

Credit Risk Management of Banking Customers Using Support Vector Machine Optimized by Genetic Algorithm with Data Mining Approach

Meisam Jafari Eskandari^{1*}, Milad Roohi²

1- Associate Prof., Faculty of Industrial Engineering, Industrial Engineering Department, Payame Noor University, Tehran, Iran
meisam_jafari@pnu.ac.ir

2- MSc. Industrial Engineering Department, Faculty of Industrial Engineering Payame Noor University, Tehran, Iran
milad.roohi@gmail.com

Abstract

Credit risk management, credit scoring and risk assessment of customers is an important issue in banking industry. Credit scoring is important because if the banks fail to earn resource allocation and create a balance between the processes of mobilizing and allocating resources, they are typically faced with many problems in the future. According to official figures released by the Central Bank of Iran in recent years, the rate of bad loans increased, since the systemic strict validation to evaluate and measure the credit risk of customers do not exist. In this paper, we try to predict customer recovery rate index with data mining techniques. Markedly, in recent years in the world new method for measuring customer risk rather than default probability measure has been considered, but due to low precision of forecasting models widely different approaches in research and modeling is investigated. The method used in this paper is support vector regression model whose parameters selection is optimized with genetic algorithm.

Keywords: Credit risk, Recovery rates, Rialclaims, Support vector machine

مدیریت ریسک اعتباری مشتریان بانکی با استفاده از روش ماشین بردار تصمیم بهبود یافته با الگوریتم ژنتیک با رویکرد داده کاوی

میشم جعفری اسکندری^{۱*}، میلاد روحی^۲

۱- استادیار گروه مهندسی صنایع، دانشکده مهندسی صنایع، دانشگاه پیام نور تهران، ایران

meisam_jafari@pnu.ac.ir

۲- کارشناسی ارشد گروه مهندسی صنایع، دانشکده مهندسی صنایع، دانشگاه پیام نور تهران، ایران

milad.roohi@gmail.com

چکیده

مدیریت ریسک اعتباری، رتبه‌بندی اعتباری و ارزیابی میزان ریسک مشتریان، در کنار جذب منابع از اهمیت بالایی برای بانک‌ها برخوردار است؛ زیرا اگر بانک‌ها با تخصیص بهینه منابع و کسب درآمد بین فرایند تجهیز و تخصیص منابع خود نتوانند توازن ایجاد کنند، در آینده با مشکلات زیادی روبه‌رو می‌شوند. براساس آمارهای رسمی منتشر شده از سوی بانک مرکزی ج.ا.ی. در سال‌های اخیر، میزان مطالبات معوق بانک‌ها بسیار افزایش یافته است؛ زیرا سیستم اعتبارسنجی دقیقی برای ارزیابی اعتبار و اندازه‌گیری میزان ریسک مشتریان وجود ندارد. در این پژوهش، الگوریتمی با استفاده از روش‌های داده کاوی برای پیش‌بینی شاخص نرخ وصول مشتریان ارائه می‌شود. رویکردی که در سال‌های اخیر در دنیا به‌عنوان روشی جدید برای اندازه‌گیری ریسک مشتریان به‌جای اندازه‌گیری احتمال نکول مد نظر قرار گرفته است. نتایج نشان می‌دهد الگوی پیشنهادی این پژوهش، دقت بیشتری دارد. به‌طور کلی، هدف پیش‌بینی درصد وصول مطالبات قراردادهای با احتمال ریسک مطالباتی بالا قبل از اعطای تسهیلات است.

واژه‌های کلیدی: ریسک اعتباری، ماشین بردار تصمیم، مطالبات ریالی، نرخ وصول

مقدمه

بلکه از این منظر بررسی می‌شوند که اگر فرضاً احتمال نکول یا بازپرداخت نکردن به وجود بیاید، نرخ وصول مطالبات چه میزان خواهد بود؟ و یا به عبارتی، چند درصد از بدهی مشتری وصول می‌شود، در این پژوهش به پیش‌بینی این شاخص توجه شده است. براساس نتایج و الگوی به‌دست آمده و اجرای آن بر تسهیلات جاری و مطالبات‌نشده، مشتریان را می‌توان دسته‌بندی کرد و به مشتریان خوش حساب یا غیر مطالباتی تسهیلات با نرخ سود کمتر اعطا و مشتریان بدحساب را به تدریج، ابتدا از اعطای تسهیلات مجدد جلوگیری و با تدوین برنامه مدونی براساس میزان وثایق دریافتی برای وصول مطالبات آنها اقدام کرد. از این نتایج برای راه‌اندازی سامانه اعطای تسهیلات به صورت هوشمند و تسری آن به تمامی واحدهای بانک هنگام اعطای تسهیلات و الگوسازی رفتار مشتری جدید براساس ویژگی‌های اعتباری و مقایسه با مشتریان گذشته با ویژگی مشابه می‌توان استفاده کرد.

پرسش اساسی این پژوهش، یافتن پارامترهای بهینه الگوهای پیش‌بینی مجموعه SVR برای پیش‌بینی میزان مطالبات قراردادهای با ریسک زیاد قبل از اعطای تسهیلات است. هدف پژوهش، ارائه الگوی جامعی برای پیش‌بینی درصد میزان وصول مطالبات قراردادهای با ریسک زیاد است و نوآوری پژوهش در مقایسه با پژوهش‌های داخلی در درجه اول پیش‌بینی نرخ وصول مشتریان اعتباری در صورت مطالبات شدن برای اولین بار با استفاده از داده‌کاوی در ایران و همچنین پیش‌بینی مبلغ زیان ناشی از وصول نشدن قراردادهای با نرخ وصول کمتر از ۱ به‌جای صرفاً طبقه‌بندی مشتریان به بخش‌های مختلف است که به مورد دوم در پژوهش‌های قبلی در داخل کشور

