مجله علمی-پژوهشی
رایانش نرم و فناوری اطلاعات
جلد ۴، شماره ۳، سال ۱۳۹۴

امتیازبندی رفتاری مشتریان بانک با استفاده از رویکرد داده‌کاوی و فرآیند تحلیل سلسله مراتبی

شهریار محمدی[۱]، روجیار پیرمحمدیانی[۲]
[۱]عضو هیات علمی گروه صنایع دانشگاه صنعتی خواجه نصیرالدین طوسی، mohammadi@kntu.ac.ir
[۲]دانشجوی دکترای فناوری اطلاعات دانشگاه صنعتی خواجه نصیرالدین طوسی، Rpirmohamadiani@kntu.ac.ir

چکیده - در این مقاله با استفاده از رویکرد داده‌کاوی و فرآیند تحلیل سلسله مراتبی، مدلی جهت امتیازبندی رفتاری مشتریان حقیقی به منظور ارزیابی ریسک اعتباری و ایجاد دانش سازمانی در خصوص اعطای تسهیلات اعتباری، ارائه شده است. بدین منظور، در بخش اول یک روش بهبود یافته برای انتخاب متغیرها و آماده‌سازی داده‌ها برای ورود به مدل بکارگرفته شد. در این بخش با استخراج فیلدهای جدید سعی نموده‌ایم، در حد امکان همه‌ی تعاملات مشتری با بانک لحاظ شود. سپس با استفاده از فرآیند تحلیل سلسله مراتبی[۱]، اهمیت داده‌ها مورد ارزیابی قرار گرفت و داده‌های مناسب برای ورود به مدل آماده شد. در فاز مدل‌سازی پژوهش، امتیاز رفتاری مشتریان با توجه به رفتار بازپرداخت مشتریان و میزان دیرکرد آن‌ها به نحو مناسبی تعیین شد. تکنیک‌های داده‌کاوی استفاده شده ترکیبی از روش‌های رده‌بندی و روش‌های متوازن‌سازی می‌باشد. فرآیند داده‌کاوی مربوط به مدل، بر روی مجموعه داده‌های متعلق به موسسه‌ی مالی بخش خصوصی انجام شده است. روش بکاررفته در این پژوهش ضمن داشتن نتایج بهتر نسبت به سایر روش‌ها یک روش پیشنهادی برای بانک‌ها خواهد بود تا با بهره‌گیری از تحلیل اطلاعات موجود، مشتریان خود را بهتر و دقیق‌تر شناسایی اعتباری نموده و نسبت به گذشته، ریسک اعتباری خود را کاهش دهند.

کلید واژه‌ها-ریسک اعتباری، امتیازبندی رفتاری، داده‌کاوی، رده‌بندی.

۱. مقدمه

موسسات مالی و بانک‌ها از سازمان‌هایی هستند که به لحاظ ماهیت کار خود بسیار نیازمند فرآیندهای مدیریت ریسک می‌باشند [1]. در این میان ریسک اعتباری که در رابطه با مشتریان می‌باشد از اهمیت ویژه‌ای برخوردار است و مدیران باید راهکار مناسبی برای ارزیابی و شناسایی ریسک مشتریان، ارائه کنند. تا از این طریق، تخصیص کارآمد تسهیلات اعتباری ممکن شود [3]. در واقع مدیریت ریسک اعتباری، همان مفاهیم مطرح شده در فاز اول مدیریت ارتباط با مشتری، در عرصه‌ی بانکداری و امور مالی است. فاز اولیه‌ی فرآیند مدیریت ارتباط با مشتری، شناسایی مشتریان می‌باشد که شامل دسته‌بندی مشتریان و تجزیه و تحلیل بازار هدف است [4]. در بحث مدیریت ریسک اعتباری نیز، شناسایی مشتریان و تفکیک آن‌ها به گروه‌های خوش‌حساب و بدحساب و تخمین ریسک مرتبط به هر یک، از اهمیت ویژه‌ای برخوردار است [5].

با توجه به حجم عظیم داده‌های مشتریان، داده‌کاوی می‌تواند جهت دسته‌بندی و پیش‌بینی رفتار مشتریان مورد استفاده قرار گیرد. در واقع امروزه داده‌کاوی از جمله مهمترین ابزارهای در دسترس این سازمان‌ها برای ارزیابی و شناسایی ریسک‌های متنوع پیشرو است [6]. تحقیقات متنوعی برای اعتبارسنجی مشتریان بانک‌ها و موسسات مالی، با استفاده از مدل‌های پارامتریک و ناپارامتریک، جهت امتیازبندی مشتریان صورت گرفته است. در یک تقسیم بندی، دو روش کلی وجود دارد: (۱) امتیازبندی اعتباری[۲] جهت شناسایی مشتریان، در خصوص متقاضیان و مشتریان جدید و (۲) امتیازبندی اعتباری جهت مدیریت مشتریان موجود، که در این صورت امتیازبندی رفتاری[۳] نامیده می شود [7]. در واقع بهبود مدل‌های امتیازبندی یک مزیت رقابتی برای بانک‌ها محسوب می‌شود که امکان اتخاذ تصمیمات پیشگیرانه را میسر می‌سازد [8].

اکثر روش‌های امتیازبندی اعتباری و رفتاری موجود، گویای احتمال قصور در یک دوره‌ی مشخص است و برای کمی‌سازی این احتمال، به تفکیک مشتریان به دو گروه خوش‌حساب و بدحساب اکتفا نموده‌اند. چنین مدل‌هایی ممکن است از لحاظ فیلتر کردن ریسک‌های بالاتر مثمرثمر باشد ولی هچنان احتمال زیان بالا است.

این مسائل مورد توجه کمیته‌ی بال؛کمیته نظارت بر مقررات بانکی[۴]؛ هم بوده است و این کمیته اقدام به تعریف توصیه‌نامه‌هایی در این حوزه نموده است. این کمیته به

بانک‌ها توصیه کرده است تا از رویکرد امتیازبندی داخلی[۵] جهت محاسبه ریسک اعتباری استفاده نمایند [9]. همچنین علاوه بر راهکارهای اولیه برای ارزیابی ریسک بر محاسبه‌ی اطلاعات تکمیلی، نظیر نرخ برگشت، تخمین زمان قصور و غیره تاکید کرده است.

به همین دلیل، اهمیت این نوع سیستم‌های امتیازبندی داخلی در بانک‌ها خیلی بیشتر از سابق شده است. ولی برای محاسبه چنین پارامترهایی لازم است که روش‌های امتیازبندی رفتاری و اعتباری توسعه داده شوند بنابراین لحاظ کردن روش بهبود یافته برای انتخاب متغیرها، آماده‌سازی داده‌ها و تخمین معیارهایی مناسب‌تر برای ایجاد دید همه جانبه نسبت به مشتری، خصوصا مشتریان حقیقی الزامی است.

در این مقاله سعی شده است با لحاظ کردن قواعد خاص بانک‌های داخلی، یک سیستم امتیازبندی داخلی را در قالب روش امتیازبندی رفتاری، برای مشتریان حقیقی موسسه‌ی مالی مورد نظر ارائه دهیم. به این ترتیب یک سیستم رده‌بندی که بتواند نوع مشتریان را شناسایی کند، لحاظ شده است.

در این تحقیق به منظور ارزیابی ریسک اعتباری مشتریان، ضمن جمع‌آوری و نظام‌مند کردن نظر و معیار خبرگان، ارتباطات موجود در مجموعه داده‌ها بررسی شده و عوامل موثر و میزان اهمیت آن‌ها شناسایی شده است. سپس برای امتیازبندی مشتریان، با لحاظ کردن سیاست های اعتباری بانک مورد بررسی، مطابق نظر کارشناسان با توجه به دیرکردها و رفتارهای بازپرداختی، ۵ رده برای انواع مختلف مشتری تعریف شده است. سیستم رده‌بندی تعریف شده با توجه به پارامترهای لحاظ شده تشخیص می‌دهد که هر مشتری به کدامیک از دسته ها تعلق دارد. راهکار ارائه شده بر اساس تلفیق و ترکیب فرآیند تحلیل سلسله مراتبی و الگوریتم‌های داده‌کاوی و متوازن‌سازی صورت می‌گیرد. فیلدهای لحاظ شده برای مدل امتیازبندی پیشنهادی بر اساس داده‌های موجود در پایگاه داده‌ی مورد مطالعه می‌باشد و سعی شده است بر اساس حداقل پارامترهای موجود، مدل با کارایی مناسب پیشنهاد شود.

در ادامه در بخش ۲، برخی از پژوهش‌هایی که در این زمینه انجام شده و با موضوع تحقیق مرتبط بوده، در قالب یک دسته‌بندی مورد بررسی قرار گرفته است. دربخش ۳، چارچوب پیشنهادی سیستم امتیازبندی رفتاری ارائه شده است. در بخش ۴ نیز مراحل اجرایی تحقیق و نتایج بدست آمده بیان شده است.

۲. مرور کارهای انجام شده

توجه به عواملی که منجر به افزایش ریسک اعتباری بانک‌ها و موسسات مالی می‌گردد، از اهمیت خاصی برخورداراست. یکی از اجزء اصلی ریسک در حوزه ریسک اعتباری، احتمال نکول[۶] می‌باشد. احتمال نکول این است که طرف قرارداد، به تمام یا بخشی از تعهداتش، خواسته یا ناخواسته عمل نکند [10]. این احتمال نکول نهایتا با در نظر گرفتن یک آستانه[۷] مشتریان بانک را به دو دسته خوب و بد تقسیم می‌نماید.

