده توسط تمام ویژگی‌های موجود در $\acute{T}_+$ قبل و بعد از اضافه شدن نمونه ارزیابی‌نشده می‌دهد. از این اطلاعات طی فرایند ارزیابی، به‌طور مشترک با اطلاعات ارائه‌شده توسط آنتروپی محلی Λ در معادله ۳ استفاده می‌کنیم.

رویکرد تفاوت آنتروپی

رویکرد تفاوت آنتروپی پیشنهادی (EDA) قادر است مجموعه‌ای از نمونه‌های ارزیابی‌نشده مشتریان (مشتریان جدید متقاضی تسهیلات) را به‌عنوان مشتری قابل اعتماد یا غیرقابل اعتماد ارزیابی و طبقه‌بندی کند. در این قسمت الگوریتم یک مجموعه $\acute{T}_+$ از نمونه‌های قابل اعتماد با ویژگی‌های کاهش‌یافته (حذف ویژگی‌های کم‌اهمیت در مرحله انتخاب ویژگی)، مجموعه $\hat{\acute{T}}$ از نمونه‌های ارزیابی‌نشده با همان ویژگی‌های کاهش‌یافته و یک آستانه آموزش‌دیده قبلی Θ به‌عنوان ورودی می‌گیرد. سپس، مجموعه E را به‌عنوان خروجی برمی‌گرداند که شامل تمام نمونه‌های $\hat{\acute{T}}$ است که بسته به اطلاعات آنتروپی محلی و سراسری Λ و γ به‌عنوان قابل اعتماد یا غیرقابل اعتماد طبقه‌بندی شده‌اند.

در مرحله بعد، الگوریتم مقدار آنتروپی محلی Λ_a از نمونه‌های خوش‌حساب قبلی با ویژگی‌های کاهش‌یافته در $\acute{T}_+$ را همان‌طور که توضیح داده شد، محاسبه می‌کند، سپس، آنتروپی سراسری γ را در همان مجموعه داده به دست می‌آورد. به‌طور کلی، قابلیت اطمینان یک نمونه جدید را از نظر مقایسه آنتروپی اندازه‌گیری‌شده در مجموعه‌ای از نمونه‌های قابل اعتماد قبلی، قبل و بعد از افزودن نمونه بررسی‌شده، ارزیابی می‌کند.

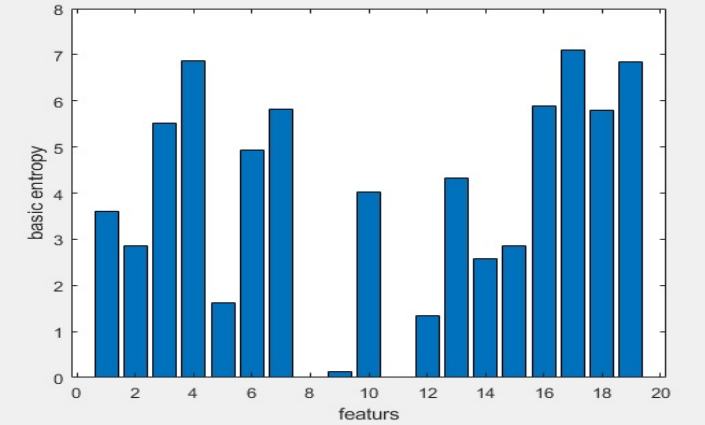
همان‌طور که آنتروپی عدم قطعیت یک متغیر تصادفی را اندازه‌گیری می‌کند، آنتروپی بزرگ‌تر در مجموعه شامل نمونه بررسی‌شده نشان می‌دهد که دارای داده‌های مشابهی در ویژگی‌های آن است و آن را به‌عنوان قابل اعتماد طبقه‌بندی می‌کنیم. در غیر این صورت، حاوی داده‌های متفاوتی است و ما نمونه را غیرقابل اعتماد می‌دانیم. مجموعه E در مرحله آخر برگردانده می‌شود که نمونه‌ها را به تفکیک قابل اعتماد و غیرقابل اعتماد بودن بازمی‌گرداند.