در سال‌های اخیر، توانایی تولید، ضبط و ذخیره داده‌ها بسیار افزایش یافته است. اطلاعاتی که در این داده‌ها می‌تواند نهفته باشد، بسیار مهم است. در دسترس بودن حجم بالای داده و نیاز به تبدیل آنها به دانش، صنعت فناوری اطلاعات را برای استفاده از داده‌کاوی تشویق کرده و به سمت آن سوق داده است. صنعت بانکداری در مسیر کسب و کار خود در سراسر جهان، دستخوش تغییر فوق‌العاده‌ای شده است. همچنین این صنعت شروع به شناخت تکنیک‌ها و مهارت‌های استفاده از داده‌کاوی برای استفاده در رقابت در بازار بانکی کرده است. بانک‌ها با استفاده از ابزار داده‌کاوی به بخش‌بندی مشتریان، مطالعه سوددهی، رتبه‌بندی اعتباری و پیش‌بینی پرداخت تسهیلات و وصول مطالبات، بازاریابی، شناسایی و کشف تقلب در تراکنش‌ها و غیره توجه کرده‌اند [۱۰]. مؤسسات اعتباری برای پیش‌بینی وضعیت افرادی که در آینده از عهده انجام تعهدات خود بر نخواهند آمد، مایل به ارزیابی مشتریان هستند. در هر دو حالت (بررسی درخواست‌های وام جدید و کنترل وام‌گیرندگان قبلی) احتمال بازپرداخت در دوره وام‌دهی تخمین زده می‌شود و در نتیجه، مشتریان براساس تخمین حاصل درباره ناتوانی در بازپرداخت به سطوح متفاوتی از ریسک رتبه‌بندی خواهند شد. این روش به‌عنوان تعیین ریسک و یا طبقه‌بندی اعتبار شناخته می‌شود [۱۴]. در سال‌های اخیر در مباحث مربوط به ریسک اعتباری، مفهوم تازه‌تری برای پیش‌بینی مدنظر قرار گرفته است و آن شاخصی با نام نرخ وصول^۱ است؛ به زبان ساده‌تر، صرفاً مشتریان براساس میزان ریسک یا احتمال بازپرداخت نکردن وام طبقه‌بندی و رتبه‌بندی نمی‌شوند؛

به دفعات توجه شده است؛ زیرا در حال حاضر، بانک‌ها در شرایطی مجبور به پرداخت تسهیلات به برخی مشتریان هستند و صرفاً قرار گرفتن مشتری در طبقه پرریسک، ملاک پرداخت نکردن تسهیلات نیست و باید با پیش‌بینی دقیق مشخص شود در صورتی که تسهیلات به آنها اعطا شود، در آینده چه میزان از مبلغ بدهی مشتری با گذراندن مراحل قانونی قابل وصول خواهد بود.

مبانی نظری

در مطالعات لوترمن^۱، به الگوریتم‌های مختلف رگرسیون در الگوسازی شاخص درصد مطالبات قابل وصول توجه شده است. ۲۴ روش الگوسازی براساس الگوریتم‌های رگرسیون خطی، لجستیک، کمترین مربعات و ... بررسی و همچنین با بررسی روش‌های دیگر نظیر شبکه عصبی و ماشین بردار تصمیم مشخص شد که این دو روش نسبت به روش‌های سنتی خطی، کارایی بیشتری دارند [۸]. با بررسی کاربردهای تحلیل تکنیک‌های ابقا زمانی در الگوسازی شاخص مذکور و روش‌هایی نظیر رگرسیون کاکس، خطی و لجستیک برای پیش‌بینی این شاخص استفاده شده است [۱۵]. با مقایسه الگوهای رگرسیون برای تخمین شاخص نرخ وصول با مقایسه الگوهای مختلف با تأکید بر الگوهای رگرسیونی نظیر رگرسیون ساده، لجستیک، درخت تصمیم و ... می‌توان نتیجه گرفت با توجه به نوع داده‌ها هیچ کدام از الگوها لزوماً به صورت در خور توجهی بهتر از دیگر الگوها نیستند [۲]. اطلاعات حسابداری برای وام‌دهندگان در اقسام قراردادهای بدهی، آنها را در تخصیص مناسب شاخص مذکور به مشتریان می‌تواند مجهز کند. با داشتن اطلاعات حسابداری

قراردادهای بدهی مشتریان ۴۷ ماه قبل از نکول، درصد مطالبات غیر قابل وصول را به صورت طبیعی قبل از وقوع می‌توان به دست آورد. همچنین مشخص می‌شود افزایش نرخ بهره وام، رابطه مستقیمی با این شاخص و میزان بدهی نیز ارتباط زیادی دارد [۱]. یاو^۲ و همکاران (۲۰۱۵) با استفاده از روش رگرسیون بردار تصمیم به الگوسازی شاخص درصد مطالبات وصول‌ناپذیر گرفتند. در این پژوهش، تکنیک مذکور و الگوریتم دیگر بررسی و در انتها، مشخص شد این تکنیک بسیار اعتمادپذیرتر از سایر روش‌های الگوسازی شاخص مذکور است [۱۶]. در پژوهش گرتلر^۳ و هیبلن^۴ (۲۰۱۳) بهبود پیش‌بینی شاخص درصد مطالبات غیر قابل وصول بانک بررسی شد. در این مطالعه با استفاده از روش‌هایی نظیر نمونه‌گیری مغرضانه، مشخصه‌های متفاوت وامی با توجه به نوع پایان دوره مطالبات و تنظیمات اطلاعاتی متفاوت براساس وضعیت نکول بررسی شد [۵]. باستوس^۵ (۲۰۱۰) برای پیش‌بینی درصد مطالبات غیر قابل وصول وام‌ها در بانک، شاخص مذکور را با استفاده از پیش‌بینی نرخ وصول بررسی و از روش رگرسیون در بخش پارامتریک و درخت تصمیم در بخش ناپارامتریک استفاده کرده است. در نتایج پژوهش نشان داده شده است درخت‌های تصمیم، جایگزین مناسبی برای روش‌های پارامتریک در الگوسازی شاخص درصد مطالبات وصول‌ناپذیر هستند [۳]. حاجی کرد و همکاران (۱۳۹۵) با استفاده از الگوی ماشین بردار تصمیم و الگوی هیبریدی الگوریتم ژنتیک برای پیش‌بینی ریسک اعتباری و تقسیم‌بندی مشتریان به دو دسته خوش حساب و بدحساب استفاده کرده‌اند که

2 Yao, X

3 Gürtler, M

4 Hibbeln, M

3 Bastos, J A

1 Loterman, G

نتیجه نشان داد الگوی بهینه‌سازی شده ماشین بردار تصمیم با الگوریتم ژنتیک، تأثیر بهتری در پیش‌بینی ریسک اعتباری و دسته‌بندی مشتریان به خوش حساب و بدحساب دارد [۹]. کرانی و آقایی‌پور (۱۳۹۳) نظریه تحلیل بقا در مدیریت ریسک اعتباری دریافت کنندگان تسهیلات (مطالعه موردی: بانک مسکن) را بررسی و احتمال‌های نکول آن را براساس الگوی خطرهای متناسب کاکس و برآوردگر حد حاصل‌ضربی تعمیم‌یافته برآورد کرده‌اند [۷]. نظرپور و رضایی (۱۳۹۲) عقود اسلامی و الگوی پرداخت تسهیلات را بررسی کردند و دریافتند عقود اسلامی غیرمشارکتی (مبادله‌ای) در مقایسه با عقود مشارکتی، ریسک کمتری در بردارد؛ اما این عقود نیز ریسک است و بانک‌های اسلامی را در معرض ریسک اعتباری قرار می‌دهد و بانک‌ها در صورت استقرار نداشتن یک نظام مدیریت ریسک اعتباری متناسب، میزان ریسک اعتبارات و تعیین زیان‌های احتمالی بازپرداخت نکردن وام‌ها را نمی‌توانند تشخیص بدهند و در نتیجه، سرمایه خود را نخواهند توانست تخصیص بهینه کنند [۱۱].

