-- Journal of Economic Research----------
Vol(issue), PP.
.atu.ac.ir
DOI:

ATU PRESS

Original Research / Review / ...
Received:
Accepted:
ISSN:
eISSN:

Predicting Loan Default Behavior Using Machine Learning Models: A Case Study of Resalat Bank

Seyed Fakhredin Fakhrehosseini
Associate Professor, Department of Accounting, TONIAU, Islamic Azad University, Tonekabon, Iran

Abstract

In recent years, the surge in credit risk among customers and the escalation of economic crises have contributed to rising loan default rates in the banking sector. In this context, identifying high-risk customers has become a top priority for banks in credit risk management. This research employs real customer data from Resalat Bank, spanning from May 22, 2020, to March 18, 2025, to predict loan default behavior using machine learning models. The dataset was processed based on demographic and credit history attributes, including age, gender, loan amount, number of installments, and payment delays. To evaluate performance, Logistic Regression models with L1 and L2 regularization, Neural Networks, Random Forest, and XGBoost were employed. Furthermore, the Synthetic Minority Over-sampling Technique (SMOTE) was utilized to address data class imbalance. The results indicate that decision tree-based models, particularly XGBoost and Random Forest, outperform Logistic Regression and Neural Networks in terms of Recall, Precision, and F1-score metrics. These findings suggest that combining relevant data with robust algorithms designed for imbalanced datasets can significantly improve the accuracy of default risk prediction.

Keywords: Credit Loans Risk; Financial Default; Machine Learning.

JEL Classification: G21; C45; C38; C55; G17

مقاله پژوهشی/مروری/سخن سردبیر
تاریخ دریافت: ۰۰/۰۰/۰۰؛ تاریخ پذیرش: ۰۰/۰۰/۰۰
eISSN: ISSN:

--پژوهشنامه اقتصادی---------
دوره ؟، شماره ؟، نام فصل، سال، ص ص
.atu.ac.ir
DOI:

پیش‌بینی رفتار نکول مالی تسهیلات اعتباری با استفاده از مدل‌های یادگیری ماشین (مطالعه موردی بانک رسالت)

سید فخرالدین فخرحسینی[1]
گروه حسابداری، واحد تنکابن، دانشگاه آزاد اسلامی، تنکابن، ایران.

چکیده

در سال‌های اخیر، افزایش ریسک تسهیلات‌گیری مشتریان و تشدید اثرات بحران‌های اقتصادی، موجب رشد نرخ نکول در تسهیلات بانکی شده است. در این زمینه، شناسایی مشتریان پرریسک به یکی از اولویت‌های اصلی بانک‌ها در مدیریت ریسک اعتباری تبدیل شده است. این پژوهش با استفاده از داده‌های واقعی مشتریان بانک رسالت در بازه ۱۳۹۹/۰۲/۰۲ تا ۱۴۰۳/۱۲/۲۸، به پیش‌بینی رفتار نکول تسهیلات اعتباری با بهره‌گیری از مدل‌های یادگیری ماشین می‌پردازد. داده‌های پژوهش براساس اطلاعات پایه جمعیت‌شناختی و سابقه اعتباری مشتریان، شامل سن، جنسیت، مبلغ تسهیلات، تعداد اقساط و تأخیر در پرداخت پردازش شده‌اند. برای ارزیابی عملکرد، مدل‌های رگرسیون لجستیک با منظم‌سازی L1 و L2، شبکه عصبی، جنگل تصادفی و XGBoost به‌کار گرفته شدند. همچنین برای مقابله با نامتوازن بودن داده‌ها از تکنیک SMOTE استفاده شد. نتایج نشان داد مدل‌های مبتنی بر درخت تصمیم، به‌ویژه XGBoost و جنگل تصادفی، در شاخص‌های بازخوانی، دقت و F۱ عملکرد بهتری نسبت به رگرسیون لجستیک و شبکه عصبی دارند. این یافته‌ها نشان می‌دهد داده‌ها همراه با الگوریتم‌های مناسب برای داده‌های نامتوازن، می‌تواند دقت پیش‌بینی ریسک نکول را به‌طور معناداری افزایش دهد.

کلیدواژه‌ها: ریسک تسهیلات اعتباری، نکول مالی، یادگیری ماشین.

[1] sf.fakhrhosseini@iau.ac.ir

مقدمه

تسهیلات بانکی از یک‌سو مهم‌ترین منبع درآمد عملیاتی بانک‌ها و از سوی دیگر یکی از اصلی‌ترین کانون‌های ایجاد ریسک در ترازنامه محسوب می‌شود. افزایش مطالبات غیرجاری و تحقق زیان‌های اعتباری می‌تواند از مسیر تضعیف کیفیت دارایی‌ها، کاهش سودآوری و محدودسازی توان وام‌دهی، پیامدهایی فراتر از یک بانک ایجاد کرده و به شکنندگی مالی در سطح نظام بانکی منجر شود (Khandani et al., 2010). در چنین شرایطی، پیش‌بینی به‌موقع نکول مشتریان و تفکیک دقیق متقاضیان کم‌ریسک از پرریسک، یک الزام مدیریتی و نظارتی برای کنترل زیان مورد انتظار، ارتقای کارایی تخصیص اعتبار و پشتیبانی از ثبات مالی است (BCBS[1], 2017). از منظر عملی، بانک‌ها برای تصمیم‌گیری اعتباری ناچارند با عدم‌قطعیت قابل توجهی مواجه شوند؛ زیرا رفتار بازپرداخت مشتریان نه‌تنها تابع توان مالی، بلکه متاثر از شوک‌های درآمدی، تغییرات محیط کلان و ویژگی‌های رفتاری و قراردادی است(Zhang & Yu, 2024) .

ارزیابی اعتباری در بسیاری از بانک‌ها هنوز به اتکای روش‌های سنتی و قواعد خبرگی انجام می‌شود؛ رویکردهایی که عمدتاً بر قضاوت اعتباری، بررسی وثایق و شاخص‌های محدود مالی تکیه دارند و اگرچه می‌توانند در موارد ساده مفید باشند، اما در مواجهه با داده‌های حجیم، ناهمگن و پویا کارایی محدودتری دارند (Thomas et al., 2017). هم‌زمان، توسعه فناوری‌های داده‌محور موجب شده است که مدل‌های امتیازدهی اعتباری مبتنی بر یادگیری ماشین به‌عنوان مکمل یا جایگزین روش‌های کلاسیک مطرح شوند؛ زیرا می‌توانند روابط غیرخطی و الگوهای پیچیده میان متغیرها را شناسایی کنند (Lessmann et al., 2015). با این حال، جایگزینی کامل روش‌های تفسیرپذیر با مدل‌های پیچیده همواره مطلوب نیست؛ بانک‌ها و نهادهای ناظر معمولاً به مدل‌هایی نیاز دارند که علاوه بر دقت، قابلیت توضیح و پایداری تصمیم را نیز فراهم کنند (BCBS, 2020).

در ادبیات پیش‌بینی نکول، رگرسیون لجستیک همچنان یکی از ابزارهای اصلی است، زیرا احتمال نکول را به‌صورت صریح مدل‌سازی می‌کند و تفسیر ضرایب آن برای سیاست‌گذاری اعتباری و توجیه تصمیمات امکان‌پذیر است (Hosmer et al., 2013). با این حال، در داده‌های بانکی با تعداد زیاد ویژگی‌ها، همبستگی میان متغیرها و وجود سیگنال‌های

[1] Basel Committee on Banking Supervision

ضعیف، مدل‌های کلاسیک در معرض بیش‌برازش یا کاهش قدرت پیش‌بینی قرار می‌گیرند؛ ازاین‌رو استفاده از رگرسیون‌های جریمه‌شده برای انتخاب ویژگی و کنترل پیچیدگی، به‌عنوان راهکاری معتبر مورد توجه قرار گرفته است (Desai et al., 1996). از سوی دیگر، شبکه‌های عصبی و الگوریتم‌های مبتنی بر درخت تصمیم مانند جنگل تصادفی[۱] و XGBoost به دلیل انعطاف بالا در تقریب توابع و توانایی کشف الگوهای غیرخطی، در بسیاری از پژوهش‌ها عملکرد برتری نسبت به روش‌های سنتی گزارش کرده‌اند (Desai et al., 1996). با این وجود، چالش کلیدی آن است که مسئله نکول در عمل غالباً با «داده‌های نامتوازن» مواجه است و همین امر می‌تواند موجب شود مدل‌ها در شناسایی کلاس اقلیت (نکول) ضعیف عمل کنند (He & Garcia, 2009). بنابراین تمرکز بر معیارهای ارزیابی مناسب (مانند AUC، بازخوانی و امتیاز F۱) و به‌کارگیری راهبردهای مقابله با نامتوازنی داده نظیر تکنیک SMOTE، برای دستیابی به یک مدل کارا ضروری است (Lessmann et al., ۲۰۱۵).

پژوهش حاضر با هدف بهبود پیش‌بینی نکول مشتریان بانکی، یک چارچوب ارزیابی و مقایسه‌ای ارائه می‌کند که در آن عملکرد مدل‌های متنوع یادگیری ماشین شامل رگرسیون لجستیک جریمه‌شده، شبکه عصبی، جنگل تصادفی و XGBoost مورد تحلیل قرار می‌گیرد. در این پژوهش، از رگرسیون لجستیک جریمه‌شده به‌عنوان جزء تفسیرپذیر و از مدل‌های غیرخطی برای شکار الگوهای پیچیده رفتار مشتری استفاده می‌شود. علاوه‌بر این، برای رفع چالش نامتوازنی داده‌ها، از ترکیب راهبردهای وزن‌دهی و تکنیک بیش‌نمونه‌گیری SMOTE بهره گرفته شده است. انتظار می‌رود این رویکرد چندمدلی ضمن کاهش خطاهای نوع دوم، دقت و کارایی تصمیم‌گیری اعتباری را به‌طور معناداری ارتقا دهد. ساختار مقاله به این ترتیب است که پس از این مقدمه، ابتدا مبانی نظری و پیشینه ارائه می‌شود، سپس داده‌ها، روش‌شناسی و مدل‌های پیشنهادی تشریح می‌گردد، در ادامه نتایج تجربی و مقایسه عملکرد مدل‌ها گزارش می‌شود و در نهایت، جمع‌بندی و پیشنهادهای کاربردی ارائه خواهد شد.

