به کارگیری الگوریتم‌های یادگیری ماشین برای اعتبار سنجی مشتریان بانک‌ها

حسین فقیه علی آبادی
کارشناسی ارشد مهندسی نرم افزار، گرایش شبکه‌های کامپیوتری، دانشگاه ارومیه

علی قنبری زاده
کارشناسی ارشد فناوری اطلاعات، گرایش تجارت الکترونیک، دانشگاه امیرکبیر

چکیده

در سال‌های اخیر اعطای تسهیلات بانکی با مشکلاتی مواجه شده است که استفاده از سیستم های اعتبار سنجی را برای بانک‌ها ضروری نموده است. با استفاده از تحلیل اطلاعات مربوط به مشتریان، بانک‌ها با استفاده از فرآیند داده‌کاوی می‌توانند به اعتبار سنجی متقاضیان تسهیلات و طبقه‌بندی آن‌ها به مشتریان خوب یا بد اعتبار را پرداخت کرد. درسیستم‌های بانکداری سنتی، مدیران اعتباری اغلب میزان اعتبار مشتریان را با توجه به تجربه خود می‌سنجند ولی در نظام بانکداری نوین، با محدودیت زمانی و تعدد فزاینده مشتریان مواجه هستیم. برای حل این مشکل، ما در این مقاله با استفاده از الگوریتم‌های یادگیری ماشین احتمال قصور یا عدم قصور در پرداخت‌ها را محاسبه کرده و با توجه به آن مشتریان را رتبه بندی می‌کنیم. هدف این پژوهش توسعه مدلی یکپارچه با دقت بالا برای اعتبار سنجی مشتریان است. که در همه الگوریتم‌ها این حاصل شده و به دقت بالا 90 رسیده‌ایم که این نشان از برتری مدل پیشنهادی میدهد. شبکه‌های عصبی به دلیل دقت به مراتب بالاتر و حجم محاسبات پایین تر نسبت به سایر روش‌های کلاسیک در پیش بینی رفتار اعتباری افراد متقاضی تسهیلات دارای اولویت هستند.

واژگان کلیدی: انتخاب ویژگی، اعتبار سنجی، بانک، هوش مصنوعی، یادگیری ماشین

۱. مقدمه

اعتبارسنجی مشتریان بانک‌ها به فرایند ارزیابی و تجزیه و تحلیل اطلاعات مربوط به مشتریان برای تعیین اعتبار و قابلیت پرداخت آن‌ها می‌پردازد. در این فرایند، اطلاعات مالی، اقتصادی، شخصی و تاریخچه پرداخت مشتریان برای تعیین سطح ریسک و اعتبار آن‌ها مورد بررسی قرار می‌گیرد. اهداف اصلی اعتبارسنجی شامل کاهش خطرات اعطای وام، حفظ سودآوری بانک، و جلوگیری از تعریض بودجه بانک است. از روش‌های اعتبارسنجی مشتریان بانک‌ها می‌توان به استفاده از اطلاعات مالی، بررسی تاریخچه پرداخت، استفاده از مدل‌های پیش‌بینی و روش‌های آماری و همچنین اعتبارسنجی خودکار با استفاده از سیستم‌های خودکار اشاره کرد.بانک‌ها به منظور آگاهی از نیازمندی‌ها و رفتار مشتریان خود در اعطای تسهیلات اعتباری باید به شناسایی ویژگی‌های آن‌ها بپردازند که این امر منجر به کاهش ریسک‌های بانکی از جمله ریسک اعتباری می‌شود. پژوهش‌ها و کاربردهای متعددی در حوزه اعتبارسنجی برای شناسایی مشتریان خوب یا بد صورت گرفته است. روش قضاوتی در اعتبارسنجی به دلیل خطا و زمان زیاد به تدریج جای خود را به روش‌های پارامتریک و ناپارامتریک داد. روش‌های پارامتریک مثل تحلیل تمایزی و رگرسیون لجستیک از ابتدای ظهور اعتبارسنجی مورد استفاده قرار گرفتند و سپس روش ناپارامتریک و داده کاوی مثل درختان تصمیم‌گیری، شبکه‌های عصبی و سیستم‌های خبره به کار گرفته شدند(Jiang ,2023). درخت تصمیم یکی از تکنیک‌های داده کاوی با قابلیت فهم بالا و سرعت مناسب در یادگیری الگو که می‌توانند برای طبقه‌بندی مشتریان در اعتبارسنجی مفید باشند. امروزه مسئله اعتبارسنجی به یکی از مسایل مهم مدیران و کارشناسان بانکی تبدیل شده است. در یک بانک، می‌توان درخت تصمیم گیری متنوعی را برای طبقه‌بندی و اعتبارسنجی مشتریان ایجاد نمود. از ویژگی‌های مدل پیشنهادی می‌توان به افزایش دقت در ساختار و محتوای درختان تصمیم‌گیری ، بهبود الگوریتم‌های یادگیری ماشین برای فهم آسان مدل طبقه بندی مشتریان، جلوگیری از اتخاذ تصمیم‌های احتمالی غلط کارشناسان بانکی در اعتبارسنجی، کاهش نیاز به تحلیل‌های پرهزینه و زمانبر در طبقه‌بندی مشتریان، انتخاب بهینه ویژگی‌های اعتبارسنجی مشتریان و در نهایت رضایت مندی آن‌ها در ارایه تسهیلات اعتباری متناسب با هر طبقه اشاره نمود(J., Zhang, 2022)

به طور معمول، (Shen F, 2022) مراحل اعتبارسنجی در بانک‌ها یا مؤسسات مالی شامل چندین گام است که به شرح زیر می‌باشد:

۱. جمع‌آوری اطلاعات: در این مرحله، اطلاعات مورد نیاز برای ارزیابی اعتبار مشتری جمع‌آوری می‌شود. این اطلاعات می‌تواند شامل اظهارنامه مالی، گزارش مالی، مدارک هویتی و سایر مدارک مرتبط باشد.
۲. ارزیابی اطلاعات مالی: در این مرحله، تحلیل‌گران مالی اطلاعات جمع‌آوری شده را بررسی می‌کنند و اعتبار و قدرت بازپرداخت بدهی مشتری را ارزیابی می‌کنند.
۳. ارزیابی ریسک: در این مرحله، میزان ریسک مرتبط با اعطای تسهیلات به مشتری، از جمله ریسک اعتباری و ریسک مالی، مورد بررسی قرار می‌گیرد.
٤. تصمیم‌گیری: بر اساس ارزیابی‌های انجام شده، تصمیمی در مورد اعطای یا عدم اعطای تسهیلات اعتباری به مشتری گرفته می‌شود.
٥. اعلام تصمیم: در این مرحله، مشتری از تصمیم گرفته شده در خصوص تسهیلات اعتباری باخبر می‌شود و در صورت لزوم توضیحاتی درباره تصمیم گرفته شده ارائه می‌شود.