با توجه به تفکیک مشتریان به گروه‌های خوش‌حساب و بدحساب، مسائل اعتبارسنجی در حوزه‌ی مسائل رده‌بندی داده‌کاوی قرار می‌گیرد. رده‌بندهای مختلفی تا کنون ارایه شده است که از آن جمله می‌توان به شبکه‌های عصبی، ماشبن بردار پشتیبان، انواع درخت‌ها و .... اشاره کرد. در جدول ۱ مجموعه‌ای از تکنیک‌های رده‌بندی که در حوزه‌ی رتبه‌بندی اعتباری به کار برده می‌شوند، نمایش داده شده است.

جدول ۱: تکنیک‌های رده‌بندی استفاده شده برای امتیازبندی اعتباری مشتریان بانک

تکنیک‌های کلاس‌بندی	منابع استفاده کننده از این تکنیک‌ها در حوزه‌ی رتبه‌بندی اعتباری
رگرسیون لجستیک و تجزیه و تحلیل ممیزی خطی	Baesens et al.(2003)[11]
	Leea et al. 2006[16]
	Hayhoe, C. R., Leach, L., & Turner, P. R. (1999)[17]
شبکه‌های عصبی	West(2000)[18]
	Tsai, C.F., Wu, J.W(2008)[19]
	Baesens et al. (2005)[20]
	Hsieh(2005)[15]
ماشبن بردار پشتیبان	Wang, G., Hao, J., Ma, J., Jiang(2011)[21]
	Chen, F., Li, F.:(2010)[14]
	TunLi, S., Shiue, W., Huang(2006)[22]
	Huang, C.L., Chen, M.C., Wang, C.J (2007)[23]
درخت تصمیم (C4.5 CART, etc)	Arminger G, Enache D, and Bonne T (1997)[24].
	Baesens et al.(2003)[11]

البته به منظور افزایش دقت، روش‌های انتخاب مجموعه متغیرهای بهینه‌ی ورودی و متوازن‌سازی[8] نیز جهت تحلیل رفتار مشتریان، به میزان زیادی کاربرد دارند. در ادامه به توضیح هر کدام از این گروه مقالات می پردازیم.

در یک مدل داده‌کاوی، هر چقدر معیارهای مورد استفاده بهتر باشد، خروجی سیستم دقیق‌تر بوده و با واقعیت منطبق‌تر می‌باشد [2]. در برخی از مقالات جهت بهبود جواب‌های بدست آمده، متغیرهای جدیدی تعریف و بکارگرفته شده‌اند. به عنوان مثال در تحقیقی که به منظور پیش‌بینی نکول وام مشتریان حقیقی بانک تایوانی صورت گرفته است، متغیرهای جدیدی جهت مدل‌سازی لحاظ شده است. در این مقاله از دو دسته متغیر پیش‌بینی استفاده شده است: (۱) متغیرهای جمعیت-شناسی[9] و (۲) متغیرهای نگرش مالی[10]. متغیرهای جمعیت شناختی شامل ۵ متغیر: سن، جنس، شغل، تحصیلات و درآمد ماهیانه می‌باشد. متغیرهای لحاظ شده برای نگرش مالی نیز شامل ۹ ویژگی: سابقه مشتری در بانک، وضعیت پس انداز، مدت زمان نگهداری بودجه واعتبار، بی‌تفاوت نسبت به قیمت، علاقه مندی به سرمایه‌گذاری و آزمایش خدمات جدید، میزان وفاداری به بانک، نسبت اعتبار به توان مالی، برنامه‌های پرداختی دیگر و هدف وام می‌باشد. داده‌های مورد نیاز از طریق اطلاعات موجود در پایگاه داده‌ی بانک گردآوری شده است، همچنین از طریق پرسش‌نامه اطلاعات مورد نیاز در مورد نگرش اشخاص نسبت به پول جمع‌آوری شده است. با بکارگیری مدل‌های پیش‌بینی بر روی داده‌های نمونه به این نتیجه رسیده است که با درنظر گرفتن این دو دسته فاکتور در کنار هم به عنوان مشخصات ورودی برای مدل مورد نظر، صحت پیش‌بینی وام نسبت به حالتی که فقط از مشخصات افراد استفاده کنیم، افزایش می یابد [12].

برخی از مقالات نیز با استفاده از روش انتخاب ویژگی[11] تعداد ورودی‌ها را کاهش می‌دهند. در تحقیقی که برای امتیازبندی اعتباری مشتریان با استفاده از تکنیک ماشین بردار پشتیبان و بر روی مجموعه داده‌های آلمان و استرالیا در بانک اطلاعاتی UCI[12] انجام شده است، برای اینکه از ماشین بردار پشتیبان جواب بهتری بگیرد دو مورد زیر را لحاظ کرده است.

۱. تعیین متغیرهای ورودی

۲. تنظیم بهینه‌ی پارامترهای ورودی مدل ماشین بردار پشتیبان

درواقع این مقالات به ترکیب ماشین بردار پشتیبان با مکانیزم‌هایی پرداخته‌اند که پارامترها و ویژگی‌های مدل را بهینه می‌نمایند و از این طریق، عملکرد و صحت مدل را بالا می-برد[13][14].

داده‌های رتبه‌بندی اعتباری اصولا داده‌های نامتوازنی هستند. چرا که عمدتا تعداد افراد یا شرکت‌هایی که در بازپرداخت دیون خود دچار نکول می‌شوند بسیار کمتر از تعدادی است که آن‌ها را به موقع و به صورت مناسبی بازپرداخت می-نمایند. این نسبت بسته به نوع وام‌ها از یک به ۱۰ الی ۱۰۰ گزارش شده است. در ادامه به مقالاتی که به مسئله‌ی عدم‌توازن در داده‌های رتبه‌بندی اعتباری پرداخته‌اند اشاره شده است.

برون و مورس به مسئله عدم توازن در رتبه‌بندی اعتباری پرداخته‌اند [26]. در این مقاله رده‌بندهای مختلف روی ۵ پایگاه داده‌ی مختلف، که با شیوه‌های مختلف نمونه‌برداری، توازن آن‌ها تغییر یافته است آزمایش شده است. نتایج تحقیق حاکی از آن است که در حالی که پایگاه‌های داده با انجام undersamling به یک حالت توازنی رسیده‌اند برای مثال در نسبت بد ۳۰٪، روش LS-SVM نتایج بهتری را ارایه نموده است. لیکن هر چه به سمت عدم توازن بیشتری حرکت می‌کنیم دو روش random forest و gradiant boosting نتایج بهتری ارایه می نماید.

کرون و فینلی، تاثیرات اندازه نمونه و توازن را در مدل-های اعتبارسنجی مورد بررسی قرار داده‌اند[27]. در این مقاله یک نوآوری کاربردی برای ساخت مدل‌های رتبه‌بندی اعتباری ارایه شده‌است و نتیجه می‌گیرند که تخمین حداکثری بهتر از حداقلی می‌باشد و هم‌چنین با بالا رفتن تعداد نمونه‌ها از بازه ۳۰۰۰ الی ۴۰۰۰، همچنان به صورت معناداری جواب درصد صحت مدل‌ها بالاتر می‌رود.

به طور کلی یافته‌های حاصل از مرور کارهای انجام شده به صورت زیر می‌باشد:

۱. عمده تحقیقات گذشته روی رتبه‌بندی اعتباری صورت گرفته‌اند و حجم کمتری از تحقیقات روی رتبه‌بندی رفتاری متمرکز شده است.

۲. مشکل برای انتخاب بهترین ویژگی‌های ورودی و پارامترهای کرنل در روش ماشین بردار پشتیبان. در واقع اکثر روش‌های به کاربرده شده برای انتخاب مشخصه‌های ورودی تنها مبتنی بر تکنیک های پایگاه داده بوده‌اند این در حالی است که بسیاری از بانک‌ها، سیاست‌های اعتباری خود را در قالب قواعد اجرا می‌کنند لذا توجه بیشتر به این قواعد منجر به نتایج بهتر خواهد شد

۳. در حالی که بخش زیادی از اطلاعات اعتباری مشتری ساختار نامتوازن دارند، ولی متوازن‌سازی داده در مقالات بررسی شده کمتر مورد توجه قرار گرفته است.

از این رو در این مقاله سعی شده چارچوبی برای بهبود روش های امتیازبندی رفتاری ارائه شود که با بکارگیری تکنیک-های مختلف، یک سیستم امتیازبندی داخلی کارا را برای بانک‌ها فراهم آورد.

۳. چارچوب پیشنهادی تحقیق

در این قسمت، یک مدل امتیازبندی رفتاری جهت تخصیص بهینه تسهیلات اعتباری در یک موسسه‌ی مالی بخش خصوصی ارائه می‌شود. چهارچوب کاربردی پیشنهادی، ترکیبی از چند روش داده‌کاوی و تکنیک تحلیل سلسله مراتبی می‌باشد. همان‌طور که قبلاً ذکر شد شناسایی و اولویت‌بندی معیارها برای گروه‌های مختلف مشتریان، گام اصلی و ابتدایی در امتیازبندی اعتباری مشتریان محسوب می‌شود. اگرچه در مقالات بررسی شده‌ی قبلی نیز یک سری روش‌های انتخاب و آماده‌سازی ویژگی‌ها به کار برده شده‌اند اما این روش‌ها تنها از دیدگاه علم پایگاه داده انتخاب می‌شوند. لیکن اطلاعات دامنه می‌تواند به عنوان یک منبع ارزشمند جهت آماده‌سازی و بهینه‌سازی ویژگی‌ها مورد استفاده قرار گیرد. از این رو در این مقاله سعی شده است با ترکیب انتخاب ویژگی‌ها با تکنیک تحلیل سلسله مراتبی، ضمن جمع‌آوری و نظام‌مند کردن نظر و معیار خبرگان، داده‌های مناسب‌تر برای ورود به مدل آماده شوند. به منظور امتیازبندی رفتاری مشتریان ۵ رده‌ی اصلی تعریف شده است.