یافته‌های پژوهش

در این پژوهش، برای شبیه‌سازی رویکرد معرفی‌شده، از نرم‌افزار متلب به‌علت شمار زیادی از توابع که در آن وجود دارد، استفاده می‌شود. همچنین، سیستم عامل ۶۴ بیتی دارای پردازنده Intel(R)-cori5-2450M, 2.50GHz و نرم‌افزار متلب نسخه ۲۰۱۸ برای اجرای کدها استفاده شده است. در مجموع اطلاعات مالی تعداد ۲۰۴ مشتری فعال در صنعت گردشگری بین سال‌های ۱۳۹۵تا ۱۴۰۰ جمع‌آوری شده بود که برای این مشتریان ۱۹ شاخص مالی استفاده شده است.

در مرحله نخست، برای محاسبه آنتروپی پایه، از طریق رابطه ۱ می‌توان دریافت که هر شاخص تا چه میزان عدم قطعیت دارد. برای محاسبه آنتروپی هر شاخص کافی است تابع توزیع p(f) شاخص را تخمین زده و در رابطه یادشده قرار دهیم. اگر هر شاخص براساس یک رفتار مشخص مقدار پذیرد، می‌توان خروجی متغییر تصادفی را پیش‌بینی کرد.

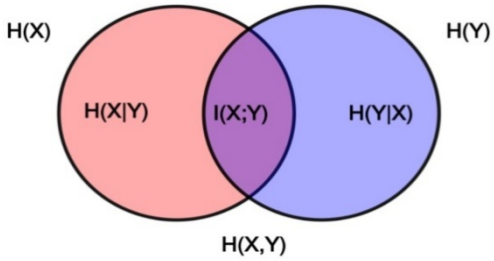
آنتروپی در نهایت به صورت یک عدد به دست می‌آید که می‌تواند مقدار بزرگ یا کوچک داشته باشد. اگر مقدار کم داشته باشد، به این معنا است که عدم قطعیت این شاخص کم بوده و دارای یک رفتار خاص است و یک فرایند تصادفی نیست و به اطلاعات کمتری برای تعیین خروجی متغییر تصادفی نیاز داریم و مقدار بزرگ آن یعنی عدم قطعیت زیاد بوده و به اطلاعات زیادی نیاز خواهیم داشت. نمودار زیر، مقادیر آنتروپی پایه هر یک از شاخص‌ها را در نرم‌افزار متلب نشان می‌دهد:



شکل ۳. مقادیر آنتروپی پایه شاخص‌های اعتبارسنجی مشتریان فعال در صنعت گردشگری

همان‌طور که اشاره شد، مقادیر برزگ آنتروپی به‌معنای عدم قطعیت زیاد است. در نمودار شکل ۳ سطرها نشان‌دهنده شاخص بررسی‌شده و ستون‌ها نشان‌دهنده مقادیر آنتروپی ۲۰۴ مشتری به‌ازای هر شاخص است. شاخص‌های ۳، ۴، ۷، ۱۶، ۱۷، ۱۸، ۱۹ (به‌ترتیب عبارت‌اند از: درآمد حاصل از فروش و ارائه خدمات، نسبت حاشیه سود، نسبت بدهی جاری، سابقه بدهی معوق، داشتن یا نداشتن گزارش حسابرسی، دارا بودن بدهی مالیاتی و روند سودآوری شرکت در سه سال گذشته) از حد آستانه مدنظر که توسط الگوریتم به دست آمده، فراتر رفته و به مقدار زیادی اطلاعات نیاز خواهند داشت.