همچنین (Guo, W, 2023)، ممکن است مراحل اعتبارسنجی براساس نوع تسهیلات و شرایط مخصوص مشتری، متغیر باشد. استفاده از الگوریتم‌های یادگیری ماشین در انتخاب ویژگی‌ها برای اعتبارسنجی مشتریان بانک ها مفید است. الگوریتم‌های انتخاب ویژگی ممکن است در بهینه محلی قرار گیرند و از طرف دیگر تعامل بین ویژگی‌ها را درنظر نمی‌گیرند و فرض می‌کنند، روابط بین ویژگی‌ها خطی بوده و ویژگی‌ها مستقل از هم می‌باشند. همچنین این الگوریتم‌ها تنها از برخی معیارها برای انتخاب ویژگی استفاده می‌کنند. در برخی از الگوریتم‌ها حساسیت بیشتری نسبت به برخی از ویژگی‌ها برای تفکیک شدن وجود دارد. که در نتیجه دارای دقت کم در طبقه‌بندی در مجموعه داده تست و ارزیابی و تناسب بیش از حد است. الگوریتم‌ها نسبت به داده‌های زیاد در ویژگی‌ها تمایل نشان می دهند. این الگوریتم‌ها در فرایند خود تنها از برخی معیارها و توابع در ساخت درخت تصمیم گیری استفاده می‌کنند. ریسک اعتباری خطر بدهی است که از ناتوانی وام گیرنده در پرداخت‌های لازم ناشی می‌شود. با وقوع بحران وام مسکن و بحران بدهی اروپا، ریسک اعتباری اهمیت بیشتری پیدا کرده است. امتیازدهی پرکاربردترین ابزار امتیازدهی مرتبط با مدل برای کاهش ریسک اعتباری است که ماهیت آن یک کار طبقه بندی است. امتیازدهی اعتبار بخش مهمی از حفظ یک محیط تجاری پایدار است. با این حال (Smirnov, 2023)، مدل‌های پیش‌بینی‌کننده می‌توانند به راحتی با چالش یک مشکل عدم تعادل کلاس به دلیل وجود سوابق بسیار کم مواجه شوند. کلاس اکثریت (غیر پیش‌فرض) به کلاسی اطلاق می‌شود که تعداد نمونه‌ها زیاد است، این با مسئله طبقه بندی سنتی متفاوت است، که بر این فرض استوار است که تعداد نمونه‌های آموزشی در کلاس‌های مختلف بسیار کم است. این مشکل باعث می‌شود که مدل تخمین‌های گمراه کننده ایجاد کند و نتایج پیش بینی را به سمت طبقه اکثریت سوگیری کند و نمونه‌های کلاس اقلیت را به طور دقیق شناسایی نکند. در همان زمان، معیارهای عملکرد برخی از مشکلات طبقه‌بندی سنتی در یک مورد مشکل CI غیرقابل اجرا می‌شوند. به عنوان مثال، اگر طبقه‌بندی‌کننده یک رکورد پیش‌فرض را به‌عنوان یک رکورد معمولی اشتباه ارزیابی کند، حتی اگر دقت پیش‌بینی کل بالایی را ارائه دهد، احتمال اینکه یک فرد، اجتناب از بازپرداخت و ایجاد زیان هنگفت برای بانک را افزایش می‌دهد برای از بین بردن مشکل مانند روش‌های نمونه‌گیری تصادفی و SMOTEها تعداد نمونه‌ها را بین کلاس‌های مختلف به روش‌هایی تغییر میدهد. رویکردهای سطح الگوریتم برای غلبه بر انحراف خروجی ها (مانند پردازش حساس به هزینه) از دیدگاه مدل، تنظیماتی را انجام می‌دهند. به‌ویژه در سال‌های اخیر، شبکه‌های متخاصم مولد توجه محققان را برای رسیدگی به مسئله به عنوان یک ابزار ترکیب‌کننده نمونه که شبیه به SMOTE است، به خود جلب کرده‌اند .در این مقاله پس از مقدمه‌ای که در این بخش بیان شد، در بخش دوم به مرور کارهای گذشته، در بخش سوم به بررسی روش پیشنهادی، در بخش چهارم به نتایج ارزیابی و در بخش پایانی به جمع بندی این مقاله می‌پردازیم.

۲. بررسی کارهای گذشته

(cohen, 2008) قابل پیش بینی بودن بازده سهام تامین کننده از شرکت‌های مشتری اصلی را نشان می‌دهند. این مقاله شواهدی از قابلیت پیش‌بینی بازده در بین شرکت‌های اقتصادی را پیدا می‌کند. این فرضیه را آزمایش می‌کنند که در حضور سرمایه‌گذاران مشروط به محدودیت‌های توجه، قیمت سهام به‌سرعت اخبار مربوط به شرکت‌های مرتبط اقتصادی را در بر نمی‌گیرد و قابلیت پیش‌بینی بازده را در سراسر دارایی‌ها ایجاد می‌کند. با استفاده از مجموعه داده‌ای از مشتریان اصلی شرکت‌ها برای شناسایی مجموعه‌ای از شرکت‌های مرتبط اقتصادی، نشان می‌دهند که قیمت سهام اخبار مربوط به شرکت‌های مرتبط را در بر نمی‌گیرد و حرکت‌های قیمتی قابل پیش‌بینی بعدی را ایجاد می‌کند.