همان‌طور که قبلا بحث شد اکثر مطالعات قبلی مشتریان را تنها به دو گروه خوش‌حساب و بدحساب تقسیم نموده‌اند که در آن مشتری بد حساب، مشتریانی هستند که بیش از ۹۰ روز از هر گونه تعهد اعتباری او سپری شده باشد این تقسیم بندی خیلی کلی بوده و بازه‌ها و حالت‌های مختلف دیرکرد در آن لحاظ نشده است. حال آنکه بانک‌ها بایستی با توجه به سیاست‌های داخلی خود تقسیم‌بندی مشتریان را با دقت بیشتری انجام دهند. در این صورت آن‌ها قادر خواهند بود با توجه به گروه رفتاری که مشتری به آن تعلق دارد، طیف مختلفی از استراتژی تعامل با مشتری را به کار گیرند. از این رو در این مقاله با لحاظ کردن سیاست های اعتباری بانک مورد بررسی، مطابق نظر کارشناسان با توجه به دیرکردها و رفتارهای بازپرداختی، ۵ رده برای انواع مختلف مشتری تعریف شده است. برای این منظور ابتدا با توجه به پارامتر میزان دیرکرد در پرداخت، اقساط به پنج گروه اقساط بدون تاخیر، اقساط جاری، اقساط با تاخیر معمولی، اقساط سررسیدگذشته و اقساط معوق طبقه‌بندی می‌گردند. پارامتر میزان دیرکرد به صورت زیر در طبقه‌بندی دخیل است :

اقساط بدون تاخیر: در این صورت پرداخت اقساط در سررسید مربوطه و یا زودتر از سررسید صورت پذیرفته است.

اقساط جاری: چنانچه تاخیر در پرداخت قسط از تاریخ سررسید قسط، کمتر از ۲ ماه باشد قسط مزبور از لحاظ پارامتر زمان، شرایط جاری بودن را دارا می‌باشد. مبالغ مربوط به اقساط واجد این شرایط، جز سرفصل تسهیلات به حساب می‌آیند. در این شرایط در صورتحساب وام تاخیر محاسبه و برای مشتری جمع تاخیرها لحاظ می‌شود.

اقساط با تاخیر معمولی: چنانچه تاخیر در پرداخت قسط از تاریخ سررسید بیشتر از ۲ ماه و کمتر از ۳ ماه باشد، قسط یا اقساطی که واجد این شرایط باشد به گروه اقساط با تاخیر معمولی منتقل خواهد شد. در صورت احراز شرایط زمانی ذکر شده، برای وام گیرنده و ضامن اخطاریه صادر می‌شود ولی مبالغ مربوط به این نوع از اقساط، جز سرفصل تسهیلات می‌باشد.

اقساط سررسید گذشته: چنانچه تاخیر در پرداخت قسط از تاریخ سررسید بیشتر از ۳ ماه وکمتر از ۴ ماه باشد، قسط یا اقساطی که واجد این شرایط باشد به گروه اقساط سررسید گذشته منتقل خواهد شد. در صورت احراز شرایط زمانی ذکر شده از تاریخ سررسید اقساط، مبلغ اقساط از سرفصل تسهیلات خارج و به گروه اقساط سررسید گذشته منتقل خواهد شد.

اقساط معوق: چنانچه تاخیر در پرداخت قسط از تاریخ سررسید بیشتر از ۴ ماه باشد، قسط یا اقساطی که واجد این شرایط باشد به گروه تسهیلات سررسید گذشته منتقل خواهد شد. در صورت احراز شرایط زمانی ذکر شده از تاریخ سررسیداقساط، مبلغ اقساط از سرفصل تسهیلات خارج و به گروه تسهیلات معوق منتقل خواهد شد .

بر اساس طبقه‌بندی صورت گرفته برای اقساط، با در نظر گرفتن آستانه‌ای، ۵ رده‌ی اصلی و مهم جهت شناسایی ریسک اعتباری مشتریان تعریف شده است. نحوه‌ی رده‌بندی و شرایط هر رده به شرح ذیل می‌باشد:

رده‌ی اول: چنانچه مشتری تمام اقساط خود را بدون تاخیر پرداخت کرده باشد و یا حداکثر ۲ قسط از نوع جاری داشته باشد آن مشتری جز رده‌ی اول محسوب می‌شود. رده‌ی اول بهترین کیفیت را از نظر بازپرداخت تسهیلات دارا است و از

کمترین ریسک برخوردار است.

رده‌ی دوم: چنانچه مشتری حداقل ۳ قسط از نوع جاری و یا حداکثر ۲ قسط از نوع تاخیر معمولی را داشته‌باشد. در صورت احراز این شرایط مشتری جز رده‌ی دوم محسوب می‌شود. رده‌ی دوم از لحاظ بازپرداخت از شرایط مطلوبی برخوردارند و فرق آن‌ها با گروه قبلی این است که حاشیه امنیتشان به گستردگی گروه قبلی نیست و ریسک بلندمدت آن‌ها مقداری بیشتر است.

رده‌ی سوم: چنانچه مشتری حداقل ۳ قسط از نوع تاخیر معمولی و یا حداکثر ۲ قسط از نوع سررسید گذشته را داشته باشد. در صورت احراز این شرایط مشتری جز رده‌ی سوم محسوب می‌شود.گروه سوم از درجه متوسط محسوب می‌شوند و در حال حاضر از نظر بازپرداخت مشکلی ندارند اما در درازمدت ممکن است دچار مشکل گردند.

رده‌ی چهارم: چنانچه مشتری حداقل ۳ قسط از نوع سررسید گذشته و یا حداکثر ۲ قسط از نوع معوق را داشته باشد. در صورت احراز این شرایط مشتری جز رده‌ی چهارم محسوب می‌شود. گروه چهارم از نظر تضمین بازپرداخت تسهیلات در شرایط مطلوبی قرار ندارند و از مشخصه‌های یک متقاضی خوب برخوردار نیستند.

رده‌ی پنجم: چنانچه مشتری حداقل ۳ قسط از نوع معوق داشته‌باشد. در صورت احراز این شرایط مشتری جز رده‌ی پنجم محسوب می‌شود. گروه پنجم از نظر بازپرداخت اصل و سود تسهیلات از تضمین خوبی چه در حال حاضر و چه در آینده برخوردار نیستند، این گروه یا در حالت نکول قرار دارند و یا اینکه بازپرداخت اقساط آن‌ها با خطرات زیادی روبه روست.

لازم به ذکر است که تقسیم‌بندی مشتریان به نحوی که ارائه شد، با توجه به دستورالعمل بانک و مطابق نظر کارشناسان و خبرگان صورت پذیرفته است. مراحل اصلی تحقیق جهت تخصیص بهینه تسهیلات اعتباری به صورت یک چارچوب کلی در شکل ۱ مشخص شده است. در ادامه به توضیح هر یک از این مراحل می‌پردازیم.

۳.۱. شناخت عوامل تاثیر گذار بر امتیاز رفتاری مشتریان حقیقی

برای ایجاد نگرش جامع و درک کامل رفتار مشتری با بانک، بایستی کلیه‌ی تراکنش‌ها و داده‌های جدا از هم مرتبط با مشتریان را در نظر گرفت. بنابراین با لحاظ کردن کلیه‌ی داده و اطلاعات جمع‌آوری شده موجود در پایگاه داده‌ی بانک و با در نظر گرفتن سیاست‌ها و برنامه‌های راهبردی بانک در اعطای تسهیلات، زیرمجموعه‌ای از مشخصه‌ها ورودی را بر اساس اهمیت آن‌ها استخراج می‌کنیم. فرض لحاظ شده برای این تحقیق بر این اساس است که مدل پیش‌بینی ما برای آن دسته از مشتریان بانک که قبلا سابقه‌ی دریافت تسهیلات اعتباری را نداشته‌اند هم قابل اعمال باشد. به همین خاطر متغیرهای انتخاب شده باید بین هر دو گروه مشتریان مشترک باشد. بر این اساس فیلدهای اطلاعاتی موجود در جداول مشتری، حساب و تراکنش در نظر گرفته شده است و معیارها تعیین شده‌اند.

بر این اساس تعداد ۹ متغیر اصلی شامل: سن، جنس، شغل، کارکرد حساب ۱۲ ماه قبل از اخذ وام (بر اساس میانگین ریال/روز)، موجودی اولیه، تعداد حساب های فعال مشتری، میانگین زمانی بین مراجعات، سابقه‌ی مشتری در بانک، میانگین موجودی سایر حساب‌ها به عنوان متغیرهای اثرگذار بر امتیاز اعتباری مشتریان حقیقی تعیین شد. در این فرآیند عوامل موثر بر اعتبار فرد مطابق نظر کارشناسان اعتباری به سه دسته: مشخصات شخصی مشتری، اطلاعات مربوط به حسابی که مشتری بر اساس آن وام گرفته و سطح فعالیت مشتریان تقسیم شدند.