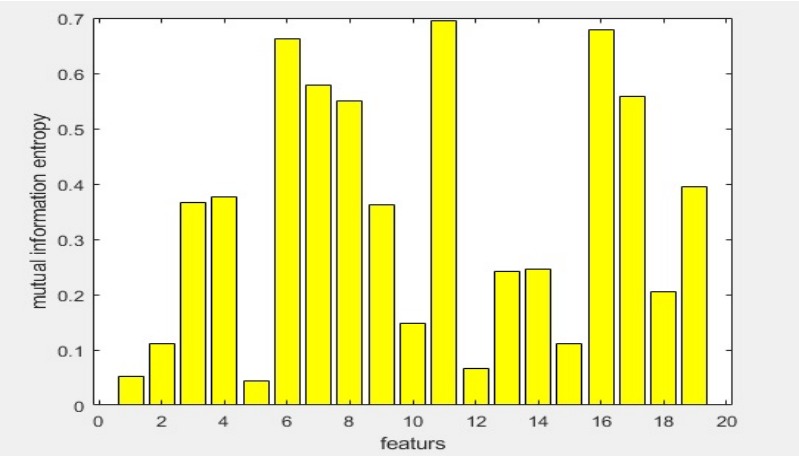
در مرحله دوم، برای به دست آوردن آنتروپی اطلاعات متقابل از آنتروپی شرطی (رابطه ۲) استفاده می‌شود. آنتروپی شرطی به این معناست که اگر اطلاعات خروجی Y را داشته باشیم چه میزان از عدم قطعیت متغیر X باقی می‌ماند (شکل ۴).



شکل ٤. آنتروپی شرطی

آنتروپی اطلاعات متقابل به‌وسیله تابع توزیع تخمین متغیر تصادفی X و تابع توزیع تخمین خروجی Y و هیستوگرام دو متغیره X و Y به دست می‌آید (هیستوگرام درباره شاخص اطلاعاتی می‌دهد که به‌وسیله آن می‌توان فهمید داده‌های هر کلاس چه رفتاری دارند و مقدار شاخص در چه محدوده‌ای است).

در این پژوهش، آنتروپی اطلاعات متقابل میزان اطلاعاتی است که یک شاخص درباره خروجی فراهم می‌کند. در نهایت، هدف، ایجاد مدلی است که با کمک این مدل بتوان خروجی را تخمین زد. از اطلاعات متقابل برای دریافت اطلاعاتی که ویژگی و خروجی درباره یکدیگر می‌دهند، استفاده می‌شود. هرچه اطلاعات متقابل بین آن‌ها بیشتر باشد، یعنی ویژگی اطلاعات بیشتری از خروجی دارد و خروجی را می‌توان راحت‌تر تخمین زد. در فرایند انتخاب ویژگی، اطلاعات متقابل تک‌تک شاخص‌ها با خروجی محاسبه شده (هر یک از شاخص‌های اعتبارسنجی با خوش‌حساب و بد حساب بودن مشتری) و در پایان اولویت با شاخص‌هایی است که اطلاعات بیشتری از خروجی می‌دهند. آنتروپی اطلاعات متقابل در نهایت عددی بین صفر و ۱ است که هرچه به ۱ نزدیک‌تر باشد، نشان‌دهنده ارتباط بیشتر و عدد صفر نشان‌دهنده این است که شاخص و خروجی هیچ ارتباطی به یکدیگر ندارند.



شکل ۵. مقادیر آنتروپی اطلاعات متقابل شاخص‌های اعتبارسنجی مشتریان فعال در صنعت گردشگری

شکل ۵ مقادیر آنتروپی اطلاعات متقابل شاخص‌ها را در نرم‌افزار متلب نشان می‌دهد. همان‌طور که ذکر شد، مقادیر زیاد اطلاعات متقابل نشان‌دهنده ارتباط بیشتر شاخص و خروجی است. در نمودار شکل ۵، سطرها نشان‌دهنده شاخص بررسی‌شده و ستون‌ها نشان‌دهنده مقادیر آنتروپی اطلاعات متقابل ۲۰۴ مشتری به‌ازای هر شاخص است. شاخص‌های شماره ۱، ۲، ۵، ۱۰، ۱۲، ۱۳، ۱۴، ۱۸ (به‌ترتیب عبارت‌اند از: مدت زمان رابطه مشتری با بانک، سرمایه مشتری، نسبت بازده دارایی، نسبت دارایی‌های جاری، تحصیلات مدیرعامل یا مالک، نوع وثیقه (سند رهنی یا سایر وثیقه‌ها)، نوع مالکیت محل فعالیت و دارا بودن بدهی مالیاتی) از حد آستانه مقدار کمتری و در نتیجه ارتباط کمتر با خروجی دارند.