(Hertzel,2008) تأثیر منفی قابل توجه پرونده‌های ورشکستگی شرکت‌ها را بر قیمت سهام تأمین کنندگان آن‌ها نشان می‌دهد. تحقیقات قبلی میزان مسری بودن ورشکستگی در صنایع را بررسی می‌کند. این مطالعه با بررسی اثرات ثروت ناشی از پریشانی بر مشتریان و تامین کنندگان قبل از ورشکستگی، تحقیقات را گسترش میدهد. به طور متوسط، اثرات مسری مهم قبل از ورشکستگی رخ می‌دهد و فراتر از رقبای صنعت در طول زنجیره تامین است. به طور خاص، ناراحتی پیش از ورشکستگی با اثرات منفی و قابل توجه قیمت سهام برای رقبا، مشتریان و تامین کنندگان همراه است. مطابق با انتظارات، اثرات مشتری و تامین کننده زمانی که سرایت صنعت شدیدتر باشد منفی‌تر است. ساختار صنعت و ماهیت محصول تخصصی نیز پیامدهای قابل توجهی برای تامین کنندگان و مشتریان دارد زمانی که یکی از اعضای اصلی زنجیره تامین مضطرب می شود.

(Kouvelis, 2018) دریافتند که شرکت‌ها رتبه‌های اعتباری شریک زنجیره تامین خود را در تامین مالی زنجیره تامین در نظر می گیرند. تأثیر رتبه‌بندی اعتباری را بر تصمیم‌های عملیاتی و مالی زنجیره تأمین با تأمین‌کننده و خرده‌فروشی که از طریق قرارداد تخفیف پرداخت زودهنگام در تعامل هستند، مطالعه می‌کنند. خرده فروش یک فرصت واحد دارد تا محصولی را از عرضه کننده سفارش دهد تا تقاضای نامشخص آینده را برآورده کند. خرده‌فروش و عرضه‌کننده دارای محدودیت سرمایه هستند و خرده‌فروش می‌تواند از وام‌های کوتاه‌مدت بانکی و اعتبارات تجاری برای نیازهای مالی خود استفاده کند، در حالی که عرضه‌کننده می‌تواند از وام‌های کوتاه‌مدت بانکی و یا پرداخت زودهنگام خرده‌فروشان استفاده کند. برای تمام تصمیمات عملیاتی مرتبط (قیمت عمده فروشی، نرخ اعتبار تجاری، وام های بانکی، و مقدار سفارش) برای شرکت‌های دارای محدودیت سرمایه تجزیه و تحلیل می‌کنند. در فروش اصلاح‌شده به بازی استکلبرگ خبرفروش با عرضه‌کننده به‌عنوان رهبر، نرخ‌های اعتبار تجاری تعادلی، قیمت عمده‌فروشی، وام‌های بانکی و مقدار سفارش را استخراج می‌کنند. این مقاله نشان می‌دهد که آستانه‌ای وجود دارد که اگر رتبه اعتباری تأمین‌کننده بالاتر از آن باشد، تأمین‌کننده اعتبار تجاری با نرخ بهره صفر ارائه می‌کند و خرده‌فروش فقط از اعتبار تجاری استفاده می‌کند. نویسندگان توضیح قابل قبولی برای عملکرد خرده فروشان رتبه بندی اعتباری بزرگ و خوب ارائه می دهند که نسبت نقدینگی کمی را حفظ می‌کنند و با تامین کنندگان کوچک در کشورهای در حال توسعه کار می‌کنند.

(Acemoglu,2015) اهمیت زنجیره‌های اعتباری تجاری را برای گسترش ورشکستگی شرکت‌ها تجزیه و تحلیل می‌کنند این مقاله استدلال می‌کند که میزان سرایت مالی از انتقال فاز را نشان می‌دهد، تا زمانی که بزرگی شوک‌های منفی تأثیرگذار بر مؤسسات مالی به اندازه کافی کوچک باشد، یک شبکه مالی با اتصال متراکم‌تر (مطابق با الگوی متنوع‌تری از بدهی‌های بین بانکی) مالی را افزایش می‌دهد. با این حال، فراتر از یک نقطه خاص، اتصالات متراکم به عنوان مکانیزمی برای انتشار شوک‌ها عمل می‌کند که منجر به یک سیستم مالی شکننده تر می‌شود. بنابراین، نتایج نشان می‌دهد که همان عواملی که در شرایط خاص به انعطاف‌پذیری کمک می‌کنند، ممکن است به عنوان منابع مهم ریسک سیستمیک تحت شرایط دیگر عمل کنند.

(Herskovic,۲۰۱۸) عوامل شبکه را به عنوان منابع ریسک سیستماتیک که در قیمت دارایی‌های تعادلی منعکس می‌شود، آشکار می کند. این مقاله قیمت گذاری دارایی را در یک مدل چندبخشی بررسی می‌کند که در آن بخش‌ها از طریق یک شبکه ورودی، خروجی به یکدیگر متصل می‌شوند. تغییرات در ساختار شبکه منابع ریسک سیستماتیک است که در قیمت دارایی‌های تعادلی منعکس می‌شود. دو ویژگی کلیدی شبکه ینی تمرکز شبکه و پراکندگی شبکه وجود دارد که برای قیمت دارایی مهم است. تمرکز شبکه میزان تسلط بر خروجی تعادل توسط چند بخش بزرگ را اندازه گیری می‌کند در حالی که پراکندگی شبکه میانگین تخصص

ورودی اقتصاد را اندازه گیری می‌کند. علاوه بر این، این دو عامل قیمت‌گذاری دارایی مبتنی بر تولید توسط ساختار شبکه تولید تعیین می‌شوند و می‌توانند از داده‌های ورودی، خروجی محاسبه شوند.

(Wenbo Wu, 2021) شواهدی یافتند مبنی بر اینکه شوک‌های اعتباری مثبت و منفی، همانطور که توسط جهش‌های شدید مبادله پیش فرض اعتباری تشدید می شوند، می‌توانند در امتداد زنجیره تامین به سایر شرکت‌های شریک منتشر شوند. چهار روش یادگیری ماشین را برای پیش‌بینی بازده مقطعی برای انتخاب صندوق تامینی اعمال می‌کنند. مدل پیش‌بینی را با مجموعه‌ای از ویژگی‌های خاص تجهیز می‌کنند که از بازده‌های تاریخی یک صندوق تامینی مشتق شده‌اند و انواع اطلاعات خاص صندوق را به دست می‌آورند. با ارزیابی عملکرد خارج از نمونه، متوجه می‌شوند که روش پیش‌بینی تقریباً در همه موقعیت‌ها به طور قابل‌توجهی از چهار شاخص تحقیقاتی صندوق سرمایه‌گذاری تامینی بهتر عمل می‌کند. در میان چهار روش یادگیری ماشین، متوجه شدند که شبکه عصبی عمیق در مجموع مؤثرترین است. با بررسی منبع مزیت روش‌شناختی روش ما با استفاده از مطالعه موردی، متوجه می‌شویم که پیش‌بینی مقطعی در بیشتر موارد از پیش‌بینی بر اساس رگرسیون سری زمانی بهتر عمل می‌کند. قابلیت‌های مدل‌سازی پیشرفته یادگیری ماشین این مزایا را بیشتر می‌کند. همچنین ویژگی‌های مبتنی بر بازده منجر به بازدهی بالاتری نسبت به معیار مجموعه‌ای از ویژگی‌های کلان می‌شوند، و روش پیش‌بینی ما بهترین عملکرد را زمانی که دو مجموعه از ویژگی‌ها با هم ترکیب می‌شوند، دارد.