مبانی نظری و پیشینه پژوهش

---

[۱] Random Forest

ریسک اعتباری یکی از مهم‌ترین انواع ریسک در نهادهای مالی است و به احتمال ناتوانی وام‌گیرنده در ایفای کامل و به‌موقع تعهدات مالی خود اشاره دارد. این نوع ریسک، هسته اصلی فعالیت واسطه‌گری مالی بانک‌ها را تحت تأثیر قرار می‌دهد، زیرا بخش عمده دارایی بانک‌ها را تسهیلات و مطالبات از مشتریان تشکیل می‌دهد. از این‌رو، هرگونه اختلال در بازپرداخت تعهدات اعتباری می‌تواند کیفیت دارایی‌ها، سودآوری، نقدشوندگی و حتی کفایت سرمایه بانک را متاثر سازد (Thomas et al., 2017). در این چارچوب، نکول به‌عنوان وضعیتی تعریف می‌شود که در آن مشتری قادر نیست اقساط یا اصل و فرع بدهی خود را مطابق شرایط قرارداد بازپرداخت کند؛ بنابراین، پیش‌بینی احتمال نکول یکی از مسائل بنیادی در مدیریت ریسک اعتباری است (Crook et al., 2007).

در ادبیات بانکداری، سنجش ریسک اعتباری عموماً در دو سطح انجام می‌شود: سطح خرد، که بر احتمال نکول مشتری یا قرارداد متمرکز است؛ و سطح کلان، که پیامدهای تجمع ریسک اعتباری را در پرتفوی یا نظام مالی بررسی می‌کند. پژوهش حاضر در سطح خرد قرار می‌گیرد و هدف آن، پیش‌بینی نکول مشتریان حقیقی بر اساس ویژگی‌های اعتباری، جمعیت‌شناختی و رفتاری است. تمرکز بر سطح خرد از آن جهت اهمیت دارد که تصمیمات اعطای تسهیلات، قیمت‌گذاری اعتباری و کنترل مطالبات غیرجاری عمدتاً در همین سطح اتخاذ می‌شوند (Henley, 1997 & Hand).

پژوهش در زمینه پیش‌بینی نکول و درماندگی مالی، از مدل‌های آماری کلاسیک آغاز شد. در مراحل اولیه، روش‌هایی مانند تحلیل تشخیصی، مدل احتمال خطی و رگرسیون لجستیک به دلیل سادگی، شفافیت و قابلیت تفسیر، کاربرد گسترده‌ای در ارزیابی اعتباری داشتند (Altman, 1968; Ohlson, 1980). رگرسیون لجستیک به‌ویژه به سبب توانایی در برآورد احتمال وقوع نکول برای متغیر وابسته دودویی، به یکی از ابزارهای اصلی امتیازدهی اعتباری تبدیل شد (Thomas et al., 2017).

با توسعه زیرساخت‌های داده و افزایش پیچیدگی رفتار مالی مشتریان، مدل‌های مبتنی بر یادگیری ماشین نیز به‌تدریج وارد حوزه ارزیابی ریسک اعتباری شدند. این مدل‌ها، از جمله درخت تصمیم، جنگل تصادفی، ماشین بردار پشتیبان، شبکه‌های عصبی مصنوعی و روش‌های تقویت‌گرادیانی، قادرند روابط غیرخطی، اثرات متقابل و الگوهای پنهان در داده‌ها را بهتر از بسیاری از مدل‌های خطی شناسایی کنند (Lessmann et al., 2015). مرورهای

نظام‌مند نشان می‌دهد که در محیط‌های داده‌ای پیچیده، روش‌های یادگیری ماشین غالباً عملکرد پیش‌بینی بالاتری نسبت به مدل‌های سنتی دارند، هرچند این برتری همواره با کاهش نسبی در تفسیرپذیری همراه است (Brown & Mues, 2012).

با این حال، ادبیات پژوهش تأکید می‌کند که هیچ الگوریتمی به‌صورت جهان‌شمول در همه زمینه‌ها بهترین عملکرد را ندارد و کارایی هر مدل به عواملی همچون کیفیت داده، ساختار متغیرها، میزان نامتوازنی کلاس‌ها و ویژگی‌های نهادی بازار بستگی دارد (Lessmann et al., 2015). از این منظر، انتخاب مدل مناسب در پیش‌بینی نکول باید نه صرفاً بر مبنای پیچیدگی الگوریتم، بلکه بر پایه تناسب آن با ماهیت مسئله و ساختار داده‌های پژوهش صورت گیرد.

پژوهش‌های تجربی در حوزه پیش‌بینی نکول و ارزیابی ریسک اعتباری نشان می‌دهد که این موضوع از دیرباز یکی از مسائل اصلی در ادبیات مالی و بانکداری بوده است. مطالعات اولیه بیشتر بر قابلیت اطلاعات مالی در پیش‌بینی درماندگی و ناتوانی در ایفای تعهدات متمرکز بودند. در این چارچوب، بیور با تحلیل نسبت‌های مالی نشان داد که برخی شاخص‌های حسابداری توانایی مناسبی در تفکیک واحدهای سالم و درمانده دارند (Beaver, 1966). پس از آن، آلتمن با ارائه مدل Z-Score نشان داد که ترکیب چند نسبت مالی می‌تواند قدرت پیش‌بینی ورشکستگی را به‌طور معناداری افزایش دهد (Altman, 1968). این مطالعات مبنای توسعه مدل‌های کمی در سنجش ریسک اعتباری و پیش‌بینی نکول شدند.

در ادامه، رگرسیون لجستیک به یکی از پرکاربردترین روش‌ها در پیش‌بینی نکول تبدیل شد؛ زیرا ضمن برآورد احتمال نکول، امکان تفسیر اثر متغیرهای توضیحی را نیز فراهم می‌کند (Hosmer et al., 2013). با این حال، گسترش داده‌های اعتباری و پیچیده‌تر شدن روابط میان متغیرها، موجب شد پژوهشگران به استفاده از روش‌های هوش مصنوعی و یادگیری ماشین روی آورند. در همین راستا، دسای و همکاران نشان دادند که شبکه‌های عصبی در پیش‌بینی درماندگی مالی و ریسک اعتباری می‌توانند عملکردی رقابتی و در برخی موارد برتر از روش‌های سنتی داشته باشند (Desai et al., 1996). ساندرز و آلن نیز با طرح رویکردهای نوین سنجش ریسک اعتباری، بر لزوم استفاده از الگوهای کمی پیشرفته در تحلیل زیان‌های اعتباری و مدیریت ریسک تأکید کردند (Altman & Saunders, 1997).

در سال‌های بعد، مطالعات مقایسه‌ای متعددی برای ارزیابی کارایی الگوریتم‌های مختلف اعتبارسنجی انجام شد. لسمن و همکاران با مقایسه گسترده روش‌های یادگیری ماشین در امتیازدهی اعتباری نشان دادند که الگوریتم‌هایی مانند جنگل تصادفی، ماشین بردار پشتیبان و شبکه‌های عصبی، در بسیاری از موارد دقت پیش‌بینی بالاتری نسبت به مدل‌های سنتی دارند (Lessmann et al., 2015). این یافته‌ها نشان داد که برای شناسایی الگوهای پیچیده، به‌ویژه در داده‌های بزرگ و غیرخطی، اتکا به یک روش کلاسیک به‌تنهایی کافی نیست. افزون بر این، پژوهش‌های جدیدتر بر نقش داده‌های بزرگ، اطلاعات تراکنشی و تحلیل‌های بلادرنگ در ارتقای دقت پیش‌بینی نکول تأکید کرده‌اند (Nahar et al., 2024). همچنین، برخی مطالعات معاصر نشان داده‌اند که ترکیب داده‌های سنتی با داده‌های رفتاری و غیرسنتی می‌تواند قدرت تفکیک مدل‌های اعتبارسنجی را افزایش دهد(Zhang, Yu, 2024).

در ادبیات داخلی نیز مطالعات متعددی به شناسایی عوامل مؤثر بر بازپرداخت و نکول مشتریان و نیز طراحی مدل‌های اعتبارسنجی در نهادهای مالی و حمایتی پرداخته‌اند. اصغری و احمدی (۱۳۹۷) با ارائه مدلی ترکیبی به بررسی عملکرد بازپرداخت تسهیلات در صندوق کارآفرینی امید پرداختند؛ آن‌ها ابتدا با استفاده از تحلیل عاملی پابرجا (Robust PCA) و الگوریتم K-Means داده‌ها را خوشه‌بندی کرده و سپس برای پیش‌بینی وضعیت اعتباری متقاضیان، مدل‌های ماشین بردار پشتیبان (SVM) و الگوریتم ترکیبی شبکه عصبی-ژنتیک را مقایسه کردند و نشان دادند که مدل ترکیبی شبکه عصبی-ژنتیک از صحت و دقت پیش‌بینی بالاتری برخوردار است. فلاح شمس و مهدوی راد (۱۳۹۱) با بررسی مشتریان حقیقی شرکت لیزینگ ایران‌خودرو، پنج متغیر کلیدی شامل درآمد ماهیانه متقاضی، مدت و میزان تسهیلات، درآمد ماهیانه ضامن و سابقه کار را به‌عنوان عوامل مؤثر بر ریسک اعتباری شناسایی کردند و با مقایسه مدل‌های اقتصادسنجی لاجیت و پروبیت، نشان دادند که مدل لاجیت کارایی و دقت بالاتری در پیش‌بینی ریسک اعتباری و طبقه‌بندی مشتریان دارد.

در همین راستا، بهزادی‌راد و همکاران با استفاده از رگرسیون گسسته نشان دادند که متغیرهایی مانند سن، جنسیت، مبلغ تسهیلات، نرخ سود و نوع وثیقه، اثر معناداری بر احتمال بازپرداخت به‌موقع دارند (بهزادی‌راد و همکاران، ۱۴۰۳). محرابی و همکاران با طراحی یک الگوریتم فازی برای رتبه‌بندی اعتباری، تلاش کردند تصمیم‌گیری کارشناسی را به چارچوبی کمی و ساخت‌یافته تبدیل کنند و نشان دادند که استفاده از منطق فازی می‌تواند

در شرایط عدم‌قطعیت به بهبود ارزیابی اعتباری کمک کند (محرابی و همکاران، ۱۴۰۲). شجاع و شوشاد با بهره‌گیری از نظریه مقدار حدی نشان دادند که درماندگی مالی شرکت‌ها می‌تواند از مسیر ارتباطات اعتباری به بخش بانکی سرایت کند و ریسک نکول را در سطح شبکه تشدید سازد (شجاع و شوشاد، ۱۴۰۰). خرمی و همکاران نیز با ترکیب مدل CBR و روش‌های تصمیم‌گیری چندمعیاره، متغیرهایی مانند چک برگشتی، درآمد و شاخص‌های رفتاری را در ارزیابی ریسک اعتباری مؤثر دانستند (خرمی و همکاران، ۱۳۹۹). افزون بر این، جندقی و باقری با استفاده از شبکه‌های عصبی و الگوریتم‌های فراابتکاری نشان دادند که روش‌های هوشمند می‌توانند در پیش‌بینی ریسک اعتباری شرکت‌ها و تعاونی‌ها دقت بالایی ارائه کنند (جندقی و باقری، ۱۳۹۹). همچنین، کمالی و همکاران گزارش کردند که مدل ZPP در پیش‌بینی نکول عملکرد بهتری نسبت به مدل ساختاری KMV ارائه می‌دهد.