توجه به این نکته در فرایند اعتبارسنجی مشتریان با استفاده از الگوریتم پیشنهادی حائز اهمیت است که براساس ضوابط اجرایی سیاست اعتباری در زمان اعطای تسهیلات و ایجاد تعهدات به‌منظور احراز توان مالی و حسن شهرت متقاضیان، استعلام از سامانه چک‌های برگشتی بانک مرکزی جمهوری اسلامی ایران و همچنین بررسی وضعیت مالیاتی متقاضیان بررسی خواهد شد و در صورت وجود هیچ‌گونه تسهیلاتی به مشتری پرداخت نخواهد شد. از این رو، در مجموعه داده جمع‌آوری‌شده مشتریان مقدار شاخص‌های سابقه چک برگشتی/ بدهی مالیاتی برای کلیه مشتریان مقدار کیفی «ندارد» درج شده است و مقادیر آنتروپی به‌دست‌آمده نشان‌دهنده بی‌اهمیت بودن شاخص‌های بیان‌شده نیست.

مراحل زیر، فرایند اعتبارسنجی با روش آنتروپی پایه و اطلاعات متقابل در نرم‌افزار متلب را نشان می‌دهد:

- تقسیم داده به دو بخش آموزش و تست
- محاسبه آنتروپی شانون هر شاخص و اطلاعات متقابل تمامی ویژگی‌ها با خروجی از طریق داده آموزشی
- انتخاب ویژگی‌های مناسب از طریق مقدار حد آستانه مناسب در آنتروپی شانون و اطلاعات متقابل (در این پژوهش فقط شاخص شماره ۱۸، شاخص دارا بودن بدهی مالیاتی از مجموعه شاخص‌ها حذف شده و بقیه شاخص‌ها با توجه به مقدار حد آستانه بهینه اهمیت داشته و در اعتبارسنجی استفاده شدند).
- حذف شاخص شماره ۱۸ از مجموعه داده مشتریانی که قرار است اعتبارسنجی شوند.
- محاسبه آنتروپی سراسری مجموعه داده اولیه

- اضافه شدن هر عضو مجموعه داده جدید برای اعتبارسنجی به مجموعه داده اولیه و محاسبه مجدد آنتروپی سراسری
- طبقه‌بندی مشتریان جدید با توجه به تغیرات آنتروپی سراسری

معیارهای ارزیابی روش پیشنهادی

در این بخش، معیارهای استفاده‌شده برای رویکرد طبقه‌بندی پیشنهادی شرح داده خواهد شد. معیارهای ارزیابی شامل موارد زیر هستند:

دقت

دقت، واضح‌ترین و ساده‌ترین معیار است که تعداد نمونه‌هایی که به‌درستی طبقه‌بندی شده‌اند و درصد پیش‌بینی‌های صحیحی که هر طبقه‌بندی‌کننده قادر به انجام آن است را نشان می‌دهد و به‌صورت زیر محاسبه می‌شود:

$$Accuracy\ \hat{T} = \frac{|\hat{T}(+)|}{|\hat{T}|}$$ رابطهٔ ۵)

که $|\hat{T}|$ مربوط به تعداد کل نمونه‌ها است و $|\hat{T}(+)|$ مربوط به تعداد نمونه‌هایی است که به‌درستی طبقه‌بندی شده‌اند.