تلاقی ویژگی روشی امیدوارکننده برای ثبت تعاملات بین ویژگی‌های خام است و به طور گسترده برای بهبود عملکرد بسیاری از کارهای پیش‌بینی، مانند نرخ استفاده می‌شود.(Hosaka, 2022) نتایج تلاقی ویژگی‌ها می‌تواند نشان‌دهنده همزمانی ویژگی‌ها و افزودن غیرخطی به داده‌ها باشد که می‌تواند عملکرد روش‌های یادگیری را به طور قابل‌توجهی بهبود بخشد. ماشین های فاکتورسازی و الحاقات آن، نمونه‌های شناخته شده‌ای از تعاملات ویژگی‌های یادگیری هستند که پیشنهاد شده‌اند. برای به دست آوردن تعاملات ویژگی‌های رتبه اول و رتبه دوم، و برای بسیاری از وظایف موثر ثابت شده است. با این حال، مدل‌سازی فقط تعاملات ویژگی با رتبه پایین، بهبود عملکرد را محدود می‌کند.

(Zhang, 2022) روش نمونه‌گیری مجدد را برای فرآیند آموزش مدل‌های یادگیری گروهی اعمال می‌کنند، در مطالعات آن‌ها، هر طبقه‌بندی‌کننده پایه با یک زیرمجموعه متوازن مجدد آموزش داده می‌شود. گسترش ریسک اعتباری شرکت در زنجیره تامین ممکن است منجر به ورشکستگی و بحران اعتباری در مقیاس بزرگ شود که با ثبات اقتصادی و اجتماعی ملی و امنیت سیستم مالی مرتبط است. بنابراین، ریسک اعتباری شرکت در زمینه زنجیره تامین نه تنها برای مؤسسات مالی بانکی، مؤسسات رتبه‌بندی اعتباری و مدیران شرکت‌ها، بلکه مورد توجه دولت‌ها است. این مقاله یک چارچوب پیش‌بینی DTE-DSA مدل درخت تصمیم با استفاده از نرخ نمونه‌گیری تفاضلی، روش نمونه‌برداری بیش از حد اقلیت مصنوعی SMOTE و AdaBoost ایجاد می‌کند که اطلاعات زنجیره تامین را برای پیش‌بینی ریسک اعتباری شرکت یکپارچه می‌کند. آزمون تجربی نشان می دهد که استفاده از اطلاعات زنجیره تامین می تواند به طور قابل توجهی امتیاز پیش بینی را بهبود بخشد. مدل DTE-DSA بهترین اثر پیش بینی را در برخورد با مشکلات عدم تعادل کلاس دارد. چارچوب آن‌ها به نتایج ایده آل در مجموعه داده‌های امتیازدهی اعتبار نامتعادل دست می‌یابد.
(Zhang, Chi, 2021) روش انتخاب تطبیقی را برای طبقه‌بندی‌کننده پایه با توجه به مقیاس مجموعه داده‌های امتیازدهی اعتبار نامتعادل در مدل‌های مجموعه پیاده‌سازی می‌کند. در حوزه امتیازدهی اعتبار، تعداد مشتریان بد به مراتب کمتر از مشتریان خوب

است. بنابراین طبقه بندی نامتعادل داده‌ها یک موضوع واقعی و حیاتی در فرآیند امتیازدهی اعتباری است. در این مطالعه، یک مدل جدید امتیازدهی اعتباری مجموعه ناهمگن برای مشکل طبقه‌بندی داده‌های نامتعادل پیشنهاد شده‌است. این مدل پیشنهادی بر اساس پنج طبقه‌بندی‌کننده استاندارد که به‌طور تطبیقی طبقه‌بندی‌کننده‌های پایه با بالاترین AUC را با توجه به توزیع داده‌ها انتخاب می‌کند، سپس همه طبقه‌بندی‌کننده‌های پایه را برای به دست آوردن یک پیش‌بینی ادغام می‌کند. در نهایت، با استفاده از پنج معیار عملکرد جامع و چهار مجموعه داده اعتباری کلاسیک، متوجه می‌شوند که مدل پیشنهادی بهتر از سایر مدل‌های پایه است. این مدل جدید را می توان برای امتیازدهی واقعی اعتبار اعمال کرد و به موسسات مالی در مدیریت ریسک اعتباری کمک کرد. (Carta, 2020) یک چارچوب انتخاب نمونه مبتنی بر آنتروپی را برای حل مشکل داده‌های ناکافی و نامتعادل در اعتبار مصرف‌کننده طراحی کردند که در یادگیری گروهی حساس به هزینه را در نظر می گیرند. با تنظیم قانون جریمه برای طبقه‌بندی اشتباه، مدل‌ها می‌توانند بر روی نمونه‌های پیش‌فرض تمرکز کرده و پیش‌بینی‌های بی‌طرفانه را تولید کنند.

۳. روش پیشنهادی

در این مقاله با استفاده از الگوریتم های LGB, CATBOOST, XGBOOST,ADABOOST, RF با استفاده از مجموعه داده Give Me Some Credit [1] به اعتبار سنجی می‌پردازیم.

۱.۳ الگوریتم های یادگیری ماشین

در این قسمت به بررسی الگوریتم‌های یادگیری ماشین می‌پردازیم:

LightGBM یک چارچوب تقویت گرادیان است که از الگوریتم‌های درخت تصمیم استفاده می‌کند. همانطور که از نام این رویکرد مشخص است، سرعت آموزش آن بسیار بالاست. علاوه بر این، قابلیت استفاده روی مجموعه دادگان آموزشی بزرگ را نیز دارد.