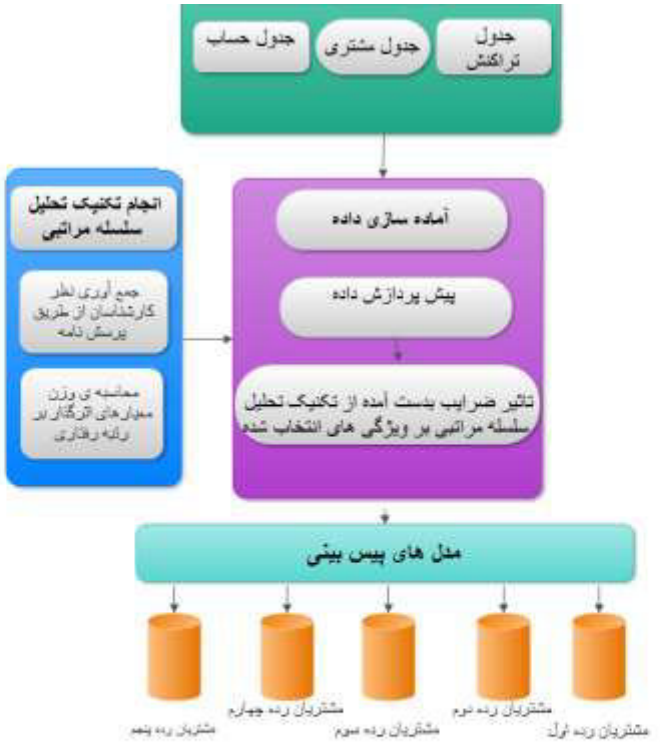
۳.۲. آماده‌سازی داده

بعد از مشخص شدن عوامل اثرگذار بر امتیاز رفتاری مشتریان در فاز قبلی، در این مرحله با توجه به داده‌های موجود، فیلدهای مورد نظر استخراج می‌شوند. وظایف آماده‌سازی در دو مرحله انجام می‌گیرد. مرحله‌ی اول پیش‌پردازش و مرحله‌ی دوم تاثیردادن ضرایب بدست آمده از فرآیند تحلیل سلسله مراتبی می‌باشد.

مرحله‌ی اول، پیش‌پردازش؛ شامل یکپارچه‌سازی داده‌ها، پاکسازی داده‌ها، شکل دادن و ایجاد فیلدهای جدید، همچنین انتقال داده برای ابزار مدل سازی می‌باشد. یکی از مراحل مهم این گام نرمال‌سازی است. مقادیر موجود در مجموعه داده شامل داده‌های اسمی و مقادیر عددی می‌باشد. روش‌های نرمال سازی هر یک، بر روی نوع خاصی از مقادیر اعمال می‌شوند. برای مقادیر عددی از فرمول نرمال‌سازی Min_Max استفاده می‌شود که در معادله‌ی (۱) نمایش داده شده است.

$$\mathrm{Value}_{\mathrm{New}} = \frac{\mathrm{Value}_{\mathrm{Old}} - \mathrm{Min}_A}{\mathrm{Max}_A - \mathrm{Min}_A} \ (New\_Max_A - New\_Min_A) + New\_Min_A$$

شکل ۱- چارچوب کلی مدل امتیازبندی رفتاری پیشنهادی



که در آن $Value_{New}$ مقدار نرمال شده برای مقدار $Value_{Old}$ می‌باشد، Min_A و Max_A به ترتیب کوچک ترین و بزرگ ترین مقدار برای ویژگی A می‌باشند. و New_Max_A و New_Min_A بزرگترین و کوچک‌ترین مقادیر بازه‌ی جدید هستند.

در مورد مقادیر اسمی روش نرمال‌سازی مخصوصی وجود ندارد و می‌بایست آن‌ها را به مقادیر باینری تبدیل نمود و سپس با استفاده از روش‌های نرمال سازی رایج آن را نرمال نمود.

مرحله‌ی دوم، تاثیردهی ضرایب بدست آمده از فرآیند تحلیل سلسله مراتبی؛ دانش و اطلاعات مسئله و نظر تصمیم-گیران می‌تواند ویژگی‌ها و داده‌های بهتر و با اهمیت‌تری برای الگوریتم کلاس‌بندی و مدل تامین نماید. بر این اساس در این پژوهش با ترکیب انتخاب ویژگی‌ها با تکنیک فرآیند تحلیل سلسله مراتبی، یک رویکرد پیشنهادی جهت آماده‌سازی داده‌های نهایی، ارائه گردیده است. چارچوب کلی رویکرد پیشنهادی جهت آماده‌سازی داده‌های نهایی در شکل ۲ نشان داده شده است . پس از آن‌که با در نظر گرفتن دانش مسئله و نظر تصمیم‌گیران معیارها و زیرمعیارهای اثرگذار بر رتبه‌ی اعتباری مشتریان مشخص شد، گام بعدی ارزیابی معیارهای موجود با استفاده از مقایسه‌ی دوبه‌دوی آن‌ها می‌باشد. برای مقایسه‌ی معیارها در این تحقیق از پرسش‌نامه‌هایی مبتنی بر قضاوت‌های عددی استفاده شده است و از این طریق نظر کارشناسان جمع‌آوری شده است. در گام بعدی با توجه به ماتریس مقایسات بدست آمده، میزان اهمیت هر معیار تعیین گردیده است [25].

شکل ۲. نهادی برای استفاده از تکنیک تحلیل سلسله مراتبی جهت آماده-سازی داده

در واقع اهمیت معیارهای مشخص شده در مرحله‌ی۳ اندازه‌گیری می‌شود. و این ضرایب در مقدار فیلدها یا ویژگی‌ها ضرب شده و داده‌ها جهت ورود به مدل آماده می شوند.

با توجه به داه‌های نرمال شده، نحوه‌ی تاثیر ضرایب بدست آمده از فرآیند تحلیل سلسله مراتبی بر روی داده‌ها به شرح زیر خواهد بود:

اگر فیلد مورد نظر پیوسته و عددی باشد از معادله‌های (۲) و (۳) استفاده می‌شود:

$$\text{Value}_{Norm} = \frac{\text{Value}_{Old} - \text{Min}_A}{\text{Max}_A - \text{Min}_A} \quad (۲)$$

$$\text{Value}_{New} = \text{Value}_{Norm} * \text{weight}_A \quad (۳)$$

که در آن weight_A، وزن محاسبه شده برای ویژگی A می‌باشد که با توجه به رویکرد فرآیند تحلیل سلسله مراتبی، محاسبه شده است.

۲- اگر فیلد مورد نظراسمی[۱۳] باشد، همان‌طور که قبلا بیان شد، صفات اسمی را به صورت باینری نمایش می‌دهیم در این صورت برای تاثیر ضرایب فرآیند تحلیل سلسله مراتبی به شرح زیر عمل می‌نماییم:

با توجه به مقدار فیلد، اگر مقدار فیلد ۱ باشد، خصیصه‌ی باینری معادل آن برابر weight_A و در غیر این صورت صفر در نظر گرفته می‌شود. که در آن weight_A، وزن محاسبه شده برای ویژگی A می‌باشد که با توجه به رویکرد فرآیند تحلیل سلسله مراتبی، محاسبه شده است.

۳.۳. مدل پیش بینی

متغیر وابسته تحقیق عبارت است از: امتیاز رفتاری مشتریان. همان‌طور که قبلا اشاره شد، بر خلاف تحقیقات پیشین که برای تقسیم‌بندی مشتریان، به دو گروه خوش‌حساب و بد-حساب اکتفا می‌کنند، در این تحقیق به منظور شناخت دقیق‌تر مشتریان، ۵ رده‌ی اصلی و مهم جهت شناسایی ریسک اعتباری مشتریان تعریف شده است. که رده‌ی اول بهترین کیفیت را از نظر بازپرداخت تسهیلات دارا است و از کمترین ریسک برخوردار است و رده‌ی پنجم از نظر بازپرداخت اصل و سود تسهیلات از تضمین خوبی چه در حال حاضر و چه در آینده برخوردار نیستند و بازپرداخت اقساط آن‌ها با خطرات زیادی روبه‌روست و گروه-های دوم و چهارم و سوم رفتار ما بین این دو رده را دارند. در این بین به دلیل افزایش تعداد رده‌های تعریف شده، ممکن است مشکل عدم توازن داده‌ها بیشتر رخ بنماید. مجموعه داده‌ی مورد استفاده در این تحقیق نیز نامتوازن است به طوری که فقط ۴ درصد افراد از رده‌ی سوم می‌باشند و اکثر نمونه‌ها، نزدیک ۴۳ درصد متعلق به رده‌ی دوم می‌باشند. این در حالی است که اغلب الگوریتم‌های رده‌بندی، فرض می‌کنند که توزیع رده‌ها یکسان است و در صورتی که داده‌ها نامتوازن باشند، این الگوریتم‌ها در تشخیص خود دچار مشکل می‌شوند و رده‌ای که تعداد داده‌های آن بسیار کم است. اغلب نادیده انگاشته می‌شود. از این رو جهت کارایی بهتر الگوریتم‌های رده‌بندی در فاز پیش‌بینی باید از روش‌های متوازن‌سازی استفاده شود. هم‌چنین تمام روش‌های ارزیابی الگوریتم‌های رده‌بندی، برای مسائل دو رده‌ای طراحی شده‌اند، حال آنکه با توجه به چند رده‌ای بودن نمونه‌ی مورد بررسی باید از معیارهای ارزیابی مسائل چند رده‌ای استفاده شود.

روش متوازن‌سازی مورد استفاده در این تحقیق ModifiedBagging می‌باشد که مبتنی بر زیرنمونه‌برداری است. در این الگوریتم، زیرنمونه‌هایی از رده‌های اکثریت تولید می‌شود. سپس هر یک از این زیرنمونه‌ها با تمام نمونه‌های رده‌ی اقلیت ترکیب شده و مجموعه‌ای برای آموزش رده‌بند تولید می‌کنند در نهایت نیز تمام رده‌بندها با هم ترکیب شده و مدل نهایی را تولید می‌کنند. شبه‌کد موجود در تصویر۳، الگوریتم ModifiedBagging را معرفی می‌کند[28].