حساسیت

این معیار، تعداد نمونه‌هایی را که به‌درستی به‌عنوان مشتریان قابل اعتماد طبقه‌بندی شده‌اند، اندازه‌گیری می‌کند و اطلاعات مهمی ارائه می‌دهد، زیرا امکان ارزیابی قدرت پیش‌بینی رویکرد را از نظر قابلیت شناسایی موارد نکول فراهم می‌کند و به‌صورت زیر محاسبه می شود:

$$Sensitiviti\hat{T} = \frac{|\hat{T}(TP)|}{|\hat{T}(TP)| + |\hat{T}(FN)|}$$ رابطهٔ ۶)

که در آن $|\hat{T}(TP)|$ مربوط به تعداد نمونه‌هایی است که به‌درستی به‌عنوان قابل اعتماد طبقه‌بندی شده‌اند و $|\hat{T}(FN)|$مربوط به تعداد نمونه‌های قابل اعتمادی است که به‌اشتباه به‌عنوان غیرقابل اعتماد طبقه‌بندی شده‌اند.

F Score

نشان‌دهنده میانگین وزنی معیارهای دقت و یادآوری است و یک معیار عملکرد مؤثر بـرای مجموعـه داده‌های نامتعادل در نظر گرفته می‌شود. چنین متریک به‌صورت زیر محاسبه می‌شود:

$$F - score(T^P, T^R) = 2.\frac{\text{Precision} \,.\, \text{Recall}}{\text{Precision} \;+\; \text{Recal}}$$ رابطهٔ ۷)

$$Precision(T^P, T^R) = \frac{|(T^R \cap T^P|}{|T^P|}$$ رابطهٔ ۸)

$$Recall(T^P, T^R) = \frac{|(T^R \cap T^P|}{|T^R|}$$ رابطهٔ ۹)

که در آن، T^P مجموعه‌ای از طبقه‌بندی‌های انجام‌شده از نمونه‌ها و T^R مجموعه‌ای را که شامل طبقه‌بندی واقعی آن‌ها است، نشان می‌دهد.

ناحیه زیر منحنی (AUC)

این معیار یک معیار عملکرد است که برای ارزیابی اثربخشی یک مدل طبقه‌بندی استفاده می‌شـود و به‌صورت زیر محاسبه می شود.

$$\theta(t_+, t_-) = \begin{cases} 1 & if\ t_+ > t_- \\ 0.5 & if\ t_+ = t_- \\ 0 & if\ t_+ < t_- \end{cases}$$ رابطهٔ ۱۰)

که در آن تمام مقایسه‌های ممکن در نمونه‌های نکول و عدم نکول نشان داده می‌شود.

مقادیر معیارهای ارزیابی این پژوهش در جدول ۲ آورده شده است. این معیارها نشـان مـی‌دهنـد که رویکرد پیشنهادی، حتی بدون استفاده از موارد نکول در مرحله آموزش داده‌ها، نتایج قابل قبـولی دارد.

جدول۲. مقادیر ارزیابی پژوهش

معیار ارزیابی	پژوهش انجام‌شده
$Accuracy$	٪۷۷/ ۶۶۶۷
$Sensitiviti$	٪۸۹/۷۴۳۶
$F - score$	٪۸۶/۳۶۳۶
AUC	٪۸۷/۱۳۴۵

نتیجه‌گیری

تکنیک‌های یادگیری ماشین اعتبارسنجی در بسیاری از زمینه‌های مالی (برای مثال، تسهیلات، بیمه‌نامه‌ها و غیره) نقش مهمی ایفا می‌کنند، زیرا از این رویکردها می‌توان برای ارزیابی ریسک‌های احتمالی فرایند تخصیص تسهیلات استفاده کرد و در نتیجه زیان ناشی از نکول وام مشتریان بدحساب را کاهش داد. با این حال، یکی از مسائل مهمی که در چنین فرایندی یافت می‌شود، مشکل عدم توازن داده در مجموعه‌های داده مشتریان بانک است که در آن تعداد موارد مشتری بدحساب بسیار کمتر از تعداد موارد خوش‌حساب است. این مسائل می‌توانند به‌طور جدی بر عملکرد الگوریتم‌های اعتبارسنجی با هدف طبقه‌بندی مشتریان جدید بانکی تأثیر بگذارند.