XGBoost یک پیاده‌سازی پیشرفته از الگوریتم‌های تقویت گرادیان است. این الگوریتم از نظر محاسبات با تقویت گرادیان تفاوت دارد. زیرا تکنیک نرمال‌سازیرا به صورت داخلی اجرا می‌کند. به بیان ساده، XGBoostبه تکنیک تقویت منظم اشاره دارد.

CatBoost (Chen, 2021) ، مخفف CatBoosting، الگوریتمی است که بر پایه درختان تصمیم و تقویت گرادیان مانند XGBoost اما با عملکرد بهتراست. CatBoost در یک شرکت روسی به نام Yandex سرچشمه گرفت. این یکی از جدیدترین الگوریتم‌های تقویتی است که در سال ۲۰۱۷ در دسترس قرار گرفت. الگوریتم‌های تقویت کننده زیادی مانند XGBoost، LightGBM و غیره وجود داشت، اما هیچ کدام به دلایل متعدد به CatBoost نزدیک نمی‌شوند. الگوریتم CatBoost الگوریتم دیگری است که بر پایه گئورنت دسنت بنا شده‌است و تفاوت‌های ظریفی دارد که آن را منحصر به فرد می‌سازد. الگوریتم CatBoostدرختان متقارن را پیاده‌سازی می‌کند که به کاهش زمان پیش‌بینی کمک می‌کنند و همچنین عمق درخت کم‌عمق‌تر به طور پیش‌فرض دارد. الگوریتم CatBoost یکی دیگر از اعضای تکنیک تقویت گرادیان در درختان تصمیم است. یکی از بسیاری از ویژگی‌های منحصربه‌فردی که الگوریتم CatBoost ارائه می‌دهد، ادغام کار با انواع داده‌های مختلف برای حل طیف گسترده‌ای از

[1] https://www.openml.org/d/45577

مشکلات داده‌ای است که مشاغل متعدد با آن مواجه هستند. یک درخت CatBoost آموزش دیده می تواند بسیار سریعتر از XGBoost یا LightGBM پیش بینی کند. از طرف دیگر، برخی از شناسایی داخلی CatBoost از داده های طبقه بندی شده زمان آموزش آن را به طور قابل توجهی در مقایسه با XGBoost کند می کند، اما هنوز هم بسیار سریعتر از XGBoost گزارش می شود. سرعت CatBoost با پشتیبانی از GPUهای توزیع شده چند سرور (که میزبان های متعدد را برای یادگیری سریع فعال می کند) و با استفاده از GPU های قدیمی تر، مقیاس پذیری را ارائه می دهد. برخی از معیارهای سرعت آموزش CPU و GPU را روی مجموعه داده های بزرگی مانند Epsilon و Higgs تنظیم کرده است. به طور کلی، CatBoost یک الگوریتم بسیار سریع، دقیق و مبتکرانه است، اما به نوعی مانند پیشینیان خود مانند XGBoost به طور گسترده مورد استفاده قرار نمی گیرد.

AdaBoost تقویت کنندگی وفقی است. اساساً ، AdaBoosting اولین الگوریتم تقویت کننده موفقیت آمیز برای کلاسه بندی باینری بود. همچنین، بهترین نقطه شروع برای درک مفهوم تقویت کنندگی است. علاوه بر این، روش‌های تقویت کننده مدرن بر اساس AdaBoost ساخته می‌شوند. به طور کلی، AdaBoost با درختان تصمیم گیری کوچک مورد استفاده قرار می‌گیرد. بدین صورت که درخت اول ایجاد می‌شود و از عملکرد آن روی هر نمونه آموزشی، برای سنجش میزان توجه درخت بعدی به نمونه‌ها استفاده می‌شود. بنابراین، درخت باید به تمام نمونه‌های آموزشی توجه کند و به داده های آموزشی که پیش بینی آن ها دشوار است، وزن بیشتری بدهد، در حالی که به داده‌هایی که پیش بینی آن ها آسان است وزن کمتری بدهد.

۳.۲ دیتاست

بانک‌ها نقش مهمی در اقتصاد بازار دارند. آن‌ها تصمیم می‌گیرند که چه کسی و با چه شرایطی می‌تواند کمک مالی دریافت کند و می‌تواند تصمیمات سرمایه گذاری را بگیرد یا شکست دهد. برای عملکرد بازارها و جامعه، افراد و شرکت‌ها نیاز به دسترسی به اعتبار دارند. الگوریتم‌های امتیازدهی اعتباری که احتمال نکول را حدس می‌زنند، روشی است که بانک‌ها برای تعیین اینکه آیا وام باید اعطا شود یا نه. این مسابقه از شرکت‌کنندگان می‌خواهد تا با پیش‌بینی احتمال بروز مشکلات مالی در دو سال آینده، وضعیت هنر در امتیازدهی اعتبار را بهبود بخشند. در ادامه به ویژگی های این دیتاست میپردازیم.

- SeriousDlqin2yrs : فرد ۹۰ روز گذشته یا بدتر از آن را تجربه کرده است.
- MonthlyIncome: درآمد ماهانه
- Number O fDependents: تعداد افراد تحت تکفل خانواده به استثنای خود (همسر، فرزندان و غیره).
- Age: سن وام گیرنده بر حسب سال
- DebtRatio: پرداخت ماهیانه بدهی، نفقه، هزینه‌های زندگی تقسیم بر درآمد ناخالص ماهانه.
- Revolving Utilization O fUnsecured Lines: کل موجودی کارت‌های اعتباری و خطوط اعتباری شخصی به جز مستغلات و بدون بدهی اقساطی مانند وام خودرو تقسیم بر مجموع محدودیت های اعتباری.
- Number Real EstateLoansOrLines: تعداد وام‌های رهنی و املاک و مستغلات از جمله خطوط اعتباری سهام خانه.
- Number Of Open Credit Lines And Loans: تعداد وام‌های باز (اقساط مانند وام خودرو یا وام مسکن) و خطوط اعتباری (مانند کارت های اعتباری).

- Number Of Time 30-59 DaysPastDueNotWorse: تعداد دفعاتی که وام گیرنده ۳۰-۵۹ روز از سررسید گذشته بوده است اما در ۲ سال گذشته بدتر نشده است.
- Number Of Time 60-89Days PastDue NotWorse: تعداد دفعاتی که وام گیرنده ۶۰-۸۹ روز از سررسید گذشته بوده است اما در ۲ سال گذشته بدتر نشده است.
- Number Of Times90 Days Late: تعداد دفعاتی که وام گیرنده ۹۰ روز یا بیشتر از سررسید گذشته بوده است

۴. ارزیابی نتایج

در این قسمت به بررسی حاصل از پیاده سازی الگوریتم‌های یادگیری ماشین بر روی دیتاست مورد نظر می‌پردازیم.