روش پژوهش

برای پیش‌بینی و الگوسازی شاخص نرخ وصول، علاوه بر مشخصات اعتباری مشتریان، اطلاعات متغیرهای اقتصادی نیز در الگوسازی استفاده و برای محاسبه این شاخص، از مشتریانی در الگوسازی استفاده می‌شود که تسهیلات دریافتی آنها به طبقه مطالبات غیرجاری (سررسید گذشته / معوق / مشکوک‌الوصول) منتقل شده است.

$$RR = \frac{DNR}{EAD} = \frac{FR}{EAD} \times \frac{FR - AC}{FR} \times (1 + r)^{-T} \quad (1)$$

RR نرخ وصول تعهدات مطالبات شده، DNR

ارزش تنزیل خالص وصول که به خالص تمام

هزینه‌های وصول مطالبات گفته می‌شود. EAD میزان بدهی در زمان ایجاد مطالبات. FR ارزش اسمی میزان وصول در دوره (در این پژوهش ارزش وثیقه‌های دریافتی به جای میزان وصولی در دوره واقعی در نظر گرفته شده است. AC هزینه‌های اداری مرتبط با فرایند وصول تعهدات مطالبات شده. r نرخ تنزیل و T زمان فرایند وصول مطالبات [۱۲]. با توجه به اطلاعات اعتباری موجود و مشخص نبودن هزینه‌های اداری، نرخ تنزیل و زمان فرایند وصول برای تک‌تک مشتریان و همچنین فرایند وصول مطالبات که بعضاً به دلیل بوروکراسی اداری، مکاتبات واحدهای حقوقی، تشکیل دادگاه در کشور ایران که کاملاً متغیر و بعضاً غیر قابل اندازه‌گیری است، فرمول مذکور با توجه به شرایط بانک‌های کشور و اطلاعات موجود بومی‌سازی می‌شود. در این پژوهش، ارزش اسمی میزان وصول به دست آمده است که از اندازه‌گیری ارزش مبلغی وثیقه‌های دریافتی با ضرایب اعلامی از سوی بانک مرکزی ج.ا.ا. براساس (دستورالعمل نحوه محاسبه ذخیره مطالبات، ۱۳۹۰) اعلام شده است و به عنوان DNR در نظر گرفته می‌شود. در حالت کلی برای محاسبه نرخ وصول واقعی باید میزان وصول واقعی اتفاق افتاده در دوره را جایگزین کرد؛ بنابراین فرمول محاسبه نرخ واقعی وصول تعهدات مطالبات شده به صورت رابطه (۲) و (۳) است:

$$RR = \frac{DNR}{EAD} = \frac{FR}{EAD} \quad (2)$$

ارزش اسمی مطلوب

= میزان بدهی حال مشتری

$$RR = \sum_{i=1}^2 (W_i \times FR_i) \quad (3)$$

ارزش وثایق مطلوب

پس انداز، سرمایه گذاری مدت دار و ... ۱۰۰، ضمانت صادرات، سهام بورس، غیر منقول از محل اجرای طرح/ خارج از محل، غیر منقول ملکی، غیر منقول کارخانه همگی ۷۰ و غیر منقول ماشین آلات و تجهیزات و کالا ۵۰ درصد است. الگوریتم استفاده شده، روش ماشین بردار تصمیم است. روش ماشین بردار تصمیم به دو صورت مبتنی بر طبقه بندی و مبتنی بر پیش بینی است. در این پژوهش، پیش بینی با روش های مبتنی بر رگرسیون انجام می گیرد. علت این موضوع، پیوستگی متغیر هدف (نرخ وصول) است و شامل زیر مجموعه ای از الگوهای پیش بینی است که شامل الگوهای مختلف - SVR ،

-SVR است؛ با این تفاوت که کلیه این الگوها با روش الگوریتم ژنتیک برای انتخاب متغیرهای الگو، بهینه سازی و در نهایت، میزان خطا و کارایی آنها مقایسه می شود. مبنای الگوسازی در این پژوهش [۴] و [۱۳] است.

روش رگرسیون بردار تصمیم (مبتنی بر پیش بینی یا تخمین عددی) برای اهداف پیوسته در داده کاوی است.

که در آن W_i ضریب ارزش وثیقه i و FR_i ارزش اسمی مطلوب است. در پژوهش حاضر در بانک مد نظر، دو نوع وثیقه $i=1,2$ با ارزش متفاوت با مشتری گرفته می شود در جدول (۱) لیست وثیقه ها به همراه ارزش وزنی آنها آمده است. ارزش وزنی سایر وثیقه ها (سفته و برات، قرارداد، چک، ضامن معتبر و امضای مدیران) صفر است (توجه شود در حالت کلی، صورت کسر، میزان وصولی در دوره است که در این پژوهش، ارزش وثیقه های دریافتی در نظر گرفته شده است)

$$(4) \text{ سایر بدهی ها} + (\text{سود آینده} - \text{اصل مبلغ وام}) = \text{میزان بدهی حال مشتری}$$

سایر بدهی ها

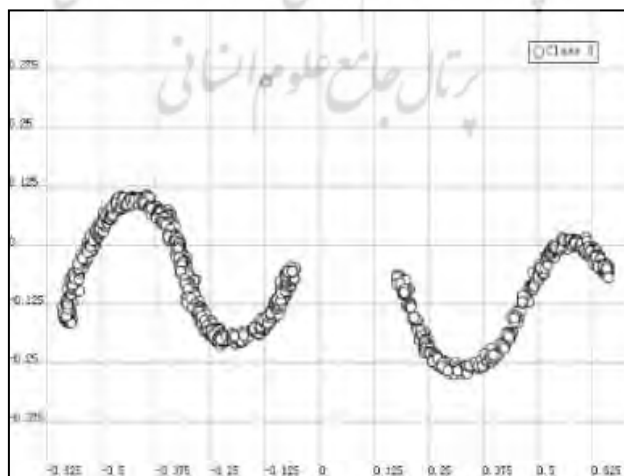
$$= \sum (\text{معوق} + \text{سررسید گذشته})$$