{Input:

A set of minority class examples S_{min}

A set of majority class examples S_{maj} $|S_{min}|<|S_{maj}|$

T, the number of subsets to sample from S_{maj}}

Method:

$i \leftarrow 0$

repeat

$i \leftarrow i+1$

Randomly sample a subset E_i from S_{maj}, $|E_i| = |S_{min}|$,

$S_i = S_{min} \cup E_i$

Learn C_i on S_i. C_i is a simple classifier

until $i = T$

Output: An ensemble $\{H_i | 1 \le i \le T\}$

شکل ۳: شبه‌کد الگوریتم ModifiedBagging.[28]

جهت رده‌بندی از یک الگوریتم رده‌بندی استفاده می‌شود که با توجه به مجموعه داده صحت بهتری داشته باشد، همچنین جهت ترکیب این رده‌بندها نوع نمونه‌گیری مبتنی بر روش-bagging می‌باشد. پس از ساخت مدل، ما باید بتوانیم میزان دقت مدل پیشنهادی خود را مورد ارزیابی قرار دهیم. برای مسائل با K رده، ماتریس پیچیدگی، به صورت یک ماتریس k*kمطابق جدول۲ خواهد بود که عناصر روی قطر اصلی، عناصری هستند که درست تشخیص داده شده‌اند و مابقی اشتباه تشخیص داده شده‌اند.

جدول۲: ماتریس پیچیدگی برای مسائل چند رده‌ای

		Predicted Class					
		C_1	C_2	.	.	.	C_k
True Class	C_1	n_{11}	n_{12}	.	.	.	n_{1k}
	C_2	n_{21}	n_{22}	.	.	.	n_{2k}
	.	.	.	.	...		.
	.	.	.	.	...		.
	.	.	.	.	...		.
	C_k	n_{k1}	n_{k2}	.	.	.	n_{kk}

بنابراین صحت، مطابق معادله‌ی (۴) برابر است با مجموع عناصر روی قطر اصلی تقسیم بر تعداد رکوردها[31].

$$\text{Accuracy} = \frac{\sum_{i=1}^{k} \mathrm{n_{ij}}}{\sum_{\mathrm{i},j=1}^{k} \mathrm{n_{ij}}} \quad (۴)$$

معیارهای بازخوانی و دقت برای هر رده مطابق معادله‌ی (۵) و (۶) محاسبه می‌شوند[23].

$$\mathrm{R}_i = \frac{TP_i}{\mathrm{TP}i+FN_i} = \frac{\mathrm{n_{ij}}}{\sum_{j=1}^{k} \mathrm{n_{ij}}} \quad (۵)$$

$$\mathrm{P}i = \frac{TP_i}{TP_i + FP_{\mathrm{i}}} = 1 - \frac{\mathrm{n_{ij}}}{\sum_{j=1}^{k} \mathrm{n_{ij}}} \quad (۶)$$

TP_i، تعداد رکوردهایی است که به درستی متعلق به رده‌ی i تشخیص داده شده‌اند. FP_i، تعداد رکوردهایی است که متعلق به رده‌ی i نبوده‌اند اما رده‌بند آن‌ها را جزء رده‌ی i تشخیص داده است. FN_i، تعداد رکوردهایی است که متعلق به رده‌ی i بوده‌اند اما رده‌بند، آن‌ها را جزء رده‌ی i تشخیص نداده است. معیارهای بازخوانی و دقت برای هر رده به صورت جداگانه محاسبه می‌شوند. در صورتی که بخواهیم این معیارها را برای کل مسئله محاسبه کنیم، دو روش میانگین‌گیری برای این کار وجود دارد: میکرو و ماکرو . روش میانگین‌گیری ماکرو بیشتر تحت تاثیر کارایی رده‌بندهای مربوط به رده‌های اقلیت قرار دارد. به همین خاطر در این تحقیق، از روش میانگین‌گیری ماکرو؛ مطابق معادله‌های (۷) و (۸)؛ برای محاسبه‌ی معیار کلی استفاده شده است. در این روش ابتدا معیار مورد نظر برای هر رده به صورت جداگانه محاسبه می‌شود و سپس میانگین آن‌ها به عنوان معیار کلی در نظر گرفته شده است.

$$\boldsymbol{R_{mac}} = 1/\mathrm{k}\sum_{i=1}^{k} R_i \quad (۷)$$

$$\boldsymbol{P_{mac}} = 1/\mathrm{k}\sum_{i=1}^{k} \mathrm{P_i} \quad (۸)$$

۴. ساختار اجرایی تحقیق

از نظر زمانی داده‌ها و اطلاعات مورد استفاده مربوط به مشتریان حقیقی است که از سال ۸۵ تا ۸۹ اقدام به دریافت وام قرض‌الحسنه از موسسه‌ی مالی مورد مطالعه نموده‌اند. تعداد رکوردهای موجود پس از پیش‌پردازش داده ۹۸۵۷ رکورد می-باشد. پس از پیش‌پردازش اولیه و یکی کردن تراکنش‌ها و داده-های جدا از هم مرتبط با مشتریان، در نهایت همه‌ی داده‌ها در یک پایگاه داده ادغام می‌شوند. ویژگی‌های نهایی، نوع آن‌ها و مقادیر ممکن برای هر یک در جدول ۳ نمایش داده شده است.

همان‌طور که گفته شد، جهت استفاده از تکنیک تحلیل سلسه مراتبی در فرآیند آماده‌سازی رویکرد پیشنهادی در شکل ۴ استفاده می‌شود. در ابتدا با توجه به عوامل تعریف شده‌ی تاثیرگذار بر امتیاز رفتاری مشتریان، اهداف، معیارها و زیرمعیارها تعیین می‌شوند. نمونه‌ی مطالعاتی این تحقیق شامل ۳ معیار گروه اصلی است که هر کدام از این معیارها دارای زیرمعیارهای مربوط به خود می‌باشند. در شکل ۴ یک نمایش گرافیکی سلسله مراتبی از این عوامل نشان داده شده است، که در راس آن هدف کلی مساله و در سطوح بعدی معیارها و زیرمعیارها قرار می‌گیرد.

برای ارزیابی این عوامل و مقایسه‌ی آن‌ها نیاز به نظرخواهی از افراد آگاه می‌باشد. به همین خاطر با ۱۰ نفر از کارشناسان بانک مورد نظر مصاحبه و داده‌های مورد نیاز از طریق پرسش‌نامه‌های تکمیل شده توسط آن‌هاجمع‌آوری شد. پرسش-نامه بر اساس قضاوت عددی طراحی شد. جهت وزن دهی

متغیرها، بعد از جمع‌آوری داده‌ها و انتقال آن‌ها به نرم‌افزار Expert Choice11، با اعمال تکنیک تحلیل سلسله مراتبی وزن معیارها و زیرمعیارها محاسبه و بیان شده است. بعد از از وارد کردن داده‌ها در نرم‌افزارExpert Choice11 پرسش نامه‌هایی را که نرخ ناسازگاری کمتر از ۰.۱ داشته اند در فرآیند تحلیل مورد استفاده قرار گرفته است و در غیر آن صورت پرسش نامه مذکور از فرآیند تصمیم گیری حذف شده است. جدول ۴ نتایج تحلیل پرسش‌نامه‌های تکمیل شده توسط کارشناسان را بیان می‌کند.

جدول ۳: مجموعه ویژگی‌های مورد استفاده برای امتیازبندی رفتاری

ردیف	نام ویژگی	نوع ویژگی	مقادیر ممکن
۱	شناسه‌ی مشتری	اسمی	شماره‌ی مشتری
۲	جنس	اسمی	۰: زن ۱: مرد
۳	سن	اسمی	۱: کمتر از ۲۵ سال ۲: ۲۶ تا ۳۵ سال ۳: ۳۵ تا ۵۰ سال ۴: بالاتر از ۵۰ سال
۴	شغل	اسمی	۱: کارمند ۲: روحانی ۳: شغل آزاد
۵	کارکرد حساب اصلی۱۲ماه قبل از اخذ وام (میانگین ریال/روز)	عددی	[۰و۱]
۶	موجودی اولیه	عددی	[۰و۱]
۷	تعداد حساب های فعال مشتری	عددی	[۰و۱]
۸	میانگین زمانی مراجعات	عددی	[۰و۱]
۹	سابقه‌ی مشتری در بانک	عددی	[۰و۱]
۱۰	میانگین کارکرد سایر حساب‌ها	عددی	[۰و۱]

با توجه به نتایج حاصل از پرسش‌نامه، درجه اهمیت شاخص‌های فوق تعیین گردید. و در نهایت پس از محاسبه‌ی میزان اهمیت با توجه به عددی و یا اسمی بودن شاخص ها، ضرایب بدست آمده مطابق رابطه‌ی ۳ بر روی داده‌ها اعمال شده است.