جمع‌بندی مطالعات پیشین نشان می‌دهد که پژوهش‌های این حوزه از دو جهت توسعه یافته‌اند: از یک‌سو، دامنه داده‌ها از نسبت‌های مالی سنتی به سمت داده‌های رفتاری، ویژگی‌های متقاضی و ضامن و داده‌های تراکنشی گسترش یافته است؛ و از سوی دیگر، روش‌ها از مدل‌های کلاسیک آماری مانند لاجیت و پروبیت به سمت الگوریتم‌های پیشرفته یادگیری ماشین، از جمله شبکه‌های عصبی، جنگل تصادفی وXGBoost ، حرکت کرده‌اند. با این حال، سه خلأ مهم همچنان باقی است: نخست، شکاف میان تفسیرپذیری و دقت پیش‌بینی؛ دوم، نامتوازن بودن شدید داده‌های اعتباری و غفلت از راهکارهای مناسب برای مقابله با آن؛ و سوم، محدود بودن استفاده از داده‌های رفتاری در بسیاری از مطالعات. بر این اساس، پژوهش حاضر می‌کوشد با استفاده از داده‌های سنتی و رفتاری، به‌کارگیری SMOTEو مقایسه چند مدل یادگیری ماشین، هم دقت پیش‌بینی نکول را بهبود دهد و هم توازن بهتری میان کارایی و قابلیت تفسیر ایجاد کند.

روش

پژوهش حاضر از نظر هدف، کاربردی و از نظر ماهیت و روش، کمی-مقایسه‌ای است. جامعه آماری این پژوهش شامل کلیه مشتریان حقیقی بانک قرض‌الحسنه رسالت است که در قالب «طرح شناخت اجتماعی» در بازه زمانی ۱۳۹۹/۰۲/۰۲ تا ۱۴۰۳/۱۲/۲۸ اقدام به دریافت تسهیلات نموده‌اند. روش نمونه‌گیری در این مطالعه به صورت کل‌شماری بوده و بر این اساس، اطلاعات مربوط به تمامی ۸۱،۵۹۷ پرونده تسهیلاتی ثبت‌شده در سامانه

بانکداری متمرکز بانک رسالت در بازه زمانی مذکور استخراج و مورد تحلیل قرار گرفت. انتخاب این بازه پنج‌ساله با هدف پوشش دادن چرخه‌های مختلف بازپرداخت و دسترسی به کامل‌ترین بانک اطلاعاتی دیجیتال طرح شناخت اجتماعی انجام شده است.

مجموعه‌داده مورد استفاده ترکیبی از متغیرهای سنتی اعتباری (اطلاعات جمعیت‌شناختی و عملکرد تسهیلات) و متغیرهای رفتاری غیرسنتی (استخراج‌شده از الگوی استفاده از خدمات دیجیتال) است. متغیرهای مستقل عددی و طبقه‌ای مورد استفاده در جدول (۱) تشریح شده‌اند.

جدول ۱. متغیرهای اصلی مورد استفاده در مدل پیش‌بینی نکول تسهیلات

نام متغیر	نوع داده	توضیح / مفهوم	نام متغیر	نوع داده	توضیح / مفهوم
AMT	عددی	مبلغ تسهیلات اعطایی (میلیون ریال)	TOT_INST_DEL	عددی	مجموع تعداد اقساط دارای تأخیر
INST_N	عددی	تعداد کل اقساط تسهیلات	BAL	عددی	مانده بدهی مشتری (میلیون ریال)
INST_AMT	عددی	مبلغ هر قسط (میلیون ریال)	AGE	عددی	سن مشتری در زمان دریافت تسهیلات
INST_PAID	عددی	تعداد کل اقساط پرداخت‌شده	EDU_LVL	طبقه‌ای	سطح تحصیلات مشتری
INST_ONTIME	عددی	تعداد اقساط پرداخت‌شده در موعد مقرر	LIT	طبقه‌ای	وضعیت سواد
risk_status	دودویی	وضعیت ریسک مشتری	GND	طبقه‌ای	جنسیت مشتری

متغیر وابسته پژوهش، وضعیت ریسک مشتری است که به صورت یک متغیر دودویی (باکلاس‌های ۰ و ۱) تعریف می‌شود. معیار تعریف نکول بر اساس آستانه ۹۰ روز تأخیر متوالی در پرداخت حداقل یکی از اقساط تسهیلات در طول دوره قرارداد (مطابق با استانداردهای کمیته بال) تدوین شده است. در این راستا، تمامی تسهیلاتی که در بازه زمانی مطالعه سررسید شده‌اند وارد مدل شده‌اند. مشتریانی که پیش از اتمام دوره تسهیلات، در هر

مقطعی سابقه تأخیر متوالی بالای ۹۰ روز داشته‌اند، حتی در صورت تسویه کامل در مراحل بعدی، همچنان به عنوان نکول‌کننده (کلاس ۱) طبقه‌بندی می‌شوند تا رفتار ریسکی واقعی آن‌ها در طول دوره اعتبار منعکس گردد.

✦ داده‌های نامتوازن در پیش‌بینی نکول

داده‌های نامتوازن به مجموعه‌داده‌ای گفته می‌شود که در آن تعداد مشاهدات متعلق به طبقات مختلف تفاوت زیادی دارد. در داده‌های اعتباری، معمولاً تعداد مشتریان غیرنکول بسیار بیشتر از مشتریان نکول است. این ویژگی موجب می‌شود الگوریتم‌های طبقه‌بندی متعارف، به‌دلیل تمرکز بر طبقه اکثریت، عملکرد مناسبی در شناسایی طبقه اقلیت نداشته باشند (Reychav et al., 2019). برای مقابله با نامتوازنی داده‌ها از دو رویکرد اصلی استفاده شده است: بیش‌نمونه‌گیری به روش SMOTE و وزن‌دهی طبقات در مدل رگرسیون لجستیک.

۱- بیش‌نمونه‌گیری[1] به روش SMOTE

روش SMOTE یکی از پرکاربردترین روش‌های متوازن‌سازی داده‌های نامتوازن است که با ایجاد نمونه‌های مصنوعی برای طبقه اقلیت، تعداد مشاهدات این طبقه را در داده‌های آموزشی افزایش می‌دهد (Chawla et al., 2002). در این روش، برای هر مشاهده از طبقه اقلیت، تعدادی از نزدیک‌ترین همسایه‌ها انتخاب می‌شوند و نمونه‌های جدید در امتداد خط بین آن مشاهده و همسایه‌هایش تولید می‌گردند. رابطه تولید نمونه مصنوعی به‌صورت زیر است:

$$x_{new} = x_i + \delta(x_{zi} - x_i), \quad \delta \sim U(0,1) \qquad (۱)$$

که در آن، x_i یک نمونه از طبقه اقلیت و x_{zi}یکی از نزدیک‌ترین همسایه‌های آن است.

۲- وزندهی طبقات

علاوه بر بیش‌نمونه‌گیری، یکی دیگر از راهکارهای مناسب برای مواجهه با داده‌های نامتوازن، استفاده از وزن‌دهی به طبقات در فرایند یادگیری مدل است. در این روش، برای خطاهای مربوط به طبقه اقلیت وزن بیشتری در نظر گرفته می‌شود تا مدل در فرایند برازش، توجه بیشتری به شناسایی موارد نکول داشته باشد (He and Garcia, 2009). در این پژوهش، این رویکرد در مدل رگرسیون لجستیک به‌کار گرفته شده و نتایج آن در قالب حالت‌های «با وزن‌دهی» و «بدون وزن‌دهی» گزارش شده است.

[1] Oversampling

✦ ساخت مدل پیش‌بینی نکول تسهیلات

۱- رگرسیون لجستیک با تنظیم[1]:

رگرسیون لجستیک برای تبدیل نتایج رگرسیون خطی با دامنه‌ی ((-∞, ∞)) به بازه‌ی (0, 1) از یک تابع-S شکل[2] استفاده می‌کند. در رگرسیون لجستیک، هدف این است که مقدار (Y) به احتمال بین ۰ و ۱ تبدیل شود تا نشان دهد احتمال تعلق نمونه به طبقه‌ی مثبت (مثلاً «نکول» در تسهیلات) چقدر است. این تابع نرمال‌سازی‌کننده، تابع لجستیک یا تابع سیگموئید[3] نام دارد. اصل الگوریتم رگرسیون لجستیک در شکل ۱ نمایش داده شده است و فرمول تابع آن به‌صورت زیر است:

$$g(z) = \frac{1}{1+e^{-z}} \qquad (۲)$$

رگرسیون لجستیک تابعی از e به‌صورت $y = \frac{1}{1+e(-x)}$ است؛ یعنی:

- هنگامی که $x > 0$، با افزایش x، مقدار y به‌سرعت به ۱ نزدیک می‌شود.
- اگر $x < 0$، با کاهش x، مقدار y به‌سرعت به ۰ نزدیک می‌شود.
- اگر $x = 0, y = \frac{1}{2}$

به دلیل ویژگی خاص رگرسیون لجستیک (که خروجی آن پیوسته و بین بازه‌ی ۰ تا ۱ است)، از آن برای ارزیابی صحت عملکرد الگوریتم‌های یادگیری استفاده می‌شود.

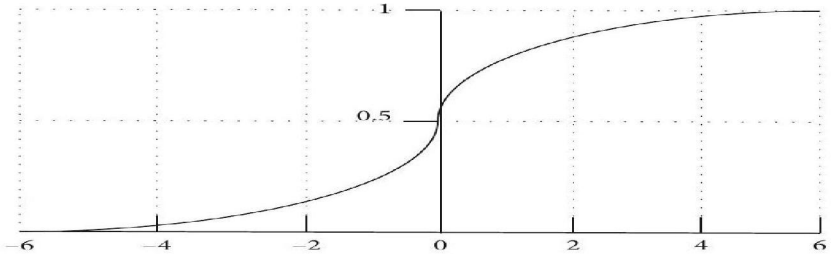
مزیت اصلی رگرسیون لجستیک، علاوه بر تشخیص درستی یا نادرستی نتایج، این است که می‌تواند میزان فاصله‌ی نتیجه‌ی فعلی از پاسخ درست را نیز اندازه‌گیری کند و از این طریق، فرآیند آموزش مدل را در جهت درست هدایت کند. بنابراین، معمولاً این الگوریتم در کنار الگوریتم استیگ[4] برای بهینه‌سازی استفاده می‌شود.