$$(5) \text{ سود دریافتی} + \text{مشکوک الوصول} +$$

$$\text{سود دریافتی مطالبات} +$$

$$\text{وجه التزام دریافتی} +$$

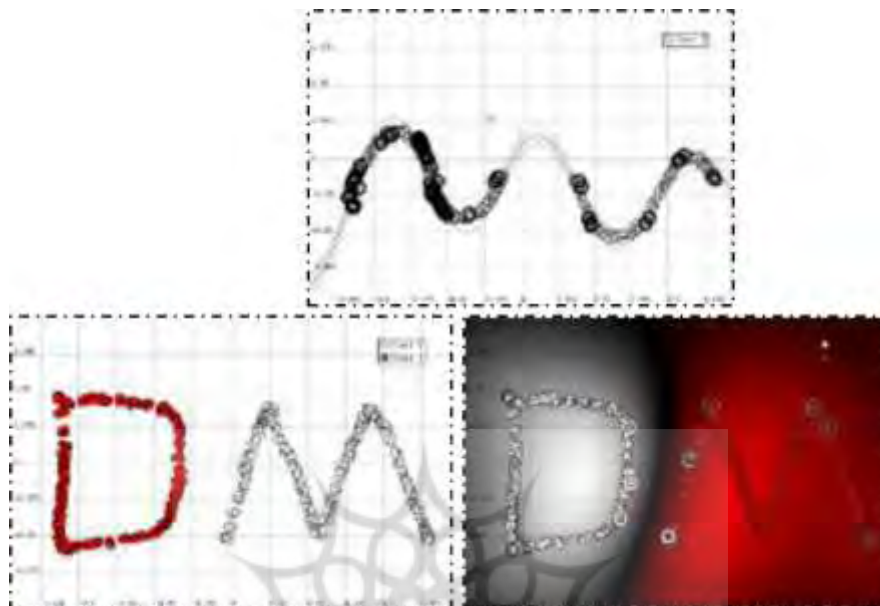
ارزش وثیقه های مندرج در بانک اطلاعات مشتریان اعتباری با توجه به ضرایب اعلامی بانک مرکزی ج.ا.ا. براساس (دستورالعمل نحوه محاسبه ذخیره مطالبات، ۱۳۹۰) غیر منقول ۷۰ درصد، سپرده (قرض الحسنه



شکل (۱) الگوسازی به روش SVR

بد و خوب با تفکیک هدف از این الگو برای پیش‌بینی می‌توان استفاده کرد.

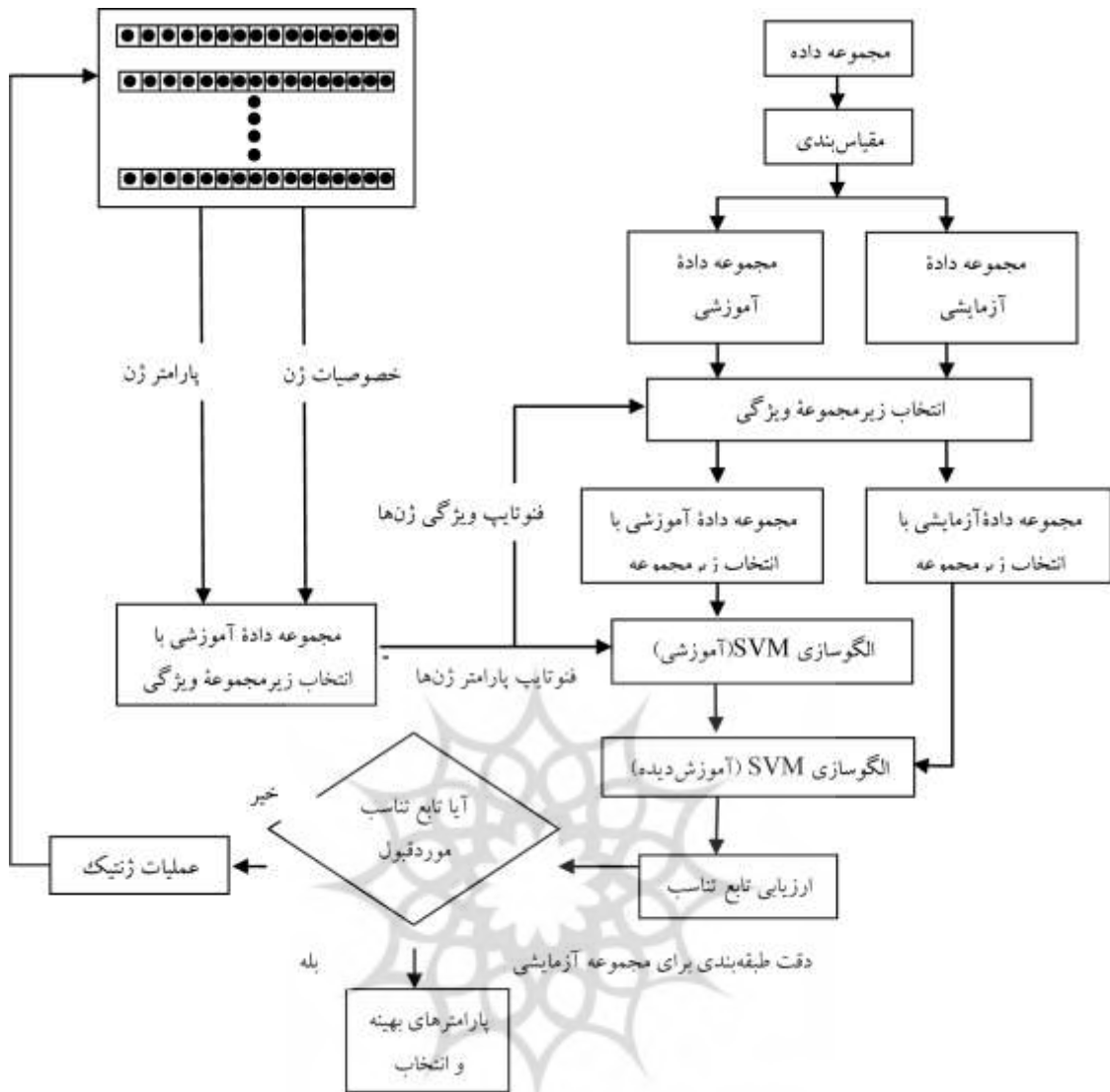
روش الگوسازی طبقه‌بندی بردار تصمیم (مبتنی بر طبقه‌بندی) برای پیش‌بینی اهداف گسسته و یا عدد صحیح به کار گرفته می‌شود. برای پیش‌بینی مشتریان



شکل (۲) الگوسازی به روش SVC

کرد؛ به عبارتی، طرح کروموزوم، تابع تناسب و نوع معماری سیستم برای انتخاب ویژگی مبتنی بر الگوریتم ژنتیک و بهینه‌سازی پارامترها است. الگوریتم کلی که در تمامی قسمت‌ها تقریباً با اندکی تغییر برای الگوسازی استفاده شده است، به صورت شکل (۳) نشان داده شده است.