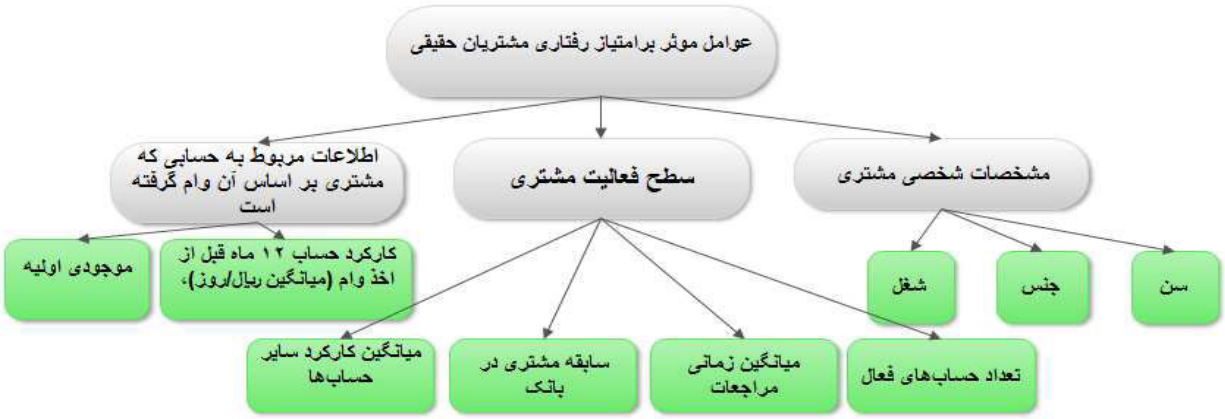
۴-۱. پیش‌بینی امتیاز رفتاری

در ادامه بخش به معرفی نتایج به دست آمده از اجرای الگوریتم خواهیم پرداخت. پیاده سازی الگوریتم‌های پیشنهادی در نرم‌افزار کلمنتاین انجام شده است. پیش از عملیات رده‌بندی، ابتدا باید مجموعه داده به دو بخش آموزش و آزمون تقسیم شود. برای تولید مجموعه‌ی آموزش و آزمون از روش نمونه‌برداری طبقه‌بندی شده[۱۴] استفاده شده است. همان‌طور که قبلا بیان شد تعداد رکوردهای موجود پس از پیش‌پردازش داده ۹۸۵۷ رکورد می‌باشدکه با استفاده از روش نمونه‌برداری طبقه‌بندی شده، داده‌ها به دو قسمت آموزش به میزان ۷۰٪، و اعتبارسنجی به میزان ۳۰٪ تقسیم شد.

جدول ۴: نتایج تحلیل پرسش نامه های تکمیل شده توسط کارشناسان

شاخص	میزان-اهمیت شاخص	ترتیب اولویت
کارکرد حساب ۱۲ ماه قبل از اخذ وام (میانگین ریال/روز)	۰/۴۴۲	۱
موجودی اولیه	۰/۰۸۶	۳
تعداد حساب های فعال مشتری	۰/۰۴۶	۶
میانگین زمانی بین مراجعات	۰/۰۲۴	۷
سابقه‌ی مشتری در بانک	۰/۰۵۳	۵
میانگین موجودی سایر حسابها	۰/۲۶۵	۲
سن	۰/۰۱۵	۸
جنس	۰/۰۱۴	۹
شغل	۰/۰۵۴	۴

شکل ۴: ساختار سلسله مراتبی معیارها در رویکرد پیشنهادی



با اعمال ضرایب بدست آمده از تحلیل سلسله مراتبی فرآیندی بر روی داده‌های نرمال، مجموعه داده جدیدی بدست آمد. در این بخش کارایی الگوریتم‌های رده‌بندی روی این مجموعه داده با داده‌های نرمال بدون این ضرایب مقایسه شده است. بر این اساس، الگوریتم‌های رده‌بندی مختلف اجرا و نتایج بر اساس معیار صحت با هم مقایسه شده‌اند.

جدول ۵: مقایسه نتایج کارایی الگوریتم‌های مورد بررسی

تکنیک	مدل/های مورد استفاده	صحت داده‌های آموزش	صحت داده‌های آزمون
داده‌های نرمال	درخت C5.1	۶۶/۵۸	۶۶/۴۳
	بردارهای ماشین پشتیبان(SVM)	۵۹/۴۷	۵۹/۲۵
	رگرسیون لجستیک	۵۹/۷۱	۵۹/۰۵
	شبکه عصبی	۶۳/۰۶	۶۲/۶۸
داده‌های حاصل از تاثیر ضرایب AHP	درخت C5.1	۶۸/۸۷	۶۷/۷۲
	ماشین بردار پشتیبان	۵۰/۶۴	۵۰/۶۱
	رگرسیون لجستیک	۶۱/۹۲	۶۰/۲۲
	شبکه عصبی	۶۴/۲۷	۶۴/۱۶

بر اساس جدول ۵، کارایی الگوریتم‌های رده‌بندی درخت تصمیم C5، رگرسیون لجستیک و شبکه عصبی، با داده‌های بدست آمده از تاثیر ضرایب AHP، بهبود یافته است. در این بین، تغییرات الگوریتم‌های C5 و رگرسیون لجستیک جالب توجه‌تر می باشد و در شبکه عصبی تغییرات کمتر می‌باشد. این مورد هم در مورد داده‌های آموزش و هم داده‌های آزمون صادق است. ستون سوم جدول ۵ نتایج حاصل از اعمال مدل بر روی مجموعه‌ی آزمون می‌باشد.

اما کارایی الگوریتم ماشین بردار پشتیبان، کاهش پیدا کرده است. این الگوریتم، از یک نگاشت غیرخطی برای تبدیل داده‌های مجموعه‌ی آموزش به ابعاد بالاتر استفاده می‌کند. با این بعد جدید، الگوریتم بدنبال یک ابرصحفه‌ی[15] جداساز بهینه می-گردد. این ابرصحفه، یک مرز تصمیم‌گیری است که رکوردهای یک رده را از رده‌ی دیگر جدا می‌کند. از طریق یک نگاشت مناسب غیرخطی به ابعاد بالاتر، داده‌های دو رده همیشه می-توانند به وسیله‌ی یک ابر صحفه جدا شوند. ماشین بردار پشتیبان، ابرصحفه را با استفاده از بردارها[16] (رکوردهای مجموعه‌ی آموزش) و حاشیه‌ها[17] (که بوسیله‌ی بردارها تعریف می‌شوند) پیدا می‌کند[30]. برای مسائل چند رده‌ای، رده‌بندهای ماشین بردار پشتیبان، می‌توانند با هم ترکیب شوند. در صورتی که در یک مسئله، n رده داشته باشیم، n رده‌بند به صورت جداگانه آموزش داده می‌شوند. هر رده‌بند برای یک رده. به عبارت دیگر هر بار یکی از رده‌ها به عنوان رده‌ی مثبت و مابقی رده‌ها به عنوان رده‌ی منفی در نظر می‌گیریم و به این ترتیب برای هر رده، یک رده‌بند تولید می‌شود و سرانجام رده‌بندهای بدست آمده با هم ترکیب می‌شوند. اگرچه با توجه به روال کاری شرح داده شده، زمان یادگیری ماشین بردار پشتیبان نسبت به

سایر الگوریتم‌ها کندتر است و این مسئله در مورد مسائل چند رده‌ای محسوس‌تر می‌باشد، با این وجود در مقالات بررسی شده در این حوزه که اکثرا بر روی مجموعه داده‌های استاندارد UCI[۱۸] انجام گرفته، الگوریتم ماشین بردار پشتیبان نتایج خوبی کسب کرده بود. اما در این مجموعه پس از اعمال ضرایب فرآیند تحلیل سلسله مراتبی بر روی داده‌ها این الگوریتم جزء ضعیف‌ترین الگوریتم‌ها بود. با توجه به نتایج بدست آمده، الگوریتم ماشین بردار پشتیبان با روش‌های آماده‌سازی مبتنی بر انتخاب ویژگی‌ها بهتر عمل کرده و اعمال ضرایب خارجی و روش پیشنهادی برای این الگوریتم مناسب نمی‌باشد.

با توجه به نتایج بدست آمده، الگوریتم C5 بالاترین کارایی را روی مجموعه‌ی آموزش دارند این الگوریتم همچنین روی داده‌های آزمون هم خوب عمل می‌کند و بیشترین دقت را دارد.

۴-۲. نتایج حاصل از متوازن‌سازی با بکارگیری الگوریتم Modifiedbagging

در بخش دوم سعی می‌کنیم داده‌ها را متوازن نماییم. جهت متوازن‌سازی از الگوریتم ModifiedBagging استفاده شده است. در این الگوریتم، زیرنمونه‌هایی از رده‌های اکثریت تولید می‌شود. سپس هر یک از این زیرنمونه‌ها با تمام نمونه‌های رده‌ی اقلیت ترکیب شده و مجموعه‌ای برای آموزش رده‌بند تولید می‌کنند. در مسئله‌ی مورد مطالعه، ۵ رده تعریف شده که رده‌ی چهارم رده‌ی اقلیت بوده و سایر رده‌ها، رده‌های اکثریت را تشکیل می‌دهند. بنابراین ما باید با انتخاب ضرایب مناسب، زیرنمونه‌هایی از رده‌های اکثریت را انتخاب کنیم تا در ترکیب با رده‌ی اقلیت، موجب بهبود کارایی مدل‌های پیش‌بینی شوند. در نهایت نیز تمام رده‌بندها با هم ترکیب شده و مدل نهایی را تولید می‌کنند. روش‌های یادگیری نامتوازن نسبت به رده‌بند پایه حساس هستند و هر یک از آن‌ها با رده‌بندهای خاصی بهتر عمل می‌کنند. برای مثال زیرنمونه‌برداری تصادفی با درخت تصمیم بهتر عمل می‌کند. در حالی که روش بیش‌نمونه برداری تصادفی با رگرسیون لاجستیک بهتر عمل می کند .[29] با توجه به این نکته که الگوریتمی که جهت متوازن‌سازی داده‌ها به کار برده می شود روش Modifiedbagging می‌باشد که مبتنی بر تکنیک زیرنمونه‌برداری می‌باشد کار را با الگوریتم C.5 دنبال می کنیم. این الگوریتم همچنین قبل از متوازن سازی نیز نسبت به سایر الگوریتم‌ها از صحت بالاتری برخوردار بوده است.

نتایج بدست آمده از اجرای الگوریتم C.5، به ازای انتخاب ضرایب مختلف برای هر یک از رده‌ها در جدول۶ آمده است.