رگرسیون لجستیک به دلیل سادگی و قابلیت تفسیرپذیری، یکی از مدل‌های مرجع در پیش‌بینی ریسک نکول به شمار می‌رود(Yu et al., 2019) . با این حال، در مسائل اعتباری با نرخ نکول پایین و داده‌های نامتوازن، عملکرد مدل ساده با محدودیت مواجه است؛ از این رو در این پژوهش، از نسخه منظم‌شده رگرسیون لجستیک به همراه وزندهی طبقات استفاده شده است.

---

[1] Regularized Logistic Regression
[2] Sigmoid Function
[3] Sigmoid
[4] Steig

شکل ۱. رگرسیون لجستیک



ضریب جریمه[1]:

برای جلوگیری از بیش‌برازش، به تابع هزینه مدل یک جمله جریمه افزوده می‌شود تا بزرگی ضرایب محدود شود. L1 وL2 ، مفاهیم کنترلی هستند که به عنوان جریمه[2] به تابع هزینه اضافه می‌شوند تا پارامترهای مدل محدود شوند و از بیش‌برازش جلوگیری شود. برای مدل‌های رگرسیون خطی، رابطه (۳) تابع هزینه رگرسیون Lasso و رابطه (۴) تابع هزینه رگرسیون Ridge است.

$$\min_{w} \frac{1}{2n_{\text{samples}}} \left\|X_{w-y}\right\|_2^2 + \alpha\|w\|_1 \qquad (۳)$$

$$\min_{w} \left\|X_{w-y}\right\|_2^2 + \alpha\|w\|_2^2 \qquad (۴)$$

که در آن، W نشان‌دهنده‌ی ضریب ویژگی‌ها[3] و α پارامتر تنظیم (ضریب جریمه) تعریف می‌شود. به‌طور کلی، L1 می‌تواند از مدل‌های کم‌هزینه و ساده[4] برای انتخاب ویژگی‌ها و L2 از بیش‌برازش مدل‌ها جلوگیری می‌کند.

الف- تنظیم L1 و انتخاب ویژگی‌ها: فرض کنید تابع هزینه زیر با تنظیم L1 داشته باشیم، همانند فرمول زیر:

$$J = J_0 + \alpha \sum_W |W| \qquad (۵)$$

یافتن کمینه تابع هزینه با استفاده از روش‌های مختلف یکی از وظایف مهم یادگیری ماشین است. همان‌طور که در شکل ۱نشان داده شده است، ضریب تنظیم α می‌تواند اندازه نمودار را به خوبی کنترل کند. هرچه منحنی α کوچک‌تر باشد، منحنی l بزرگ‌تر می‌شود (شکل ۲)، هرچه نمودار α بزرگ‌تر باشد، منحنی l کوچک‌تر می‌شود. بنابراین، جعبه سیاه تنها یک

[1] Penalty Coefficient
[2] Sanctions
[3] Characteristic Coefficient
[4] Frugal Model

نقطه بعد از مبدأ خواهد بود. این مناسب‌ترین مقدار است $(W1, W2) = (0, W)$ که (W) می‌تواند مقدار بسیار کوچکی بگیرد.

برای پارامتر تنظیم L1 ، تحت شرایط عادی، هرچه ورودی بزرگ‌تر باشد و پارامتر برابر صفر باشد، تابع جریمه می‌تواند کمینه خود را به دست آورد. به عنوان مثال ساده، فرض کنید تابع هزینه زیر با ترم تنظیم L1 داشته باشیم، همانند رابطه (۶):

$$F(x) = f(x) + \lambda\|x\|_1 \qquad (۶)$$

که در آن X پارامتری است که باید تخمین زده شود و متناظر با W و ۰ در بالا است. توجه داشته باشید که تنظیم L1 در برخی نقاط قابل مشتق‌گیری نیست، و اگر ورودی به اندازه کافی بزرگ باشد، و $(x = 0)$ کمینه تابع باشد، آنگاه می‌توان $f(x)$ را گرفت، همان‌طور که در شکل ۳ نشان داده شده است. وقتی $(x = 0)$ ، هرچه مقدار بزرگ‌تر باشد، دستیابی به کمینه $f(x)$ آسان‌تر است.

ب- تنظیم L2 و بیش‌برازش: فرض کنید تابع هزینه‌ای با قاعده L2 وجود دارد، همانند فرمول زیر:

$$J = J_0 + \alpha \sum_w w^2 \qquad (۷)$$

در مدل‌های رگرسیون خطی، یکی از چالش‌های اصلی، حساسیت ضرایب مدل (θ_j) به نوسانات کوچک در داده‌های آموزشی است. هنگامی که ضرایب مدل مقادیر بسیار بزرگی می‌گیرند، مدل به جای یادگیری الگوی کلی، به نویزهای موجود در داده‌ها حساس شده و دچار بیش‌برازش می‌شود؛ در این حالت، حتی یک تغییر جزئی در داده‌ها می‌تواند تأثیر نامتناسبی بر خروجی مدل داشته باشد.

برای مقابله با این پدیده، از روش‌های منظم‌سازی[1] استفاده می‌شود. تنظیم L2 با اضافه کردن مجازات به تابع هزینه، ضرایب (θ_j) را وادار می‌کند تا مقادیر کوچک‌تری اختیار کنند. کاهش بزرگیِ ضرایب، پایداری مدل را افزایش داده و حساسیت آن را نسبت به تغییرات کوچک در داده‌های ورودی به حداقل می‌رساند. با استفاده از این فرمول در روش نزول گرادیان، ضرایب در هر تکرار به تدریج کوچک می‌شوند. این مکانیزم به ما اجازه می‌دهد تا پیچیدگی مدل را کنترل کرده و از اثرات مخرب نوسانات داده بر نتایج پیش‌بینی نکول بکاهیم.

[1] Regularization

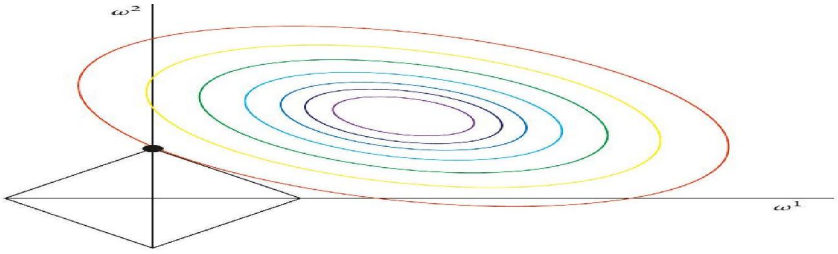
فرض کنید پارامتر مورد نیاز برابر صفر باشد، $h_\theta(x)$ تابع فرضی است و تابع هزینه رگرسیون خطی به صورت زیر تعریف می‌شود:

$$J(\theta) = \frac{1}{2\,m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 \qquad (۸)$$

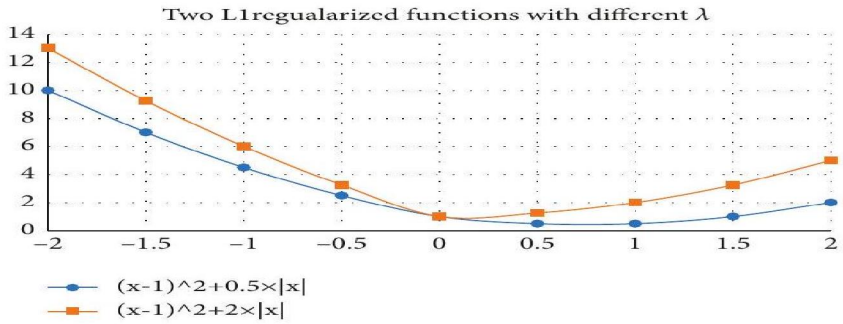
و قاعده بهروزرسانی نزول گرادیان برای پارامتر θ_j عبارت است از:

$$\theta_j = \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x_j^{(i)} \qquad (۹)$$

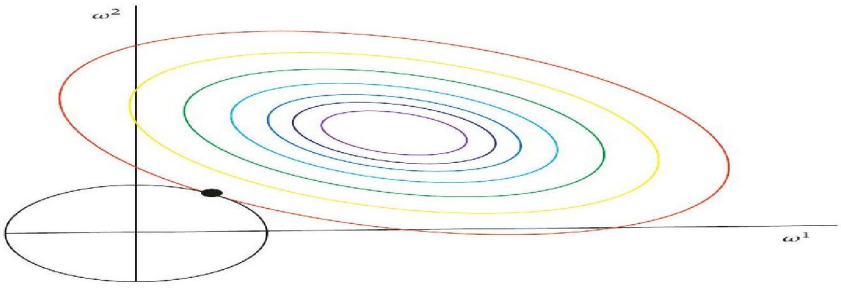
شکل ۲. تنظیم L1



شکل ۳. انتخاب پارامترهای منظم‌سازی



شکل ٤. تنظیمL2



اگر عنصر تنظیم L2 بعد از تابع هزینه اولیه اضافه شود، فرمول تکراری به شکل زیر خواهد بود:

$$\theta_j := \theta_j \left(1 - \alpha \frac{\lambda}{m}\right) - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)} \qquad (۱۰)$$

پارامترهای تنظیم در فرمول لحاظ شده‌اند. θ_j در هر تکرار در ضرب یک ضریب کمتر از ۱ قرار می‌گیرد تا به تدریج کاهش یابد، بنابراین θ_j معمولاً کاهش پیدا می‌کند. برای پارامتر تنظیمL2 ، هرچه ورودی بزرگ‌تر باشد، θ_j سریع‌تر کاهش می‌یابد. مطابق شکل ۴، هرچه ورودی فرمول بزرگ‌تر باشد، شعاع دایره L2 کوچکتر خواهد بود. اگر تابع هزینه نهایی حداکثر باشد، پارامترهای آن نیز بسیار کوچک می‌شوند.

برای حل مشکل پیش‌بینی نکول اعتباری ناشی از داده‌های بسیار نامتوازن، نتایج آموزش و پیش‌بینی مدل به راحتی قابل تنظیم هستند. بنابراین، دوره تنظیم L2 برای تحلیل مناسب‌تر است. در آزمایش، L1 و L2 از طریق تجربه و آزمایش، برای انتخاب مناسب‌ترین ضریب جریمه تنظیم می‌شوند.