در روش بردار تصمیم، پارامترهای الگو نظیر C, σ و قابل تنظیم است که قبل از الگوسازی با انتخاب یک عدد خاص، الگو بر داده‌های آموزشی فرا گرفته شود که با استفاده از روش الگوریتم ژنتیک، انتخاب این پارامترها را با رویکرد کم کردن خطای الگو در مقایسه با داده‌های تستی به‌طور بهینه می‌توان انتخاب



شکل (۳) معماری سیستم پیشنهادی برای انتخاب ویژگی و پارامترهای بهینه با استفاده از الگوریتم ژنتیک

جایگزینی بر کلیه قراردادهای مذکور انجام شده است و برای آزمایش دقت داده‌ها در مرحله بررسی دقت الگو از نمونه‌ای با حجم ۲۰۰۰ و با روش نمونه‌گیری تصادفی ساده بدون جایگزینی استفاده شده است. برای هر یک از الگوها نیز نمونه‌گیری جدید انجام شده است.

فرایند داده کاوی و الگوسازی بر داده‌های یکی از بانک‌های دولتی در فاصله مهر تا شهریورماه سال‌های ۱۳۸۹-۱۳۹۴ به صورت ماهانه به تعداد ۶۰ مقطع انجام شده است. حجم جامعه آماری مدنظر، تعداد ۲۵،۰۱۷،۰۳۶ قرارداد است. نمونه آماری برای آموزش الگو، نمونه‌ای به حجم ۵۰۰ قرارداد از جامعه مدنظر براساس روش نمونه‌گیری تصادفی ساده بدون

جدول (۱) متغیرهای استفاده شده در الگوسازی نرخ وصول

نام متغیر	شرح متغیر	ردیف
TR_ADMIN	کد منطقه	۱
TR_BR	کد شعبه	۲
C_noegharardad	کد نوع قرارداد	۳
Year_gharardad	سال قرارداد	۴
C_reshtefaaliat	کد رشته فعالیت	۵
Mablagh_pardakhtani	مبلغ پرداختی	۶
Mablagh_bazgashti	مبلغ بازگشتی	۷
Soodsahmmoshtari	سود سهم مشتری	۸
Soodsahmdolat	سود سهم دولت	۹
Nerkhsoodsahmmoshtari	نرخ سود سهم مشتری	۱۰
Nerkhsood_sahmdolat	نرخ سود سهم دولت	۱۱
Darsadmoshtari	درصد مشتری	۱۲
Tedadaghsat	تعداد اقساط	۱۳
C_noerecord	کد نوع رکورد	۱۴
C_budget	کد بودجه	۱۵
C_hadaf	کد هدف	۱۶
Mablagh_mosavab	مبلغ مصوب	۱۷
C_edareebلاغconandeh	کد اداره ابلاغ کننده	۱۸
C_mosavabe	کد مصوبه	۱۹
C_taklif	کد تکلیف	۲۰
C_moremasraf	کد مدنظر	۲۱
Ravesh_taghseet	روش تقسیط	۲۲
C_shahrmoredmasraf	کد شهر مدنظر	۲۳
Gender	جنسیت	۲۴
Inflation	نرخ تورم	۲۵
GDP	تولید ناخالص داخلی به قیمت پایه	۲۶
Arzeshvasaiegh	ارزش اسمی وثیقه‌ها	۲۷

سناریوهای مختلفی براساس الگوهای مذکور در بخش قبل در نظر گرفته می شود.

با توجه به اینکه برای الگوسازی در داده کاوی، روش های مختلفی وجود دارد؛ نظیر داده کاوی نظارت شده و غیر نظارت شده، برای الگوسازی،

جدول (۲) الگوهای زیرمجموعه ماشین بردار تصمیم به تفکیک سناریوهای استفاده شده

روش داده کاوی	یک مرحله ای	دو مرحله ای	سه مرحله ای
نظارت شده	✓	✓	-
غیر نظارت شده	-	✓	-
نظارت شده + غیر نظارت شده	-	-	✓

در روش های الگوسازی، روش یک مرحله ای مستقیماً نرخ وصول مشتریان را پیش بینی می کند؛ اما روش های دو مرحله ای و سه مرحله ای، ابتدا طبقه ریسک را با توجه به طبقه بندی با استفاده از درخت تصمیم مشخص و سپس طبقه با ریسک بالا و کم از الگوسازی حذف و طبقه با ریسک متوسط پیش بینی استفاده می شود. در روش داده کاوی نظارت شده یک مرحله ای به صورت مستقیم به الگوسازی نرخ وصول مشتریان با استفاده از روش های بخش قبل توجه می شود. الگوهای استفاده شده در این سناریو - GA-SVR و SVR است. در روش داده کاوی نظارت شده دو مرحله ای، حالت اول بخش الگوسازی به صورت دو مرحله ای انجام می گیرد؛ یعنی ابتدا، متغیر هدف (نرخ وصول) به سه طبقه تقسیم می شود. نرخ وصول با مقدار ۱ (احتمال وصول ۱۰۰ درصد)، نرخ وصول با مقدار بین ۰ و ۱، نرخ وصول با مقدار ۰ (احتمال وصول صفر). طبقات بالا به ترتیب، با ستون وضعیت (State) در فایل مشتریان با کدهای ۱، ۲ و ۳ طبقه بندی می شوند.

در روش های الگوسازی، روش یک مرحله ای مستقیماً نرخ وصول مشتریان را پیش بینی می کند؛ اما روش های دو مرحله ای و سه مرحله ای، ابتدا طبقه ریسک را با توجه به طبقه بندی با استفاده از درخت تصمیم مشخص و سپس طبقه با ریسک بالا و کم از الگوسازی حذف و طبقه با ریسک متوسط پیش بینی استفاده می شود. در روش داده کاوی نظارت شده یک مرحله ای به صورت مستقیم به الگوسازی نرخ وصول مشتریان با استفاده از روش های بخش قبل توجه می شود. الگوهای استفاده شده در این سناریو - GA-SVR و SVR است. در روش داده کاوی نظارت شده دو مرحله ای، حالت اول بخش الگوسازی به صورت دو مرحله ای انجام می گیرد؛ یعنی ابتدا، متغیر هدف (نرخ وصول) به سه طبقه تقسیم می شود. نرخ وصول با مقدار ۱ (احتمال وصول ۱۰۰ درصد)، نرخ وصول با مقدار بین ۰ و ۱، نرخ وصول با مقدار ۰ (احتمال وصول صفر). طبقات بالا به ترتیب، با ستون وضعیت (State) در فایل مشتریان با کدهای ۱، ۲ و ۳ طبقه بندی می شوند.