جدول ۶: نتایج حاصل از اجرای الگوریتم به ازای انتخاب ضرایب مختلف برای رده‌های اکثریت

دقت						بازخوانی						صحت	ضرایب انتخاب شده برای هر رده
ماکرو	۵	۴	۳	۲	۱	ماکرو	۵	۴	۳	۲	۱		
۵۷.۷۷	۵۵.۵۱	۶۸.۱۰	۵۲.۵۶	۵۶.۷۰	۵۶	۵۸.۴۳	۵۵.۹۳	۷۱.۶۲	۵۰.۹۰	۴۵.۷۱	۶۸.۰۱	۵۸.۴۵	۰.۲۶،۰.۱،۰.۲۷ ،۱ ،۰.۲۸
۷۲.۷۱	۸۰.۱۰	۶۳.۷۵	۷۱.۳۳	۷۶.۷۰	۷۱.۸۰	۷۳.۵۱	۷۶.۸۱	۶۲.۰۳	۷۵.۵	۷۸.۹۱	۷۴.۳۳	۷۳.۶۴	۰.۶ ،۰.۵،۰.۶،۱،۰.۶
۷۰.۸۹	۷۸.۰۰	۵۵.۰۷	۷۴.۹۲	۷۹.۰۶	۶۷.۴۴	۷۱.۸۸	۷۴.۰۰	۵۴.۹۳	۷۵۶۸	۸۱۶۲	۷۳.۱۹	۷۱.۸۹	۰.۸،۰.۷،۰.۸،۱،۰.۸

جدول ۷: نتایج حاصل از ترکیب اجراهای الگوریتم

دقت						بازخوانی						صحت	تعداد اجرا
ماکرو	۵	۴	۳	۲	۱	ماکرو	۵	۴	۳	۲	۱		
۷۴.۷۴	۷۹.۶۲	۶۵.۷۳	۷۳.۸۰	۷۷.۱۳	۷۷.۴۳	۷۴.۹۵	۷۶.۵۱	۶۴.۹۳	۷۷۶۸	۷۹.۲۲	۷۶.۴۴	۷۵.۰۳	۱۰
۷۵.۰۲	۷۹.۸۲	۶۶.۴۱	۷۳.۳۳	۷۸.۲	۷۷.۴۵	۷۵.۱۰	۷۵۶۳	۶۵.۰۲	۷۸.۱۹	۷۹.۵۷	۷۷.۱۲	۷۵.۱۶	۱۵
۷۵.۵۷	۷۹.۸۵	۶۶.۴۷	۷۴.۰۳	۷۹.۰۰	۷۸.۵۲	۷۶.۲۱	۷۵۶۳	۶۷.۹۳	۷۹.۰۱	۷۹.۵۷	۷۸.۹۱	۷۶.۲۳	۲۰

جدول ۸: نتایج حاصل از اجرای الگوریتم بر روی مجموعه‌های آموزش و آزمون

دقت						بازخوانی						صحت	(۳۰،۷۰)
ماکرو	۵	۴	۳	۲	۱	ماکرو	۵	۴	۳	۲	۱		
۷۷۵.۵۷	۷۹.۸۵	۶۶.۴۷	۷۴.۰۳	۷۹.۰۰	۷۸.۵۲	۷۶.۲۱	۷۵۶۳	۶۷.۹۳	۷۹.۰۱	۷۹.۵۷	۷۸.۹۱	۷۶.۲۳	آموزش
۷۴.۲۱	۷۸.۰۴	۶۴.۱۶	۷۳.۳۳	۷۸.۲	۷۷.۴۵	۷۴.۰۸	۷۳.۲۹	۶۵.۱۸	۷۶.۰۶	۷۹.۰۱	۷۶۸۹	۷۴.۰۸	آزمون

بر اساس جدول ۶، بهترین نتایج به ازای؛ $0.6, 0.5, 0.6, 1, 0.6$ بدست آمده است. به ازای انتخاب این ضرایب جهت ساخت زیرنمونه‌هایی از رده‌ی اکثریت و اقلیت، دقت و بازخوانی همه‌ی رده‌ها در حد قابل قبولی می‌باشد. از آنجا که تعداد داده‌های رده‌ی ۴ کم می‌باشد، نمی‌توان ضرایب سایر رده‌ها را خیلی کوچک در نظر بگیریم. چون در این صورت زیرمجموعه‌های حاصل نیز دچار عدم توازن شده و نتایج مطلوب حاصل نخواهد شد. به عبارتی در این حالت اگرچه دقت و بازخوانی برای رده‌ی چهارم به نحو قابل توجهی افزایش می‌یابد ولی این مقادیر در سایر رده‌ها کاهش پیدا می‌کند. در مواقعی که بخواهیم تاثیر رای درخت تصمیم را بالابریم و یا در مواقعی که پایگاه داده دو درخت تصمیم متفاوت باشد می‌توان از یک رده‌بند، بیش از یکبار استفاده نمود. در این تحقیق نیز در ادامه نتایج حاصل از درخت‌های تصمیم که در اینجا الگوریتمC.5 می‌باشد ترکیب می‌شوند. تعداد اجرا برای یافتن مدل بهینه با اجرا به ازای مقادیر مختلف بر روی داده‌های آموزش تعیین می‌گردد. نتایج حاصل از ترکیب اجرای الگوریتم‌ها به ازای تعداد اجراهای مختلف در جدول ۷ امده است.

با افزایش تعداد اجرا، کارایی رده‌بند روی مجموعه‌ی داده‌ها افزایش می‌یابد. این روند افزایش تا ۲۰ بار اجرا ادامه داشته است. اما به ازای اجرای بالاتر از ۲۰ مقدار معیارهای صحت، دقت و بارخوانی تغییر چندانی نمی‌کند. از این رو ۲۰ بار اجرا، به عنوان تعداد اجرای بهینه انتخاب می‌شود. از آنجایی که مدل فوق پس از ۲۰ بار اجرا بهینه شده‌است. مدل بهینه با همین تعداد اجرا روی مجموعه داده‌های تست اجرا خواهد شد که نتایج آن در جدول ۸ آورده شده است.

در این بخش نتایج نهایی حاصل از چارچوب پیشنهادی، بانتایج اولیه حاصل از اجرای الگوریتم‌های پایه‌ی درخت تصمیم C.5، ماشین بردار پشتیبان، شبکه‌ی عصبی و رگرسیون لجستیک مقایسه شده است. همان‌طور که جدول ۹ نشان می-دهد، کارایی چارچوب پیشنهادی از سایر الگوریتم‌های شرکت-کننده در این مقایسه بیشتر است.

جدول ۹. مقایسه نتایج کارایی الگوریتم‌های‌های مورد بررسی

مدل/های مورد استفاده	صحت داده‌های آموزش	صحت داده‌های آزمون
الگوریتم پیشنهادی	۷۶.۲۳	۷۴.۰۸
درخت C5.1	۶۶/۵۸	۶۶/۴۳
ماشین بردارهای ماشین پشتیبان(SVM)	۵۹/۴۷	۵۹/۲۵
رگرسیون لجستیک	۵۹/۷۱	۵۹/۰۵
شبکه عصبی	۶۳/۰۶	۶۲/۶۸

۵. نتیجه‌گیری

امتیازدهی نظامی است که بوسیله‌ی آن بانک‌ها و موسسات اعتباری با استفاده از اطلاعات حال و گذشته‌ی متقاضی، احتمال عدم بازپرداخت وام توسط وی را ارزیابی نموده و به متقاضی امتیاز می‌دهد. این روش مشتریان را بی‌طرفانه و بر اساس آمار و اطلاعات کمی امتیازبندی می‌نماید. سیستم امتیازدهی اعتباری مشتریان یکی از ابزارهای مهمی هست که بانک‌ها برای مدیریت و کنترل ریسک اعتباری به آن نیاز دارند.

روش امتیازدهی هم از طریق سیستم‌های امتیازبندی داخلی بانک‌ها و هم از طریق شرکت‌های سنجش اعتباری قابل محاسبه است. ولی در سطح وام‌های اعطایی به اشخاص حقیقی سیستم‌های داخلی مناسب‌تر هستند مضاف بر اینکه پیاده‌سازی موفق سیستم‌های داخلی در محیط رقابتی کنونی یک مزیت رقابتی برای بانک‌ها محسوب می شود. موید این موضوع توصیه بانک تسویه بین‌المللی و کمیته بال در سال ۲۰۰۱ به بانک‌ها جهت استقرار و اجرای سیستم امتیازبندی داخلی در کنار موسسات رتبه‌بندی است. در بانک ها و موسسات مالی حجم قابل ملاحظه‌ای از داده وجود دارد اما تا زمانی که داده‌ها به صورت رکوردهای ماهانه و تراکنش‌های روزانه می‌باشند، تحلیل پایگاه داده بانک برای مدیریت رفتار مشتریان امری مشکل می باشد. از این رو داده‌کاوی ابزار مناسبی برای تحلیل این داده و تخمین امتیاز رفتاری مشتریان حقیقی می‌باشد.

ولی اصل مهمی که در ارتباط با الگوریتم های داده‌کاوی وجود دارد، این است که هیچ الگوریتمی برای تمام داده‌ها مناسب نخواهد بود. به عبارت دیگر هیچ‌گاه نمی‌توان ادعا کرد که یک الگوریتم یا روش برای تمام داده‌ها مناسب بوده و نتیجه‌ی مطلوبی ارائه می دهد بلکه نتایج حاصل از الگوریتم‌های داده کاوی، وابستگی بسیاری به مجموعه‌ی مورد استفاده دارند.