۲- جنگل تصادفی[1]

جنگل تصادفی یک روش یادگیری جمعی مبتنی بر «مجموعه‌ای»[2] مجموعه از درخت‌های تصمیم است که با استفاده از نمونه‌گیری تصادفی و ترکیب چندین درخت، دقت پیش‌بینی را افزایش می‌دهد و از بیش‌برازش می‌کاهد (Breiman, 2001) . پیش‌بینی نهایی از طریق رأی‌گیری اکثریت[3] در میان درخت‌ها به‌دست می‌آید:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\} \qquad (۱۱)$$

که در آن $h_t(x)$ پیش‌بینی درخت t–ام است. این فرایند باعث کاهش واریانس و جلوگیری از بیش‌برازش[4] می‌شود.

۳-الگوریتم XGBoost

XGBoost یک الگوریتم تقویت گرادیانی است که با افزودن متوالی درخت‌های ضعیف، تابع هدف را بهینه می‌کند. در هر مرحله t، مدل درختی اضافه می‌شود که خطای باقی‌مانده[5] مدل قبلی را کاهش دهد(Chen & Guestrin, 2016):

---

[1] Random Forest
[2] Ensemble
[3] Majority Voting
[4] Overfitting
[5] Residual

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \qquad (۱۲)$$

که در آن l تابع زیان و Ω عبارت منظم‌ساز[1] برای کنترل پیچیدگی مدل است. این الگوریتم به‌دلیل توانایی بالا در شناسایی روابط غیرخطی و الگوهای پیچیده، در مسائل پیش‌بینی ریسک اعتباری عملکرد بسیار مطلوبی دارد (Lessmann et al., 2015).

✦ شاخص‌های ارزیابی مدل

با توجه به نامتوازن‌بودن داده‌ها و اهمیت شناسایی صحیح مشتریان نکول‌کرده، ارزیابی عملکرد مدل‌ها در این پژوهش صرفاً بر اساس دقت کلی انجام نشده است. بر این اساس، مجموعه‌ای از شاخص‌های متداول و مناسب برای مسائل طبقه‌بندی دودویی نامتوازن استفاده شده است.

۱- ماتریس درهم‌ریختگی[2]

ماتریس درهم‌ریختگی، مبنای اصلی محاسبه شاخص‌های ارزیابی است و شامل چهار حالت زیر است:

مثبت واقعی (TP): مشتری نکول‌کرده که مدل به‌درستی او را نکول‌کرده پیش‌بینی کرده است.

منفی واقعی (TN): مشتری غیرنکول که مدل به‌درستی او را غیرنکول پیش‌بینی کرده است.

مثبت کاذب (FP): مشتری غیرنکول که مدل به‌اشتباه او را نکول‌کرده پیش‌بینی کرده است.

منفی کاذب (FN): مشتری نکول‌کرده که مدل به‌اشتباه او را غیرنکول پیش‌بینی کرده است.

۲- دقت: نسبت کل پیش‌بینی‌های درست به کل مشاهدات را نشان می‌دهد:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (۱۳)$$

هرچند این شاخص در ارزیابی کلی مدل مفید است، اما در داده‌های نامتوازن به‌تنهایی کافی نیست (Brown & Mues, 2012).

---

[1] Regularization
[2] Confusion Matrix

۳- صحت: نشان می‌دهد از میان تمام مشاهداتی که مدل به‌عنوان نکول پیش‌بینی کرده، چه نسبتی واقعاً نکول بوده‌اند:

$$Precision = \frac{TP}{TP+FP} \quad (۱۴)$$

۴- فراخوانی: بیانگر توانایی مدل در شناسایی مشتریان نکول‌کرده واقعی است و در مسئله ریسک اعتباری، یکی از مهم‌ترین شاخص‌ها محسوب می‌شود:

$$Recall = \frac{TP}{TP+FN} \quad (۱۵)$$

۵- امتیاز F1: میانگین هارمونیک صحت و فراخوانی است و برای ارزیابی تعادل میان این دو شاخص به‌کار می‌رود:

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (۱۶)$$

یافته‌ها

میانگین مبلغ تسهیلات حدود ۹۴۶ میلیون ریال و میانه آن ۱۰۰۰ میلیون ریال است که نشان می‌دهد بخش عمده تسهیلات‌ها در همین محدوده قرار دارند، هرچند دامنه بسیار وسیع مقادیر، وجود تسهیلات‌های خرد و کلان را هم‌زمان نشان می‌دهد. تعداد اقساط نیز تنوع بالایی دارد و از تسهیلات‌های کوتاه‌مدت تا بلندمدت را دربر می‌گیرد. در مورد مانده بدهی، میانه صفر بیانگر آن است که بیشتر مشتریان بدهی خود را تسویه کرده‌اند، اما در برخی موارد مانده‌های بسیار بالا مشاهده می‌شود. الگوی پرداخت نیز نسبتاً مناسب است و بیشتر اقساط به‌موقع پرداخت شده‌اند. از نظر جمعیت‌شناختی، مشتریان عمدتاً میانسال، باسواد و از نظر جنسیتی نسبتاً متعادل هستند. در مجموع، داده‌ها ناهمگن و چوله‌اند و برای تحلیل ریسک اعتباری نیاز به پیش‌پردازش دقیق دارند.

جدول ۲. آمار توصیفی متغیرهای تحقیق

متغیر	میانگین	میانه	انحراف معیار	حداقل	حداکثر
AMT	946	1,000	767	1	4,000
INST_N	16.39	12	14.09	1	144
BAL	164	0	443	0	4,134
TOT_INST_DEL	354.85	2	1,473	0	18,628
INST_ONTIME	11.31	10	9.52	0	120
INST_PAID	13.57	12	9.91	0	120
INST_AMT	265	65	542	0	4,134

2	1	0.49	1	1.41	GND
7	0	1.34	3	3	EDU_LVL
1	0	0.08	1	0.99	LIT
96	14	12.15	42	44.74	AGE

ماخذ: یافته‌های پژوهش

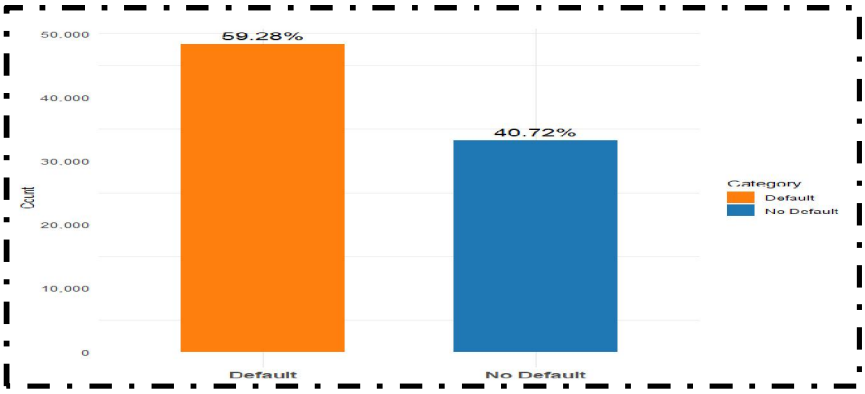
شکل (۵) توزیع متغیر وابسته وضعیت ریسک را نشان می‌دهد که در این پژوهش به‌عنوان شاخص نهایی برای تعیین احتمال نکول مشتریان بانکی تعریف شده است. بر اساس تعریف تحقیق، چنانچه پرونده تسهیلات یکی از وضعیت‌های *مشکوک‌الوصول*، *معوق* یا *سررسید گذشته* را داشته باشد، به‌عنوان نکول (1) و در غیر این صورت، یعنی در وضعیت *جاری* یا *خوش‌حساب*، به‌عنوان غیرنکول (0) طبقه‌بندی شده است.

نتایج حاصل از توزیع داده‌ها در شکل نشان می‌دهد که نسبت مشتریان خوش‌حساب به نکول‌کننده‌ها بسیار نامتوازن است. به‌طور معمول، سهم بالایی از داده‌ها مربوط به مشتریان خوش‌حساب بوده و تنها درصد کمی از کل نمونه‌ها دارای وضعیت نکول هستند. این موضوع حاکی از آن است که داده‌های مورد استفاده در پژوهش دارای عدم توازن قابل توجهی هستند؛ به بیان دیگر، حجم داده‌های متعلق به طبقه "غیرنکول" چندین برابر طبقه "نکول" است.

چنین وضعیتی در داده‌های بانکی امری رایج است، زیرا در عمل اغلب تسهیلات‌گیرندگان اقساط خود را به موقع پرداخت می‌کنند و تنها بخش کوچکی دچار تأخیر یا نکول می‌شوند. با این حال، همین اقلیت نکول‌کننده برای نظام بانکی اهمیت حیاتی دارد، زیرا می‌توانند زیان‌های مالی قابل توجهی ایجاد کنند. بنابراین، برای افزایش دقت پیش‌بینی در این شرایط، در مراحل بعدی مدل‌سازی از روش‌های وزندهی یا الگوریتم‌های مقاوم در برابر داده‌های نامتوازن مانند رگرسیون لجستیک وزن دار شده یا *SMOTE* استفاده می‌شود تا اثر طبقه‌ی اقلیت به درستی در فرآیند یادگیری مدل لحاظ شود.

در مجموع، تحلیل شکل ۵ نشان می‌دهد که بانک در اکثر موارد با مشتریان خوش‌حساب سروکار دارد، اما شناسایی و پیش‌بینی همان گروه کوچک از مشتریان پرریسک می‌تواند نقش کلیدی در بهبود سیاست‌های اعتباری و مدیریت ریسک بانک ایفا کند.

شکل ۵. توزیع طبقه‌بندی متغیر وابسته وضعیت ریسک



ماخذ: یافته‌های پژوهش

جدول ۳ نتایج آزمایش مدل طبقه‌بندی را در چهار حالت مختلف ترکیبی از ضرایب جریمه (L1 و L2) و اعمال وزن‌دهی کلاس‌ها نمایش می‌دهد. معیارهای ارزیابی شامل دقت، صحت، یادآوری و F1 هستند که هر یک جنبه‌ای از کیفیت پیش‌بینی مدل را نشان می‌دهند. ماتریس آزمون در قالب [[TN, FP], [FN, TP]] ارائه شده که در آن TN نشان‌دهنده پیش‌بینی درست منفی، FPخطای مثبت کاذب، FNخطای منفی کاذب و TP پیش‌بینی درست مثبت است.