۱.۴ پارامترهای ارزیابی

هر نمونه یا فردی در واقعیت، متعلق به یکی از کلاس‌های مثبت یا منفی است و از سوی دیگر، از هر الگوریتم که برای دسته‌بندی داده‌ها استفاده شود، در نهایت هر نمونه عضو یکی از این دو دسته‌بندی خواهد شد. بنابراین برای هر نمونه داده، یکی از چهار حالتی که در ادامه بیان شده، ممکن است اتفاق بیفتد.

- نمونه عضو دسته مثبت باشد و عضو همین کلاس تشخیص داده شود(مثبت صحیح یا True Positive)
- نمونه عضو کلاس مثبت باشد و عضو کلاس منفی تشخیص داده شود(منفی کاذب یا False Negative)
- نمونه عضو کلاس منفی باشد و عضو همین کلاس تشخیص داده شود (منفی صحیح یا True Negative)
- و در نهایت، نمونه عضو کلاس منفی باشد و عضو کلاس مثبت تشخیص داده شود(مثبت کاذب یا False Positive)

برای ارزیابی یک مدل، معیارهای بسیار زیادی موجود است که هر کدام در موارد خاص از اهمیتی بیشتری برخوردار هستند. معیار امتیاز F1 در مجموعه دادگانی که متعادل نیستند و کلاس‌های آن ها از تعداد اعضای برابری تشکیل نشده است اهمیت زیادی پیدا می‌کند. ما در این مقاله از ۴ معیار ارزیابی استفاده کرده‌ایم که توضیحات آن ها به شرح زیر است:

صحت این معیار رایج ترین معیار مورد استفاده از در ارزیابی روش‌های دسته بندی است و به صورت رایج در ارزیابی روش‌های یادگیری ماشین گزارش میشود و درصد دادگانی که درست دسته بندی شده اند را نشان می‌دهد. روش محاسبه آن به صورت زیر است:

(۱)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + TN}$$

دقت این معیار به معنی این است که چه درصد از دادگانی که به عنوان اعضای یک کلاس تشخیص داده شده‌اند در واقع متعلق به آن کلاس بوده اند. این معیار به صورت زیر محاسبه می‌شود:

(۲)

$$Precision = \frac{TP}{TP + FP}$$

بازخوانی این معیار نشان دهنده این است که چه درصدی از دادگان یک کلاس، تشخیص داده شده‌اند. این معیار به صورت زیر می‌باشد:

(۳)

$$Recall = \frac{TP}{TP + FN}$$

امتیاز F۱ این معیار میانگین همساز Recall و Precision است. این معیار به صورت زیر محاسبه می‌شود:

(۴)

$$F1 - Score = 2\frac{Precision \times Recall}{Recall + Precision}$$

۲.۴ نتایج

در ابتدا به پیش پردازش داده ها پرداختیم که در شکل ۱ درصد داده ها train و در شکل ۲ داده های test را نشان میدهد.

	Null Values	% Missing Values
MonthlyIncome	29731	19.820667
NumberOfDependents	3924	2.616000

شکل۱: داده‌های train

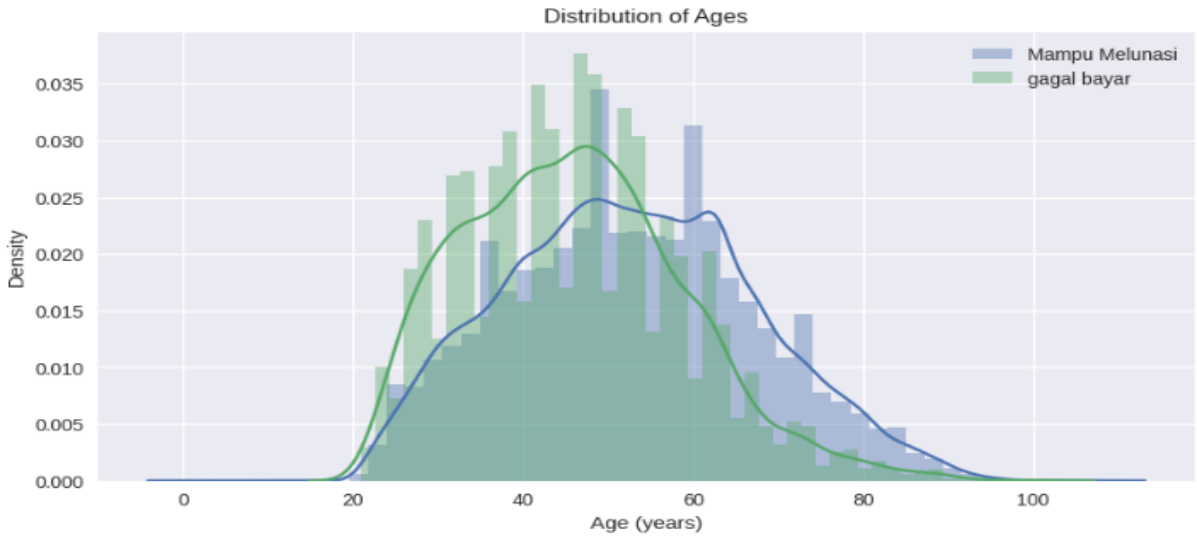
	Null Values	% Missing Values
SeriousDlqin2yrs	101503	100.000000
MonthlyIncome	20103	19.805326
NumberOfDependents	2626	2.587116

شکل۲: داده‌های test

برای ارتباط بین متغیرهای با همبستگی به عنوان یک مقدار بین ۱- و۱+ نشان داده می‌شود که در آن۱+ نشان دهنده بالاترین همبستگی مثبت، ۱- نشان دهنده بالاترین همبستگی منفی و ۰ نشان دهنده عدم همبستگی است که در شکل ۳ به این موضوع پرداخته شده است. در شکل ۴ هم به عنوان نمونه به بررسی ویژگی سن پرداختیم.

```
Most Correlations:
 age                                   -0.115386
NumberOfOpenCreditLinesAndLoans        -0.029669
MonthlyIncome                          -0.017151
DebtRatio                              -0.007602
NumberRealEstateLoansOrLines           -0.007038
RevolvingUtilizationOfUnsecuredLines   -0.001802
Unnamed: 0                              0.002801
NumberOfDependents                      0.046869
NumberOfTime60-89DaysPastDueNotWorse    0.102261
NumberOfTimes90DaysLate                 0.117175
NumberOfTime30-59DaysPastDueNotWorse    0.125587
SeriousDlqin2yrs                        1.000000
```