از این رو دغدغه‌ی اصلی در این تحقیق، چگونگی بکارگیری داده‌کاوی برای امتیازبندی رفتاری مشتریان در بانک مورد مطالعه بوده است. در این تحقیق نیز ابتدا به مطالعه‌ی کارهای مرتبط پرداخته شده است. مطالعات زیادی هم در این زمینه وجود دارد که اکثر آن‌ها روی مجموعه داده‌ی استاندارد UCI می‌باشد. در مقالات بررسی شده‌ی قبلی، یک سری روش های انتخاب و آماده‌سازی ویژگی ها به کار برده شده‌اند اما این روش‌ها تنها از دیدگاه علم پایگاه داده انتخاب می شوند. لیکن دانش و اطلاعات مسئله و نظر تصمیم‌گیران می‌تواند ویژگی‌ها و داده‌های بهتر و با اهمیت‌تری برای الگوریتم کلاس‌بندی و مدل تامین نماید. بر این اساس در این پژوهش با بکارگیری ضرایب بدست آمده از تحلیل سلسله مراتبی ، یک رویکرد پیشنهادی جهت آماده‌سازی داده‌های نهایی، ارائه گردیده است. نتایج حاصل از اجرای الگوریتم‌های رده‌بندی مختلف بر روی مجموعه داده‌ی مورد مطالعه نشان داد که این پیشنهاد، باعث کارایی الگوریتم های رده‌بندی، درخت تصمیم C.5 ، شبکه عصبی و رگرسیون لجستیک شده است ولی کارایی الگوریتم ماشین بردار پشتیبان کاهش پیدا کرده و همچنین الگوریتم C.5 نسبت به بقیه نتایج بهتری تولید می کند.

در مقالات بررسی شده تقسیم‌بندی به این صورت بوده که مشتریان به دو گروه خوش‌حساب و بدحساب تقسیم می شوند. ولی به منظور شناخت دقیق تر رفتار مشتریان در این تحقیق ۵ رده اصلی و مهم جهت شناسایی ریسک اعتباری مشتریان تعریف شد. با افزایش تعداد رده‌ها توازن داده‌ها کاهش پیدا کرده است. به همین خاطر در مسئله ی مورد بررسی به منظور بهبود کارایی الگوریتم‌های رده‌بندی، از روش متوازن سازی modifiedbagging استفاده شده است.

modifiedbagging یک روش مبتنی بر زیرنمونه‌برداری است. بر این اساس پس از انتخاب ضرایب مناسب برای هر یک از رده ها و ترکیب این نتایج، شاهد بهبود کارایی الگوریتم رده‌بندی به میزان قابل توجهی هستیم.

نتایج نهایی حاصل از اجرای چارچوب پیشنهادی به این شرح می‌باشد؛ از میان الگوریتم های رده بندی الگوریتم درخت تصمیم C.5 بیشترین میزان صحت را دارا می باشد و این الگوریتم ، مناسب ترین الگوریتم برای پیش بینی امتیاز رفتاری مشتریان می باشد. روش های متوازن سازی نسبت به آماده سازی، موجب بهبود کارایی به میزان قابل توجه تری می شوند. نهایتا ترکیب این تکنیک‌ها در کنار یکدیگر موجب شده است که قدرت پیش بینی الگوریتم رده‌بندی نسبت به حالت پایه به میزان چشم‌گیری افزایش یابد.

مراجع

[۱] پارکر ج. ، ترجمه: عبدالمجید انصاری، " ابعاد مدیریت ریسک، تعاریف و کاربردهای مدیریت ریسک در خدمات مالی"، تازه های اقتصادی، شماره ۱۳۸۲،۶۹.

[۲] گلزار. آقایی، علی. صنیعی فرد، "ارایه یک مدل سیستم خبره تصمیم-گیر در تخصیص کارآمد اعتبار به پروژه‌های فناوری اطلاعات در نظام بانکی کشور". سومین کنفرانس بین‌المللی توسعه نظام تامین مالی در ایران، ۱۳۸۹، صفحات۱ -۳۰.

[3] T. Bellotti. and J. Crook, " Modelling and estimating Loss Given Default for Credit Card", CRC working paper, pp. 2-20, 2008.

[4] E.W.T. Ngai, Li Xiu and D.C.K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification". Expert Systems with Applications, Vol. 36, pp. 2592-2602 ,2009.

[5] L. K. Keung, " An intelligent CRM system for identifying high-risk customers: An ensemble data mining approach".ICCS, Springer-Velage, Berlin Heidelberg, pp. 486-489, 2007.

[6] B.W, Lin, "Information technology capability and value creation" Evidence from the US banking industry. Technology in Society,Vol. 29, pp. 93-106, 2010.

[7] C. Th. Lyn, "A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers", International Journal of Forecastin,.Vol. 16, pp. 149–172, 2000.

[8] H. Hakanes and I. Schanabel," Credit risk transfer and bank competition "Journal of Financial Intermediation,vol. 19, pp. 308-332, 2010.

[9] Basel committee on Banking," Studies on the Validation of Internal Rating Systems".working paper. 14 .February , 2006.

[10] S.K. Bagchi, credit Risk Management. Jaico Publishing House, pp. 278, 2006.

[11] B. Baesense, T. Van Gestel, S. Viaene,M. stepanova, J. Suykens and J. Vanthienen, "Benchmarking state-of-art classification algorithms for credit scoring," Journal of Operational Research Socity, Vol. 54, pp. 627-635, 2003.

[12] M. ch. Tsai, sh. P. Lin, Ch. Ch. Cheng and Y. P. Lin, "The Consumer Loan Default prediction model- An application of DEA-DA and neural network". Expert System with Applications,vol. 36, pp. 11682-11690, 2009.

[13] Ch. L. Huang. and M. Ch. Chen, , "Credit scoring with a data mining approach based on support vector machines," Expert systems with applications, Vol.3, pp.847-856, 2011.

[14] F.L. Chen and F.C.Li, "Combination of feature selection approach with SVM in credit scoring," Expert System with Applications, Vol. 37, No. 7, pp. 4902-4909, 2010.

[15] N.Ch. Hsieh, "Hybrid mining approach in the design of credit scoring models," Expert Systems with Applications, Vol.28, pp. 655–665, 2005.

[16] T. Leea, C.C Chiub, Y.C, Chouc ands C.J. Lud," Mining the customer credit using classification and regression tree and multivariate adaptive regression splines," Expert System with Applications, Vol. 50, PP. 1113–1130, 2006.

[17] C. R. Hayhoe, L. Leach and P. R.Turner, "Discriminating the number of credit cards held by college students using credit and money attitudes," Journal of Economic Psychology, Vol.20 ,pp. 643–656, 1999.

[18] D. West, "Neural network credit scoring models," Computers and Operations Research, Vol. 27, pp. 1131–1152, 2000.

[19] C.F. Tsai and J.W. Wu, "Using Neural Network Ensembles for Bankruptcy Prediction and Credit Scoring," Expert Systems with Applications, Vol. 34, pp. 2639–2649, 2008

[20] B. Baesens, T. V. Gestel, M. Stepanova, D. V. D. Poel and J. Vanthienen, "Neural network survival analysis for personal loan data," Journal of the Operational Research Society, Vol. 56 , pp. 1089–1098, 2005.

[21] G. Wang, J. Hao, J. Ma and H. Jiang," A Comparative Assessment of Ensemble Learning for Credit Scoring," Expert System with Applications,Vol. 38, pp.223–230 2011.

[22] S. TunLi, W. Shiue and M.H. Huang, "The Evaluation of Consumer Loans Using Support Vector Machines," Expert System with Application, Vol.30,pp. 772–782, 2006.

[23] C.L. Huang, M.C. Chen and C.J. Wang, "Credit Scoring with Data Mining Approach Based on Support Vector Machine", Expert System with Application 37, 847–856, 2007

[24] G. Arminger , D. Enache and T. Bonne," Analyzing Credit Risk Data: A Comparison of Logistic Discrimination, Classification Tree Analysis, and Feedforward Networks", Comput Statist 12, pp. 293–310, 1997.

[25] T. L. Saaty, 'Decision-making with the AHP: Why is the principal eigenvector necessary', European Journal of Operational Research, pp.85-91, 2003.

[26] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalance credit scoring data sets," Expert systems with applications, Vol. 39, pp. 3346-3353, 2012.

[27] S. F. Crone and S. Finlay, Instance sampling in credit scoring: An empirical study of sample size and balancing. International journal of forecasting, Vol. 28, pp. 224-238, 2012.

[28] Y. C. Tzeng, K. Sh. Chen and N. S. Chou," Using Modified Bagging and Boosting Algorithms in Multiple Classifiers System for Remote Sensing Image Classification," Journal of Phtgrammetry and Remote Sensing Vol. 12, No.3, pp. 241-256, September 2007.

[29] J. V. Hulse, T. M. Khoshgoftaar and A. Napolitano, "Experimental Perspectives on Learning from Imbalanced Data," In Proceedings of the 24th int. conf. on Machine learning. pp. 935-942, 2007.

[30] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques. Morgan Kaufmann, 2006.

[31] Y. Sun, M. S. Kamel and Y. Wang, "Boosting for Learning Multiple Classes with Imbalanced Class Distribution", Sixth Int. Conf on Data Mining, pp 592-602, 2006.

1 Analytic hierarchy process
2 Credit Scoring
3 Behavioral Scoring
4 Committee on Banking Supervision
5 InternalRating-Based Approach(IRB)
6 Probability of Default(POD)
7 cut off points
8 balancing
9Money attitudeFeature
10Money attitudeFeature
11 Feature selection
12 UCI machine learning repository. Available at: http://archive.ics.usi.edu/ml
13 Nominal
14 Stratified Sampling
15 Hyperplane
16 Vectors
17 Margins