در حالت اول، مدل با ضریب جریمه L1 و بدون اعمال وزن‌دهی اجرا شده است. این مدل توانسته با دقت کامل (۱.۰۰)، یادآوری ۰.۹۸۷۱ و F1 معادل ۰.۹۹۳۵، عملکردی تقریباً بی‌نقص ارائه دهد. ماتریس آزمون نشان می‌دهد که تنها ۵ مورد FP و هیچ مورد FN وجود داشته است، که بیانگر قدرت بالای مدل در تشخیص صحیح نمونه‌هاست.

در حالت دوم، ضریب جریمه L2 بدون وزن‌دهی اعمال شده است. این مدل نیز عملکردی بسیار نزدیک به حالت اول دارد؛ با دقت ۰.۹۹۷۴، یادآوری ۰.۹۸۹۷ و F1 برابر با ۰.۹۹۳۵. تعداد خطاها اندکی کمتر شده و تنها یک مورد FN ثبت شده است. این مدل از نظر یادآوری اندکی بهتر از L1 عمل کرده، هرچند دقت آن کمی پایین‌تر است.

در حالت سوم، مدل با ضریب جریمه L1 و اعمال وزن‌دهی اجرا شده است. با وجود حفظ نسبی عملکرد، دقت به ۰.۹۶۶۸ و F1 به ۰.۹۷۱۷ کاهش یافته است. تعداد خطاها افزایش یافته و ۹ مورد FP و ۱۳ مورد FN ثبت شده‌اند. این نشان می‌دهد که وزن‌دهی باعث کاهش دقت مدل شده، هرچند یادآوری همچنان در سطح قابل قبول باقی مانده است.

جدول ۳. مقایسه عملکرد مدل در چهار حالت ترکیبی از ضرایب جریمه و وزن‌دهی

مدل	ماتریس آزمون	دقت	یادآوری	صحت	F1
L1بدون وزن‌دهی	[[15932, 5], [0, 382]]	1	0.9871	0.9997	0.9935
L2بدون وزن‌دهی	[[15931, 4], [1, 383]]	0.9974	0.9897	0.9997	0.9935
L1با وزن‌دهی	[[15919, 9], [13, 378]]	0.9668	0.9767	0.9987	0.9717
L2با وزن‌دهی	[[15910, 58], [22, 329]]	0.9373	0.8501	0.9951	0.8916

ماخذ: یافته‌های پژوهش

در حالت چهارم، مدل با ضریب جریمه L2 و همراه با وزن‌دهی اجرا شده است. این مدل ضعیف‌ترین عملکرد را در میان چهار حالت دارد؛ دقت ۰.۹۳۷۳، یادآوری ۰.۸۵۰۱ و F1 برابر با ۰.۸۹۱۶. تعداد خطاها به‌طور قابل توجهی افزایش یافته و ۵۸ مورد FP و ۲۲ مورد FNثبت شده‌اند. این افت در عملکرد نشان می‌دهد که ترکیب L2 با وزن‌دهی برای این مجموعه داده مناسب نبوده است.

مقایسه کلی چهار حالت نشان می‌دهد که مدل‌هایی که بدون وزن‌دهی اجرا شده‌اند، به‌ویژه L1، عملکرد بسیار بهتری دارند. اعمال وزن‌دهی در هر دو حالت باعث افت دقت و F1 شده است، اما این افت در مدل L2 شدیدتر بوده و منجر به کاهش قابل توجه در یادآوری نیز شده است. از نظر ضریب جریمه، L1در هر دو حالت بهتر از L2 عمل کرده است، به‌ویژه زمانی که وزن‌دهی اعمال نشده باشد.

در نتیجه، بهترین عملکرد پیش‌بینی در این آزمایش با مدل رگرسیون لجستیک با ضریب جریمه L1 و بدون وزن‌دهی حاصل شده است. استفاده از وزن‌دهی، به‌ویژه همراه باL2 ، برای این مجموعه داده توصیه نمی‌شود، زیرا باعث کاهش قابل توجه در کیفیت پیش‌بینی می‌شود.

جدول ۴ مقایسه مدل‌های یادگیری ماشین نشان می‌دهد: مدل XGBoost بهترین عملکرد را در این جدول نشان می‌دهد. دقت برابر با ۱ است که نشان‌دهنده توانایی کامل مدل در پیش‌بینی درست نمونه‌ها است. یادآوری برابر ۰.۹۸۹۷ و دقت برابر ۰.۹۹۷۴ است، به این معنی که مدل تقریباً تمام نمونه‌های مثبت را شناسایی کرده و تعداد بسیار کمی خطای مثبت دارد. F1 برابر ۰.۹۹۳۵ است که تعادل بالای بین دقت و یادآوری را تأیید می‌کند. بنابراین، XGBoostبهترین انتخاب برای پیش‌بینی‌های دقیق و قابل اعتماد است.

مدل جنگل تصادفی همراه با SMOTE نیز عملکرد بسیار خوبی دارد. دقت برابر ۱ و یادآوری و دقت هر دو ۰.۹۸۹۷ هستند. این مدل به خوبی عدم تعادل داده‌ها را با استفاده از SMOTEجبران کرده و توانسته تقریباً همه نمونه‌های مثبت و منفی را درست تشخیص دهد. F1برابر ۰.۹۸۹۷ نشان می‌دهد که مدل تعادل خوبی بین شناسایی نمونه‌های مثبت و کم کردن خطاهای مثبت دارد.

جنگل تصادفی بدون SMOTE نیز عملکرد قابل قبولی دارد. دقت بسیار بالا (۰.۹۹۹۹) و دقت برابر ۱ نشان می‌دهد که مدل پیش‌بینی‌های مثبت را بدون خطا انجام داده، اما یادآوری کمی کمتر و برابر ۰.۹۸۷۱ است که به معنای از دست دادن تعداد کمی از نمونه‌های مثبت است. F1 برابر ۰.۹۹۳۵ است و نشان می‌دهد تعادل خوبی بین دقت و یادآوری وجود دارد.

مدل رگرسیون لجستیک با ضریب جریمه L1 عملکرد متوسط‌تری نسبت به مدل‌های درختی دارد. دقت برابر ۰.۹۹۸۳ است که هنوز بسیار بالا است، اما یادآوری به ۰.۸۹۹۲ کاهش یافته است، یعنی برخی نمونه‌های مثبت شناسایی نشده‌اند. دقت برابر ۰.۹۴۳۱ و F1 برابر ۰.۹۲۰۶ است، که نشان می‌دهد تعادل بین یادآوری و دقت نسبت به مدل‌های درختی کمتر است.

مدل رگرسیون لجستیک با ضریب جریمه L2 ضعیف‌ترین عملکرد را در میان مدل‌های جدول دارد. دقت برابر ۰.۹۹۶۶ و یادآوری برابر ۰.۸۵۰۱ است که کاهش محسوس در شناسایی نمونه‌های مثبت را نشان می‌دهد. دقت برابر ۰.۹۳۷۳ و F1 برابر ۰.۸۹۱۶ است، که نشان می‌دهد این مدل در مقایسه با مدل‌های درختی و حتی لجستیک L1 کمتر توانمند است.

مدل‌های درختی (XGBoostو جنگل تصادفی) بهترین عملکرد را دارند و تقریبا تمام نمونه‌ها را درست پیش‌بینی می‌کنند. رگرسیون لجستیک هرچند هنوز قابل قبول است، در شناسایی نمونه‌های مثبت و حفظ تعادل معیارها ضعیف‌تر است، به ویژه وقتی از ضریب L2 استفاده می‌شود.

جدول٤. مقایسه مدل‌های یادگیری ماشین

مدل	دقت	یادآوری	صحت	F1
XGBoost	1	0.9897	0.9974	0.9935
RF + SMOTE	1	0.9897	0.9897	0.9897
Random Forest	0.9999	0.9871	1	0.9935
Logistic Regression (L1)	0.9983	0.8992	0.9431	0.9206
Logistic Regression (L2)	0.9966	0.8501	0.9373	0.8916

ماخذ: یافته‌های پژوهش

شکل ۶ به مقایسه عملکرد مدل‌های مختلف یادگیری ماشین می‌پردازد و معیارهای کلیدی مانند دقت، یادآوری، F1 و احتمالاً AUC را نمایش می‌دهد. از روی شکل مشخص است که مدل‌های درختی مانند XGBoost و جنگل تصادفی عملکرد بسیار بالایی دارند و تقریباً بهینه عمل می‌کنند، زیرا مقادیر دقت، یادآوری و F1 نزدیک به ۱ هستند. این نشان می‌دهد که این مدل‌ها توانایی بسیار خوبی در تشخیص نمونه‌های مثبت و منفی دارند و خطاهای پیش‌بینی آنها بسیار کم است.

مقایسه مدل‌ها با توجه به معیارها:

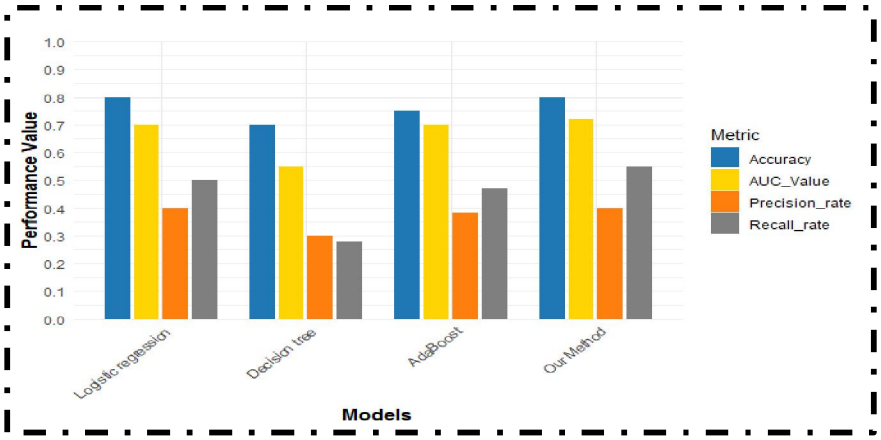
XGBoost:بهترین عملکرد کلی را دارد و تقریباً همه معیارها در حد بسیار عالی قرار دارند، به ویژه F1 که تعادل بین دقت و یادآوری را نشان می‌دهد.

جنگل تصادفی ساده و :RF + SMOTE این مدل‌ها نیز عملکرد نزدیک به XGBoost دارند، با کمی تفاوت در تعادل بین دقت و یادآوری. RF + SMOTE به دلیل استفاده از تکنیک SMOTE برای تعدیل عدم تعادل داده‌ها، در شناسایی نمونه‌های مثبت عملکرد بهتری نسبت به جنگل تصادفی ساده دارد.

رگرسیون لجستیک (L1 وL2) این مدل‌ها عملکرد پایین‌تری نسبت به مدل‌های درختی دارند. به ویژه رگرسیون لجستیک باL2 ، کاهش قابل توجهی در یادآوری و F1 نشان می‌دهد، که بیانگر ضعف مدل در شناسایی نمونه‌های مثبت است.