جدول (۳) طبقه بندی وضعیت نرخ وصول مشتریان

شرح	طبقه	برچسب شرح طبقه
نرخ وصول ۱	۱	LOW
نرخ وصول بین ۰ و ۱	۲	MEDIUM
نرخ وصول ۰	۳	HIGH

الگوهای استفاده شده در این سناریو - C5.0 +GA-SVR و SVR است. در روش داده کاوی غیر نظارت شده دو مرحله ای حالت دوم، ابتدا داده ها خوشه بندی می شوند و سپس بر اساس خروجی هر خوشه، نمونه گیری بر هر خوشه و

پیش بینی انجام می گیرد. روش یک مرحله ای در این سناریو بی معنی است؛ زیرا خوشه بندی به صورت خودکار، مرحله اول را انجام می دهد (خوشه بندی + پیش بینی). الگوهای استفاده شده در این بخش عبارت است از K-MEANS+GA-SVR و

(خوشه‌بندی+طبقه‌بندی+پیش‌بینی). الگوهای K-MEANS+C5.0 استفاده شده در این سناریو به شرح K-MEANS+C5.0 +GA- -SVR و +GA- -SVR است.

یافته‌ها

در ادامه، به صورت تفکیک شده، نتایج هر سناریو به تفکیک الگو آمده است.

GA- -SVR +K-MEANS در روش داده‌کاوی نظارت شده و غیرنظارت شده سه مرحله‌ای، ابتدا داده‌ها برچسب‌گذاری و سپس براساس سناریوی الگوسازی براساس روش داده‌کاوی غیرنظارت شده خوشه‌بندی می‌شوند و سپس براساس خروجی هر خوشه، طبقه‌بندی انجام می‌شود و در نهایت، پیش‌بینی بر جامعه طبقه‌بندی شده انجام می‌شود. الگوسازی در این سناریو به صورت سه مرحله‌ای است

جدول (۴) نتایج الگوی GA-ν-SVR - داده‌کاوی نظارت شده یک مرحله‌ای

شرح پارامتر	مقدار	مقدار بهینه
بازه تغییرات C	۱۰-۰/۰۰۰۱	۶/۶۰۱۷۱۵
بازه تغییرات	۲-۰/۰۰۱	۰/۰۲۲۶۳۷۰۲
بازه تغییرات	۱-۰	۰/۳۹۳۲۷۷۸
تعداد تکرار الگوریتم	۱۰	-
حجم جمعیت	۴۰	-
احتمال جهش	۰/۹	-
احتمال عبور	۰/۱	-
مقدار تابع تناسب	-۰/۰۳۵۹۲۱۵۷	-

جدول (۵) نتایج الگوی GA-ε-SVR - داده‌کاوی نظارت شده یک مرحله‌ای

شرح پارامتر	مقدار	مقدار بهینه
بازه تغییرات C	۱۰-۰/۰۰۰۱	۵/۹۹۳۲۳۹
بازه تغییرات	۲-۰/۰۰۱	۰/۰۲۷۶۹۸۱۷
بازه تغییرات	۲-۰/۰۱	۰/۱۷۷۱۲۸۸
تعداد تکرار الگوریتم	۱۰	-
حجم جمعیت	۵۰۰	-
احتمال جهش	۰/۸	-
احتمال عبور	۰/۰۵	-
مقدار تابع تناسب	-۰/۰۳۵۶۳۵۴۸	-

جدول (۶) نتایج الگوی GA-ε-SVR - داده کاوی نظارت شده دومرحله‌ای

شرح پارامتر	مقدار	مقدار بهینه
بازه تغییرات C	۰/۰۰۰۱-۱۰	۵/۱۵۵۱۲۶
بازه تغییرات	۰/۰۰۱-۲	۰/۰۱۵۸۰۵۸
بازه تغییرات	۰/۰۱-۲	۰/۵۶۱۸۸۰۴
تعداد تکرار الگوریتم	۱۰	-
حجم جمعیت	۵۰۰	-
احتمال جهش	۰/۸	-
احتمال عبور	۰/۰۵	-
مقدار تابع تناسب	-۰/۰۲۶۱۵۴۶۸	-

جدول (۷) نتایج الگوی GA-ν-SVR - داده کاوی نظارت شده دومرحله‌ای

شرح پارامتر	مقدار	مقدار بهینه
بازه تغییرات C	۰/۰۰۰۱-۱۰	۸/۸۵۸۹۷۶
بازه تغییرات	۰/۰۰۱-۲	۰/۰۰۰۲۷۲۸۷۳
بازه تغییرات	۰-۱	۰/۳۴۶۹۳۹۹
تعداد تکرار الگوریتم	۱۰	-
حجم جمعیت	۴۰	-
احتمال جهش	۰/۹	-
احتمال عبور	۰/۱	-
مقدار تابع تناسب	-۰/۰۲۵۵۰۶۴۱	-

جدول (۸) ستون‌های استفاده شده در فرایند خوشه‌بندی (نرمال سازی شده)

شرح	نام متغیر	وضعیت
ارزش اسمی وثیقه‌ها	Arzeshvasaiegh_Transformed	نرمال شده
کد نوع قرارداد (عقد تسهیلاتی)	C_noegharardad	نرمال شده
کد رشته فعالیت (بخش اقتصادی)	C_reshtefaaliat	نرمال شده
مبلغ تسهیلات پرداختی	Mablagh_Pardakhtani_Transformed	نرمال شده
مبلغ تسهیلات بازگشتی	Mablagh_Bazgashti_Transformed	نرمال شده

جدول (۹) نتایج خوشه‌بندی براساس تقسیم‌بندی مجموعه داده آموزشی و آزمایشی

بخش	مجموعه داده آموزشی	مجموعه داده آزمایشی	جمع کل
بخش اول	۱۴/۱۲۱/۸۶۴	۷/۶۰۶/۶۸۴	۲۱/۷۲۸/۵۴۸
بخش دوم	۲/۱۳۶/۶۷۵	۱/۱۵۱/۸۰۷	۳/۲۸۸/۴۸۲
جمع کل	۱۶/۲۵۸/۵۳۹	۸/۷۵۸/۴۹۱	۲۵/۰۱۷/۰۳۰
سهم درصد از کل	%۶۵	%۳۵	

جدول (۱۰) نتایج خوشه‌بندی براساس تقسیم‌بندی مجموعه داده آموزشی و آزمایشی

تکرار	میزان خطا
۱	۰/۵۶۷
۲	۰/۵۷
۳	۰/۳۳۵
۴	۰/۰۹۶
۵	۰/۰
۶	۰/۰

کل مجموعه هر بخش با توجه به اینکه به دو بخش تقسیم شده است، در بخش داده‌های آموزشی آموزش دیده و پس از کنترل بر داده‌های آزمایشی در صورت کمترین خطا انتخاب می‌شود.