این پژوهش، رویکرد جدیید از اعتبارسنجی را پیشنهاد می‌کند که از معیارهای مبتنی بر آنتروپی برای ساختن مدلی که قادر به طبقه‌بندی یک نمونه جدید بدون توجه به مشتریان بدحساب گذشته است، استفاده می‌کند. این رویکرد با مقایسه رفتار آنتروپی مشتریان خوش‌حساب موجود قبل و بعد از افزودن نمونه بررسی‌شده (مشتری جدید متقاضی تسهیلات) کار می‌کند. به این ترتیب، این رویکرد توانسته است به شیوه‌ای پیشگیرانه در خصوص عدم توازن داده‌ها که اثربخشی رویکردهای متعارف اعتبارسنجی را کاهش می‌دهد، عمل کند.

منابع و مآخذ

الف. فارسی

حسین‌زاده کاشان، علی و گروسی، فاطمه (۱۳۹۹). ارائه مدل ترکیبی الگوریتم مورچگان باینری و ماشین بردار پشتیبان (BACO-SVM) برای انتخاب ویژگی و طبقه‌بندی مشتریان بانکی به همراه مطالعه موردی. ***راهبرد مدیریت مالی***، ۸(۲)، ۷۱-۹۲.

فلاح پور، سعید؛ راعی، رضا و هندیجانی زاده، محمد (۱۳۹۳). رویکرد شبکه عصبی مبتنی بر کلونی زنبور عسل مصنوعی. ***مهندسی مالی و مدیریت اوراق بهادار***، ۵(۲۱)، ۳۳-۵۳.

ب. انگلیسی

Batista, G. E., Prati, R. C. & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ***ACM SIGKDD explorations newsletter,*** 6(1), 20-29.

Carta, S., Ferreira, A., Recupero, D. R., Saia, M., & Saia, R. (2020). A combined entropy-based approach for a proactive credit scoring. ***Engineering Applications of Artificial Intelligence,*** 87, 103292.

Cecchetti, S. G., & Schoenholtz, K. L. (2010). How Central Bankers See It: The First Decade of European Central Bank Policy and Beyond. ***In Europe and the Euro*** (pp. 327-374). University of Chicago Press.

Frame, W. S., Srinivasan, A., & Woosley, L. (2001). The effect of credit scoring on small-business lending. ***Journal of money, credit and banking,*** 33(3), 813-825.

Goh, R. Y., & Lee, L. S. (2019). Credit scoring: a review on support vector machines and metaheuristic approaches. ***Advances in Operations Research,*** 1-30. https://doi.org/10.1155/2019/1974794

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. ***Journal of the Royal Statistical Society: Series A (Statistics in Society),*** 160(3), 523-541.

Harris, T. (2015). Credit scoring using the clustered support vector machine. ***Expert Systems with Applications,*** 42(2), 741-750.

Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. ***Expert systems with applications,*** 33(4), 847-856.

Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. ***Decision support systems,*** 37(4), 543-558.

Kiss, F. (2003). Credit scoring processes from a knowledge Management perspective. ***Periodica Polytechnica Social and Management Sciences***, 11(1), 95-110.

Lee, J. Y., Munkhdalai, L., & Ryu, K. H. (2020). A hybrid credit scoring model using neural networks and logistic regression. ***In Advances in Intelligent Information Hiding and Multimedia Signal Processing*** (pp. 251-258). Springer, Singapore.

Rahman, A., Ignatius, J., Hatami-Marbini, A., Dhamotharan, L., & Khoshnevis, P. (2018). A fuzzy decision support system for credit scoring. ***Neural Computing and Applications,*** 29(10), 921-937.

Sharma, A., Vishraj, B., Ahlawat, J., Mittal, T., & Mittal, M. (2020). Impact of COVID-19 outbreak over Medical Tourism. ***IOSR Journal of Dental and Medical Sciences,*** 19(5), 56-58.

Shen, F., Zhao, X., Li, Z., Li, K., & Meng, Z. (2019). A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. ***Physica A: Statistical Mechanics and its Applications,*** 526, 121073.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers**.** ***International journal of forecasting,*** 16(2), 149-172.

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). ***Credit Scoring and its Applications***. SIAM Pub.