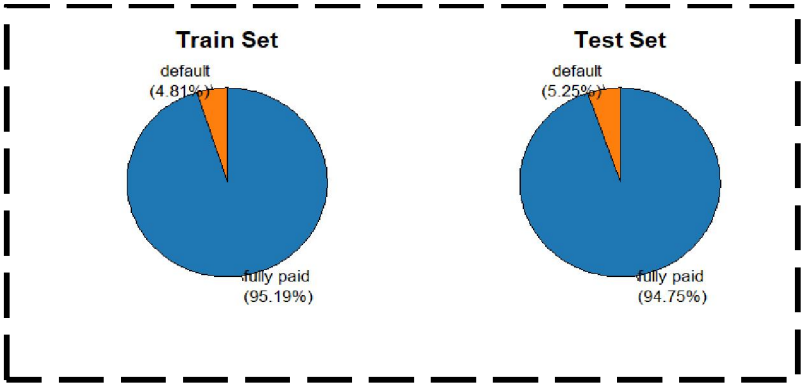
شکل ۶ نشان می‌دهد که برای این مجموعه داده، مدل‌های درختی به ویژه XGBoost بهترین گزینه هستند و مدل‌های رگرسیون لجستیک، هرچند هنوز قابل قبول هستند، در شناسایی نمونه‌های مثبت و حفظ تعادل معیارها کمتر موفق بوده‌اند. همچنین می‌توان نتیجه گرفت که استفاده از تکنیک‌هایی مانند SMOTE در مدل‌های درختی باعث بهبود یادآوری می‌شود، اما هنوز XGBoost بدون نیاز به تغییرات اضافی عملکرد بهینه دارد.

شکل ٦. مقایسه عملکرد مدل‌های مختلف



ماخذ: یافته‌های پژوهش

شکل ۷. نسبت پیش‌بینی صحیح مدل با ضریب جریمه L1 همراه با وزن‌دهی



ماخذ: یافته‌های پژوهش

شکل ۷ نتایج عملکرد مدل پیشنهادی مبتنی بر رگرسیون لجستیک با ضریب جریمه L1 و وزن‌دهی را در دو مجموعه آموزش و آزمون نشان می‌دهد. همان‌طور که مشاهده می‌شود، مدل توانسته است ۹۵.۱۹٪ از نمونه‌های مجموعه آموزش و ۹۴.۷۵٪ از نمونه‌های مجموعه آزمون را به‌درستی طبقه‌بندی کند. این نتایج نشان‌دهنده دقت بالا و تعمیم‌پذیری مناسب مدل است، چرا که اختلاف بسیار کم بین نرخ موفقیت در دو مجموعه (کمتر از ۰.۵٪)، بیانگر عدم بیش‌برازش است. موفقیت این مدل در شرایطی حاصل شده که داده‌ها به‌شدت نامتوازن هستند (تنها حدود ۵٪ مشتریان معوق)، که نشان می‌دهد استراتژی وزن‌دهی به‌خوبی توانسته

است چالش نامتوازنی را مدیریت کند و عملکرد مدل را در شناسایی هر دو کلاس (به‌ویژه کلاس اقلیت) بهبود بخشد.

بحث و نتیجه‌گیری

ریسک اعتباری، از تلفیق روش‌های یادگیری ماشین، پردازش داده‌های کلان و تکنیک‌های مدیریت داده‌های نامتوازن استفاده کرده است. داده‌های مورد استفاده مربوط به مشتریان حقیقی بانک رسالت در سال‌های ۱۴۰۱ و ۱۴۰۲ بوده و شامل دو دسته اصلی اطلاعات شامل اطلاعات ساختاری و جمعیت‌شناختی (مانند سن، جنسیت، سطح تحصیلات، مبلغ تسهیلات، تعداد اقساط، مانده بدهی و شاخص‌های پرداخت)، و متغیر وابسته (هدف) به‌صورت دودویی، مشتریان خوش‌حساب و مشتریان نکول‌کننده (دارای وضعیت معوق، مشکوک‌الوصول یا سررسید گذشته) تعریف شد. داده‌ها دارای عدم توازن شدید بودند؛ تنها حدود ۵٪ از مشتریان در گروه نکول قرار داشتند، در حالی که بیش از ۹۵٪ خوش‌حساب بودند. علاوه بر این، متغیرهای مالی (مانند مبلغ تسهیلات، تأخیر و مانده بدهی) دارای توزیع چوله به راست و پراکندگی بالا بودند که نیازمند پیش‌پردازش دقیق و انتخاب هوشمند الگوریتم بود. یافته‌های تجربی نشان می‌دهد که مدل‌های مبتنی بر درخت تصمیم، به‌ویژه XGBoost و جنگل تصادفی، عملکرد بهتری نسبت به مدل‌های رگرسیون لجستیک (حتی با جریمه‌های L۱ و L۲) دارند. این برتری به‌ویژه در شاخص‌های یادآوری و F1-Score مشهود است که برای تشخیص صحیح مشتریان پرریسک (طبقه اقلیت) حیاتی هستند. همچنین مشخص شد که اعمال وزن‌دهی به کلاس‌ها در مدل‌های رگرسیون لجستیک، به‌جای بهبود عملکرد، در برخی موارد (به‌ویژه با جریمه L۲) منجر به کاهش دقت و یادآوری شده است. در مقابل، استفاده از تکنیک SMOTE در کنار مدل‌های درختی، توانسته است عدم تعادل داده‌ها را به‌خوبی جبران کرده و عملکرد مدل را در شناسایی نکول‌کنندگان بهبود بخشد.

در مجموع، این پژوهش تأکید می‌کند انتخاب الگوریتم مناسب و توجه به ویژگی‌های داده‌های بانکی (مانند نامتوازن بودن و پراکندگی شدید) نقش کلیدی در موفقیت سیستم‌های هوشمند مدیریت ریسک دارد.

یافته‌های پژوهش حاضر با روند تحول مطالعات بین‌المللی پیش‌بینی نکول همسو است. پژوهش‌های کلاسیک بیور (Beaver, 1966) و آلتمن (Altman, 1968) بر نقش نسبت‌های

مالی و شاخص‌های حسابداری در تشخیص درماندگی مالی تأکید داشتند. نتایج این پژوهش نیز اهمیت متغیرهایی مانند مبلغ تسهیلات، مانده بدهی، تعداد اقساط و سابقه تأخیر را تأیید می‌کند؛ بااین‌حال، نشان می‌دهد که ترکیب این متغیرها با ویژگی‌های جمعیت‌شناختی و داده‌های رفتاری دیجیتال، قدرت پیش‌بینی نکول را افزایش می‌دهد. همچنین، برتری XGBoost و جنگل تصادفی بر مدل‌های خطی با نتایج دسای و همکاران (Desai et al., 1996) و لسمن و همکاران (Lessmann et al., 2015) سازگار است و توان مدل‌های غیرخطی را در شناسایی تعاملات پیچیده تأیید می‌کند. با وجود این، رگرسیون لجستیک به دلیل تفسیرپذیری و امکان برآورد احتمال نکول همچنان اهمیت دارد و می‌تواند در کنار مدل‌های درختی استفاده شود. یافته‌ها همچنین با مطالعات جدید درباره ارزش داده‌های تراکنشی و رفتاری همخوان است (Nahar et al., 2024)، هرچند استفاده از این داده‌ها مستلزم پیش‌پردازش معتبر، رعایت حریم خصوصی و کنترل سوگیری است.

در مطالعات داخلی نیز نتایج با یافته‌های اصغری و احمدی (۱۳۹۷) و جندقی و باقری (۱۳۹۹) درباره برتری روش‌های هوشمند و غیرخطی در پیش‌بینی ریسک اعتباری هم‌راستا است؛ با این تفاوت که پژوهش حاضر بر مدل‌های درختی تجمیعی، داده‌های رفتاری و مسئله عدم توازن شدید طبقات تمرکز دارد. یافته‌ها همچنین اهمیت متغیرهای مالی و سوابق بازپرداخت گزارش‌شده در پژوهش فلاح شمس و مهدوی‌راد (۱۳۹۱) را تأیید می‌کند، اما نشان می‌دهد که در داده‌های نامتوازن و چندبعدی، برتری لاجیت بر پروبیت الزاماً به معنای برتری آن بر الگوریتم‌های یادگیری ماشین نیست.

مطالعه بهزادی‌راد و همکاران (۱۴۰۳) بیشتر بر تبیین آماری اثر متغیرهایی مانند سن، جنسیت، مبلغ تسهیلات و نوع وثیقه تمرکز دارد؛ درحالی‌که پژوهش حاضر به توان پیش‌بینی ترکیبی متغیرها و شناسایی مشتریان پرریسک توجه می‌کند. رویکرد فازی محرابی و همکاران (۱۴۰۲) نیز از نظر مدیریت عدم‌قطعیت و تفسیرپذیری مکمل رویکرد داده‌محور پژوهش حاضر است. افزون بر این، نتایج با پژوهش خرمی و همکاران (۱۳۹۹) درباره اهمیت سوابق اعتباری و شاخص‌های رفتاری سازگار است، اما مفهوم اطلاعات رفتاری را به الگوهای دیجیتال مشتریان گسترش می‌دهد.

در مقایسه با شجاع و شوشاد (۱۴۰۰)، پژوهش حاضر ریسک نکول را در سطح خرد و مشتری بررسی می‌کند، درحالی‌که مطالعه آنان بر سرایت ریسک در سطح شبکه متمرکز

است؛ ازاین‌رو، این دو رویکرد سطوح مکمل تحلیل ریسک اعتباری را پوشش می‌دهند. همچنین، مقایسه با مطالعه کمالی و همکاران درباره مدل‌های ZPP و KMV نشان می‌دهد که انتخاب مدل برتر به نوع داده، جامعه آماری، تعریف نکول و سطح تحلیل وابسته است. در مجموع، ارزش‌افزوده پژوهش حاضر در ترکیب داده‌های بانکی و رفتاری، مدیریت عدم توازن طبقات و استفاده از مدل‌های غیرخطی برای پیش‌بینی نکول مشتریان حقیقی بانک رسالت است.

پیشنهادهای کاربردی

۱. استفاده از مدل‌های درختی در سیستم‌های اعتبارسنجی بانکی: بانک‌ها می‌توانند از مدل‌های XGBoost یا جنگل تصادفی به‌عنوان هسته اصلی سیستم‌های تصمیم‌گیری اعتباری خود استفاده کنند، چرا که این مدل‌ها ضمن دقت بالا، قابلیت تفسیرپذیری نسبی و مقاومت در برابر داده‌های نامتوازن را نیز دارند.