شکل۳: همبستگی داده‌ها



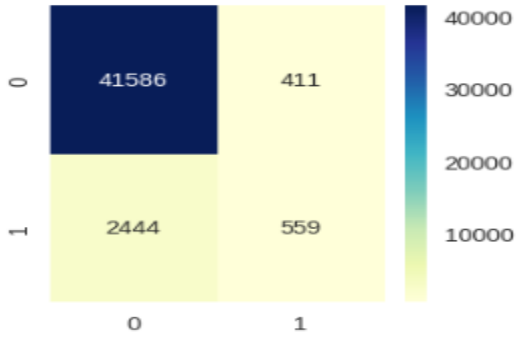
شکل ۴: بررسی ویژگی سن

در شکل ۵ دیتاست آماده شده رو به عنوان ورودی به الگوریتم‌ها می‌دهیم تا خروجی مورد نظر که دقت بالا می‌باشد را بدست آوریم.

#	Column	Non-Null Count	Dtype
0	Unnamed: 0	150000 non-null	int64
1	SeriousDlqin2yrs	150000 non-null	int64
2	RevolvingUtilizationOfUnsecuredLines	150000 non-null	float64
3	age	150000 non-null	int64
4	NumberOfTime30-59DaysPastDueNotWorse	150000 non-null	int64
5	DebtRatio	150000 non-null	float64
6	MonthlyIncome	150000 non-null	float64
7	NumberOfOpenCreditLinesAndLoans	150000 non-null	int64
8	NumberOfTimes90DaysLate	150000 non-null	int64
9	NumberRealEstateLoansOrLines	150000 non-null	int64
10	NumberOfTime60-89DaysPastDueNotWorse	150000 non-null	int64
11	NumberOfDependents	150000 non-null	float64

شکل۵: دیتاست نهایی

در شکل ۶ ماتریس درهم ریختگی حاصل از پیاده سازی الگوریتم LGB و در شکل ۷ پارامترهای ارزیابی حاصل از این الگوریتم را نمایش می‌دهیم.



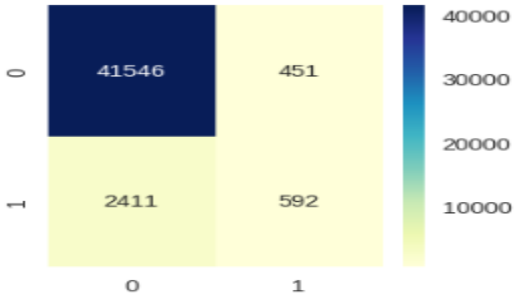
شکل ۶: ماتریس درهم‌ریختگی

	precision	recall	f1-score	support
0	0.94	0.99	0.97	41997
1	0.58	0.19	0.28	3003
accuracy			0.94	45000
macro avg	0.76	0.59	0.62	45000
weighted avg	0.92	0.94	0.92	45000

شکل ۷: پارامترهای ارزیابی

در شکل ۸ ماتریس درهم ریختگی حاصل از پیاده سازی الگوریتم XGboost و در شکل ۹ پارامترهای ارزیابی حاصل از این الگوریتم را نمایش می‌دهیم.

```
Training Accuracy : 0.9437809523809524
Testing Accuracy : 0.9364
```



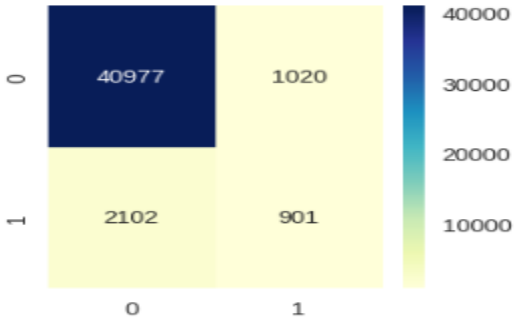
شکل ۸: ماتریس درهم ریختگی

	precision	recall	f1-score	support
0	0.95	0.99	0.97	41997
1	0.57	0.20	0.29	3003
accuracy			0.94	45000
macro avg	0.76	0.59	0.63	45000
weighted avg	0.92	0.94	0.92	45000

شکل ۹: پارامترهای ارزیابی

در شکل ۱۰ ماتریس درهم ریختگی حاصل از پیاده سازی الگوریتم CATBOOST و در شکل ۱۱ پارامترهای ارزیابی حاصل از این الگوریتم را نمایش می‌دهیم.

```
Training Accuracy : 0.9648
Testing Accuracy : 0.9306222222222222
```



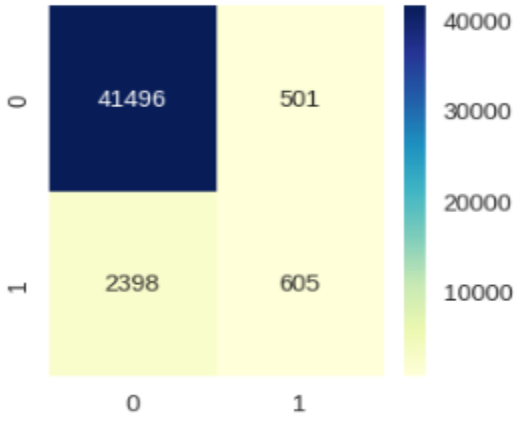
شکل ۱۰: ماتریس درهم ریختگی

	precision	recall	f1-score	support
0	0.95	0.98	0.96	41997
1	0.47	0.30	0.37	3003
accuracy			0.93	45000
macro avg	0.71	0.64	0.66	45000
weighted avg	0.92	0.93	0.92	45000

شکل ۱۱: پارامترهای ارزیابی

در شکل ۱۲ ماتریس درهم ریختگی حاصل از پیاده سازی الگوریتم ADABOOST و در شکل ۱۳ پارامترهای ارزیابی حاصل از این الگوریتم را نمایش می‌دهیم.

```
Training Accuracy : 0.9366190476190476
Testing Accuracy : 0.9355777777777777
```