جدول (۱۱) نتایج خوشه‌بندی حجم خوشه‌ها و نسبت بزرگ‌ترین خوشه به کوچک‌ترین خوشه

درصد	مقدار	شرح
٪۱۳/۱	۲/۱۳۶/۶۷۵	حجم کمترین خوشه
٪۸۶/۹	۱۴/۱۲۱/۸۳۴	حجم بیشترین خوشه
	۶/۶۱	نسبت بیشترین خوشه به کمترین خوشه

جدول (۱۲) نتایج اجرای الگوی GA-ε-SVR بر مبنای انتخاب پارامترها با روش الگوریتم ژنتیک بر اساس روش داده کاوی غیر نظارت شده و پیش بینی نظارت شده

شرح پارامتر	مقدار (خوشه اول)	مقدار بهینه (خوشه اول)	مقدار (خوشه دوم)	مقدار بهینه (خوشه دوم)
بازه تغییرات C	۰/۰۰۰۱-۱۰	۲/۸۴۰۰۴۷		
بازه تغییرات	۰/۰۰۱-۲	۰/۰۷۵۱۶۹۷۱		
بازه تغییرات	۰/۰۱-۲	۰/۰۱۶۳۷۸۷۳		
تعداد تکرار الگوریتم	۵	-		پاسخی از الگوی بهینه سازی در هیچ حالتی دریافت نمی شود و الگو، جواب بهینه ندارد.
حجم جمعیت	۵۰	-		
احتمال جهش	۰/۸	-		
احتمال عبور	۰/۱	-		
مقدار تابع تناسب	-۰/۰۳۲۹۵۰۷۸	-		

جدول (۱۳) نتایج اجرای الگوی GA-v-SVR بر مبنای انتخاب پارامترها با روش الگوریتم ژنتیک بر اساس روش داده کاوی غیر نظارت شده و پیش بینی نظارت شده

شرح پارامتر	مقدار (خوشه اول)	مقدار بهینه (خوشه اول)	مقدار (خوشه دوم)	مقدار بهینه (خوشه دوم)
بازه تغییرات C	۰/۰۰۰۱-۱۰	۳/۳۱۵۰۳۱		۹/۰۲۹۳۶۷
بازه تغییرات	۰/۰۰۱-۲	۰/۰۳۹۵۱۵۳۷		۰/۰۱۷۹۲۴۴۳
بازه تغییرات	۰-۱	۰/۶۸۰۵۱۵۹		۰/۲۱۵۴۸۷
تعداد تکرار الگوریتم	۱۰	-		۱۰
حجم جمعیت	۵۰	-		۵۰
احتمال جهش	۰/۸	-		۰/۸
احتمال عبور	۰/۱	-		۰/۱
مقدار تابع تناسب	-۰/۰۱۷۸۰۸۸۹	-		-۰/۰۶۰۹۱۸۷۱

جدول (۱۴) نتایج اجرای الگوی GA-ε-SVR براساس روش داده کاوی نظارت شده و غیرنظارت شده سه مرحله‌ای

شرح پارامتر	مقدار (خوشه اول)	مقدار بهینه (خوشه اول)	مقدار (خوشه دوم)	مقدار بهینه (خوشه دوم)
بازه تغییرات C	۰/۰۰۰۱-۱۰	۸/۶۰۲۲۱۴	۰/۰۰۰۱-۱۰	۴/۴۴۹۲۹۷
بازه تغییرات	۰/۰۰۱-۲	۰,۰۰۵۹۵۰۷۰۶	۰/۰۰۱-۲	۰/۰۱۰۵۶۴۴۱
بازه تغییرات	۰/۰۱-۲	۰,۰۸۳۵۹۷۲۵	۰/۰۱-۲	۰/۱۵۲۷۱۸
تعداد تکرار الگوریتم	۱۰	-	۱۰	-
حجم جمعیت	۵۰۰	-	۵۰۰	-
احتمال جهش	۰/۸	-	۰/۸	-
احتمال عبور	۰/۰۵	-	۰/۰۵	-
مقدار تابع تناسب	-۰/۰۳۰۴۹۵۲۸	-	-۰/۰۳۳۲۳۲۴۳	-

جدول (۱۵) نتایج اجرای الگوی GA-ν-SVR براساس روش داده کاوی نظارت شده و غیرنظارت شده سه مرحله‌ای

شرح پارامتر	مقدار (خوشه اول)	مقدار بهینه (خوشه اول)	مقدار (خوشه دوم)	مقدار بهینه (خوشه دوم)
بازه تغییرات C	۰/۰۰۰۱-۱۰	۲/۲۳۱۸۲۸	۰/۰۰۰۱-۱۰	۳/۴۸۷۶۷۱
بازه تغییرات	۰/۰۰۱-۲	۰/۱۱۶۹۸۶۱	۰/۰۰۱-۲	۰/۰۲۶۲۰۴۵۷
بازه تغییرات	۰-۱	۰/۲۸۸۰۸۶۹	۰-۱	۰/۸۳۷۹۷۱۲
تعداد تکرار الگوریتم	۱۰	-	۱۰	-
حجم جمعیت	۴۰	-	۴۰	-
احتمال جهش	۰/۹	-	۰/۹	-
احتمال عبور	۰/۱	-	۰/۱	-
مقدار تابع تناسب	-۰/۰۳۰۸۹۳۷۶	-	-۰/۰۳۳۴۷۲۵	-

نتایج و پیشنهادها

برای راحتی فراخوانی الگوها به صورت جدول (۱۶) کدبندی می شود.

جدول (۱۶) کدبندی الگوهای استفاده شده

نام الگو	کد الگو
-SVR	M1
-SVR	M2
C5.0 +GA- -SVR	M3
C5.0 +GA- -SVR	M4
K-MEANS+GA- -SVR	M5
K-MEANS+GA- -SVR	M6
K-MEANS+C5.0+GA- -SVR	M7
K-MEANS+C5.0+GA- -SVR	M8

بررسی و خطاها مقایسه می شود. برای الگوهای M1 و M2 مجموعه داده آزمایشی از کل جامعه انتخاب می شود و نیازی به تفکیک نیست. با توجه به نتایج به دست آمده، رتبه بندی الگوهای منتخب نهایی براساس کمترین خطاها و همچنین بیشترین پوشش سطح زیر منحنی به شرح جدول (۱۷) است.

در این بخش، ابتدا جامعه اصلی مشتریان دارای مطالبات غیر جاری (بدهکار به بانک) به دو بخش آموزشی و آزمایشی تقسیم می شود؛ سپس تمامی الگوهای به دست آمده با تنظیمات بهینه بر نمونه ای تصادفی با حجم ۲۰۰۰ مشاهده از جامعه اصلی (مجموعه داده آزمایشی مربوط به خود) اجرا و نتایج

جدول (۱۷) الگوهای منتخب براساس رتبه بندی کمترین میزان خطا

رتبه	کد الگو	MAE	MSE	RMSE
۱	M1	۰/۰۷۲۴۴۲۳۲	۰/۰۳۲۰۱۴۷۹	۰/۱۷۸۹۲۶۸
۲	M6	۰/۰۷۲۰۷۱۷	۰/۰۳۵۸۱۶۳	۰/۱۸۷۳۵۰۱
۳	M2	۰/۰۵۶۲۰۰۸۱	۰/۰۳۹۷۷۴۶۹	۰/۱۹۹۴۳۵۹

به ویژه الگوهای پیشنهادی در پژوهش حاضر اقدام کنند و قبل از اعطای تسهیلات شعبه با وارد کردن اطلاعات لازم مشتری و مقایسه و الگوسازی با داده های گذشته و بانک اطلاعاتی بانک و یا حتی به طور کامل تر با جامعه سیستم بانکی کشور با تأمل و بررسی بیشتری به اعطای تسهیلات اقدام کنند. بانک ها و مؤسسات مالی با روش مذکور در پژوهش حاضر، علاوه بر پیش بینی احتمال مطالبات شدن هر مشتری و در صورت مطالبات شدن پیش بینی میزان وصول قرارداد مطالبات شده، میزان زیان بانک ناشی از وصول نشدن قراردادهای اعتباری را

با توجه به نتایج به دست آمده براساس اجرای الگوهای بهینه شده با استفاده از الگوریتم ژنتیک مشخص می شود الگوی -SVR بهترین الگو برای پیش بینی نرخ وصول مطالبات در روش یک مرحله ای بدون پیش بینی طبقه ریسک مشتریان است و در روش دو مرحله ای، الگوی K-MEANS+GA- -SVR بهترین الگو برای پیش بینی نرخ وصول مطالبات است. پیشنهاد می شود در حوزه بانکداری، بانک ها به راه اندازی سامانه جامع اعطای تسهیلات بر مبنای تحلیل اطلاعات مشتریان با استفاده از الگوریتم های داده کاوی

- [6] Huang, C. L. & Wang, C. J. (2006). A GA-based feature selection and parameters optimization. *Expert Systems with Applications*. (31): 231-240.
- [7] Karani, H. & Aghaei Pour, M. (2014). Application of the theory of survival analysis of credit risk management loan recipients. *Ravand Quarterly* (21):175-200
- [8] Loterman, G. (2013). *Predicting Loss Given Default* PHD Thesis: Ghent University.
- [9] Mohammadian H. K. A., Asgharzadeh Z. M. & Emam D. M. (2016). Credit risk assessment of corporate customers using support vector machine and genetic algorithm hybrid model - A case study of Tejarat Bank. *Financial Engineering & Portfolio Management*. (7): 17-32.
- [10] Moin, K. & Baseer A. D. (2012). Use of data mining in banking. *International Journal of Engineering Research and Applications (IJERA)*. (2): 738-742.
- [11] Nazarpour, M. T. & Rezaei, A. (2013). Credit risk management in Islamic banking with approach to review contracts and loan payment pattern. *Islamic Financial Research*. (2):123-156.
- [12] Resti, A. & Sironi, A. (2007). *Risk Management and Shareholders' Value in Banking*. England: John Wiley & Sons.
- [13] Sermpinis, G. Stasinakis, C. & Theofilatos, K. (2015). Modeling, forecasting and trading the EUR exchange rates with hybrid rolling genetic algorithms support vector regression forecast combinations. *European Journal of Operational Research*. (247): 831-846.
- [14] Shahrabi, J., Hadavandi, E. (2011). *Data Mining In Banking*. Tehran: Iranian Academic Center for Education Culture & Research.
- [15] Witzany, J. Rychnovsky, M. & Charamza, P. (2012). Survival analysis in LGD modeling. *European Financial and Accounting Journal*. (7): 6-27.
- [16] Yao, X. Crook, J. & Andreeva, G. (2015). Support vector regression for loss given default modelling. *European Journal of Operational Research*. (240): 528-538.
- می‌توانند به پیش‌بینی و برآورد کنند. راه‌اندازی سامانه هوشمند اعطای تسهیلات برای الگوسازی با روش‌های داده‌کاوی قبل از اعطای تسهیلات می‌تواند اقدامی در راستای کنترل انضباط اعتباری براساس الزامات بانک مرکزی ج.ا.ا و کمیته بال باشد.
- در این پژوهش با توجه به محدودیت‌های موجود، محاسبه نرخ وصول به صورت کامل انجام نشد و برای الگوسازی به دلیل مشخص نبودن میزان وصول واقعی از ارزش مبلغی وثیقه‌ها برای الگوسازی استفاده و صرفاً بر روش الگوسازی تأکید شد. پیشنهاد می‌شود در پژوهش‌های آینده با استفاده از اطلاعات کامل‌تر نظیر زمان فرایند وصول مطالبات، نرخ تنزیل، هزینه‌های وصول مطالبات هر قرارداد نرخ وصول محاسبه شود که طبیعتاً به عدد صفر نزدیک‌تر است. همچنین از روش‌های دیگر الگوسازی نیز استفاده شود.

منابع

- [1] Amiram, D. (2011). Debt contracts and loss given default. *Job Market Paper*. University of North Carolina-Chapel Hill.
- [2] Arsova, A. Haralampieva, M. & Tsvetanova, T. (2011). Comparison of regression models for LGD estimation. *Credit Scoring and Credit Control XII Edinburgh: Experian Limited*:1-23
- [3] Bastos, J. A. (2010). Forecasting bank loans loss-given-default. *Journal of Banking & Finance*, Vol. 34 (10): 2510-2517.
- [4] Chen, K.Y. & Wang, C. H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management*. (28): 215-226
- [5] Gürtler, M. & Hibbeln, M. (2013). Improvements in loss given default forecasts for bank loans. *Journal of Banking & Finance*. (37): 2354-2366.