شکل ۱۲: ماتریس درهم ریختگی

	precision	recall	f1-score	support
0	0.95	0.99	0.97	41997
1	0.55	0.20	0.29	3003
accuracy			0.94	45000
macro avg	0.75	0.59	0.63	45000
weighted avg	0.92	0.94	0.92	45000

شکل ۱۳: پارامترهای ارزیابی

5. نتایج

بانک‌ها در اعطای تسهیلات اعتباری به مشتریان خود نیازمند اعتبارسنجی آن‌ها هستند. اعتبارسنجی مشتریان یک فرآیند مهم است که بانک‌ها و مؤسسات مالی انجام می‌دهند تا اعتبار مالی و قدرت بازپرداخت بدهی مشتریان خود را ارزیابی کنند. این فرآیند شامل

بررسی تاریخچه اعتباری، درآمدها، دارایی‌ها، بدهی‌ها، سابقه پرداخت و دیگر عوامل مرتبط با وضعیت مالی مشتری می‌شود. توجه به این نکته حائز اهمیت است که اعتبارسنجی به منظور کاهش ریسک اعطای تسهیلات اعتباری و همچنین حفظ سلامت مالی بانک یا مؤسسه مالی انجام می‌شود. این فرآیند می‌تواند شامل بررسی اسناد مالی، مصاحبه شخصی، استفاده از اطلاعات موجود در سیستم‌های اطلاعاتی مرکزی و یا هر ابزار دیگری که بانک یا مؤسسه مالی مورد استفاده قرار دهد، باشد. اعتبارسنجی مشتریان نهایتاً به جلوگیری از وام‌دهی به افرادی که قادر به پرداخت بازپرداخت نیستند و به عدم توسعه بدهی‌های بدون امانت منجر می‌شود. الگوریتم‌های یادگیری ماشین می‌توانند در این زمینه به طبقه بندی مشتریان بپردازند. هدف، ارایه یک مدل مناسب اعتبارسنجی مشتریان بانک‌ها مانند برای اعطای تسهیلات اعتباری متناسب با هر طبقه بود. این مدل در قالب فرآیند توسعه در شناخت الگو و فرآیند به ساخت درخت تصمیم گیری نهایی برای اعتبارسنجی مشتریان بانک پرداخت است. با توجه به موارد گفته شده می‌توان از مدل طبقه بندی پیشنهادی برای ساخت و آزمون به منظور اعتبارسنجی مشتریان بانک استفاده نمود. هدف اصلی این پژوهش رسیدن به دقت بالا بوده است که در همه الگوریتم‌ها این حاصل شده و به دقت بالا 90 رسیده‌ایم که این نشان از برتری مدل پیشنهادی میدهد.

منابع

Jiang, C., Lu, W., Wang, Z., & Ding, Y. (2023). Benchmarking state-of-the-art imbalanced data learning approaches for credit scoring. Expert Systems with Applications, 213, 118878.Wu,

J., Zhang, Z., & Zhou, S. X. (2022). Credit rating prediction through supply chains: A machine learning approach. Production and Operations Management, 31(4), 1613-1629.

Shen, F., Yang, Z., Zhao, X., & Lan, D. (2022). Reject inference in credit scoring using a three-way decision and safe semi-supervised support vector machine. Information sciences, 606, 614-627.

Guo, W., Yang, Z., Wu, S., Wang, X., & Chen, F. (2023). Explainable enterprise credit rating using deep feature crossing. Expert Systems with Applications, 220, 119704.

Smirnov, V. S., & Stupnikov, S. A. (2023). A Deep Learning Approach to Credit Scoring Using Credit History Data. Lobachevskii Journal of Mathematics, 44(1), 198-204.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2022). Smote: Synthetic mi- nority over-sampling technique. Journal of artificial intelligence research, 16, 321–357.

Cohen, L., & Frazzini, A. (2008). Economic links and predictable returns. The Journal of Finance, 63(4), 1977-2011.

Hertzel, M. G., Li, Z., Officer, M. S., & Rodgers, K. J. (2008). Inter-firm linkages and the wealth effects of financial distress along the supply chain. Journal of Financial Economics, 87(2), 374-387.

Kouvelis, P., & Zhao, W. (2018). Who should finance the supply chain? Impact of credit ratings on supply chain decisions. Manufacturing & Service Operations Management, 20(1), 19-35.

Acemoglu, D., Ozdaglar, A., & Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. American Economic Review, 105(2), 564-608.

Herskovic, B. (2018). Networks in production: Asset pricing implications. The Journal of Finance, 73(4), 1785-1818.

Wu, W., Chen, J., Yang, Z., & Tindall, M. L. (2021). A cross-sectional machine learning approach for hedge fund return prediction and selection. Management Science, 67(7), 4577-4601..

Hosaka, T. (2019). Bankruptcy prediction using imaged financial ratios and convolutional neural networks. Expert systems with applications, 117, 287-299.

Zhang, T., & Chi, G. (2021). A heterogeneous ensemble credit scoring model based on adaptive classifier selection: An application on imbalanced data. International Journal of Finance & Economics, 26(3), 4372- 4385.

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.

Chen, Y., & Han, X. (2021, January). CatBoost for fraud detection in financial transactions. In 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE) (pp. 176-179). IEEE.

Applying machine learning algorithms to validate bank customers

Hossein faghih aliabadi

Master's degree in software engineering, computer networks, Urmia University

Ali Ghanbarizadeh

Master's degree in information technology, e-commerce major, Amirkabir University

1-1- Abstract

In recent years, the granting of bank facilities has faced problems that have necessitated the use of validation systems for banks. By using the analysis of information related to customers, banks can use the data mining process to validate the applicants for facilities and classify them as good or bad credit customers. In traditional banking systems, credit managers often measure the credit of customers according to their experience, but in the modern banking system, we are faced with time constraints and increasing number of customers. To solve this problem, in this article, we use machine learning algorithms to calculate the probability of default or non-default in payments and rank the customers according to that. The aim of this research is to develop an integrated model with high accuracy for customer validation. This has been achieved in all algorithms and we have reached a high accuracy of 90%, which shows the superiority of the proposed model. Neural networks have priority in predicting the credit behavior of people applying for facilities due to their much higher accuracy and lower calculation volume compared to other classical methods.

Keywords: Feature selection, validation, bank, artificial intelligence, machine learning