۲. استفاده هوشمند از تکنیک‌های متعادل‌سازی داده: در مواجهه با داده‌های بسیار نامتوازن (مانند نکول با فراوانی کم)، به‌جای وزن‌دهی ساده در مدل‌های خطی، از روش‌هایی مانند SMOTE ترکیبی یا روش‌های مبتنی بر یادگیری تجمیعی استفاده شود تا از ایجاد نویز و کاهش دقت جلوگیری گردد.

۲. توسعه سکوی یکپارچه تحلیل ریسک اعتباری: بانک‌ها می‌توانند یک پلتفرم هوشمند یکپارچه ایجاد کنند که بتواند به‌صورت بلادرنگ داده‌های داخلی (مانند سابقه پرداخت) و خارجی (مانند شاخص‌های اقتصادی کلان یا شبکه‌های اجتماعی) را تحلیل کرده و هشدارهای ریسک را به‌موقع صادر کند.

این پیشنهادها نه‌تنها به کاهش ریسک نکول کمک می‌کنند، بلکه می‌توانند به‌عنوان ابزاری برای افزایش شمول مالی و ارائه تسهیلات هوشمند به گروه‌های هدف مؤثر عمل کنند.

تعارض منافع

تعارض منافع وجود ندارد.

سپاسگزاری

بر خود لازم می‌دانم، مراتب تشکر صمیمانه خود را از همکاری سردبیر، اعضای هیئت تحریریه و داوران محترم به عمل آورم.

منابع

اصغری، فرزاد، و احمدی، فرید. (۱۳۹۷). ارائه مدلی ترکیبی برای خوشه‌بندی، رتبه‌بندی و پیش‌بینی عملکرد تسهیلات اعطایی مؤسسات مالی-اعتباری از منظر بازپرداخت بدهی تسهیلات (مورد مطالعه: صندوق کارآفرینی امید). *پژوهشنامه اقتصادی*، ۱۸(۷۱)، ۱۸۵-۲۲۳. doi: 10.22054/joer.2018.9833

باقری، نوشین، و حق‌شناس کاشانی، فریده. (۱۳۹۷). ارزیابی ریسک اعتباری تعاونی‌های شهری با استفاده از روش شبکه عصبی. *فصلنامه علمی-پژوهشی اقتصاد و مدیریت شهری*، ۶(۲۴)، ۱۷-۳۳. URL: http://iueam.ir/article-۱-۹۶۷-fa.html

بهزادی‌راد، مریم، محمودزاده، محمود، حیدری، علی‌عباس، و صوفی‌مجیدپور، مسعود. (۱۴۰۳). اعتبارسنجی مشتریان بانک صادرات: رهیافت امتیازدهی رگرسیون گسسته. *سیاست‌گذاری اقتصادی*، ۱۶(۳۲)، ۱۴۵-۱۷۲. doi: 10.22034/epj.2024.20815.2443

خرمی، امیر، تقوی‌فرد، محمدتقی، و خاتمی فیروزآبادی، سیدمحمدعلی. (۱۳۹۹). ارزیابی ریسک اعتباری متقاضیان تسهیلات بانکی به روش استدلال مبتنی بر مورد (CBR). *مطالعات مدیریت صنعتی*، ۱۸(۵۹)، ۷۹-۱۱۶. doi: 10.22054/jims.2020.40455.2280

شجاع‌وشوشاد، محسن، زمزدیان، غلامرضا، پورزندی، محمدابراهیم، و مینوئی، مهرزاد. (۱۴۰۰). بررسی تأثیر ریسک درماندگی شرکت بر ریسک اعتباری بانک. *دانش حسابداری و حسابرسی مدیریت*، ۱۰(۴۰)، ۳۰۹-۳۲۱.

فلاح شمس، میرفیض، و مهدوی‌راد، حمید. (۱۳۹۱). طراحی مدل اعتبارسنجی و پیش‌بینی ریسک اعتباری مشتریان تسهیلات لیزینگ (مورد مطالعه: شرکت لیزینگ ایران خودرو). *پژوهشنامه اقتصادی*، ۱۲(۴۴)، ۲۱۳-۲۳۴.

محرابی، مسیح، سرورخواه، علی، و عدالت‌پناه، سیداحمد. (۱۴۰۲). تصمیم‌گیری در خصوص اعطای تسهیلات به متقاضیان وام بانک سپه بر اساس عوامل ریسک اعتباری با درنظر گرفتن مجموعه‌های فازی مردد. *مطالعات راهبردی مالی و بانکی*، ۱(۳)، ۱۵۳-۱۶۶.

References

Alamsyah, A., Hafidh, A. A., & Mulya, A. D. (2025). Innovative credit risk assessment: Leveraging social media data for inclusive credit scoring in Indonesia's fintech sector. *Journal of Risk and Financial Management*, *18*(2), 74.

Altman, E. I., & Saunders, A. (1997). Credit risk measurement: Developments over the last 20 years. *Journal of Banking & Finance*, *21*(11–12), 1721–1742.

Asghari,F and Ahmadi,F . (2018). A Hybrid Model for Appraising and Forecasting Loan Repayments (Case Study: Karafarini Omid Fund). *Economics Research*, *18*(71), 185-223. doi: 10.22054/joer.2018.9833[In Persian].

Ayoub, G., Dang, T. H., Oh, T. I., Kim, S.-W., & Woo, E. J. (2020). Feature extraction of upper airway dynamics during sleep apnea using electrical impedance tomography. *Scientific Reports*, *10*(1), Article 1637.

Bagheri, N., & Haghshenas Kashani, F. (2018). Credit risk assessment of urban cooperatives using neural network method. *IUESA*, *6*(24), 17–33. http://iueam.ir/article-1-967-fa.html [In Persian].

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 71–111.

Behzadirad, M., Mahmoudzade, M., Heidari, A., & Sofi Majidpor, M. (2025). Validation of SADERAT Bank customers: Discrete regression scoring approach. *The Journal of Economic Policy*, *16*(32), 145–172. https://doi.org/10.22034/epj.2024.20513.2479 [In Persian].

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033

Cao, M., Qiu, G., & Xing, L. (2021). Research on prediction of big data financial risk behavior based on deep learning. *Journal of Xinyang College of Agriculture and Forestry*, *31*(3), 119–127.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Cheng, L., Yang, Y., & Zhao, K. (2020). Research and improvement of TF-IDF algorithm based on information theory. In *Proceedings of the 8th International Conference on Computer Engineering and Networks (CENet2018)* (pp. 608–616). Kunming, China.

Desai, V. S., Crook, J. N., & Overstreet Jr, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, *95*(1), 24–37.

Dewi, P. M. (2020). Credit insurance as an effort to overcome bad credit risk in modern banking economy in the industrial revolution 4.0 in Indonesia. *UNIFIKASI Jurnal Ilmu Hukum*, *7*(1), 88.

Fallah Shams,M F and Mahdavi Rad,H . (2012). Validating Model and Risk Forecasting for Leasing Customers (Case Study: Iran Khodro Leasing Company). *Economics Research*, *12*(44), 213-234. [In Persian].

Furkatovna, I. K. (2020). Ways of development and formation of Islamic banking in modern economy. *ACADEMICIA: An International Multidisciplinary Research Journal*, *10*(2), 125.

Geeitha, S., & Thangamani, M. (2020). Integrating HSICBFO and FWSMOTE algorithm-prediction through risk factors in cervical cancer. *Journal of Ambient Intelligence and Humanized Computing*, *12*(4), 3213–3225.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Ji, L., Li, J. H., Li, G. Q., Xiao, J. L., & Unrau, S. (2020). Analysis of fractured sections in shale gas wells based on PCA-logistic regression model. *Materials Science Forum*, *980*, 483–492.

Jung, H. K., Yun, S. D., Kim, T. K., & Kim, S. G. (2020). Estimating values of urban green space in Seoul: Application of hierarchical linear regression model on hedonic housing price analysis. *Journal of Environmental Policy and Administration*, *28*(3), 213–242.

Karami, A., & Igbokwe, C. (2025). The impact of big data characteristics on credit risk assessment. *International Journal of Data Science and Analytics*, 1–21.

Kenjaev, M. (2020). The role of the credit risks management in commercial banks. *Bulletin of Science and Practice*, *6*(6), 225–229.

Khorrami, A., Taghavifard, M. T., & Khatami Firouzabadi, S. M. A. (2020). Credit status assessment of bank loan applicants using CBR method. *Industrial Management Studies*, *18*(59), 79–116. https://doi.org/10.22054/jims.2018.18574.1660 [In Persian].

Kim, H., & Lee, T. H. (2021). A robust elastic net via bootstrap method under sampling uncertainty for significance analysis of high-dimensional design problems. *Knowledge-Based Systems*, *225*, Article 107117.

Korkmaz, T., Cetinkaya, A., Aydin, H., & Bariskan, M. A. (2021). Analysis of whether news on the Internet is real or fake by using deep learning methods and the TF-IDF algorithm. *International Advanced Researches and Engineering Journal*, *5*(1), 31–41.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Li, H., & Zhou, L. (2021). Measurement index of systemic financial risk and its economic forecasting ability. *Journal of Econometrics*, *1*(4), 892–903.

Macedo, F. L., Reverter, A., & Legarra, A. (2020). Behavior of the linear regression method to estimate bias and accuracies with correct and incorrect genetic evaluation models. *Journal of Dairy Science*, *103*(1), 529–544.

Mehrabi, M., Sorourkhah, A., & Edalatpanah, S. A. (2023). Decision-making regarding the granting of facilities to Sepah Bank loan applicants based on credit risk factors considering hesitant fuzzy sets. *Financial and Banking Strategic Studies*, *1*(3), 153–166. https://doi.org/10.22105/fbs.2023.181500 [In Persian].

Nahar, J., Rahaman, M. A., Alauddin, M., & Rozony, F. Z. (2024). Big data in credit risk management: A systematic review of transformative practices and future directions. *International Journal of Management Information Systems and Data Science*, *1*(4), 68–79.

Oh, S., & Lee, S. (2020). Extracting insights of classification for Turing pattern with feature engineering. *Journal of the Society for Industrial and Applied Mathematics*, *24*(3), 321–330.

Purwanti, L., Effendi, S. A., Ibrahim, M., Cahyadi, R. T., & Prakoso, A. (2024). Prediction and comparison of bankruptcy models in banking sector companies in Indonesia. *Educational Administration: Theory and Practice*, *30*(4), 4428–4440.

Qiu, S., Gao, Y., Yao, C., Liu, Z., & Li, J. (2021). An enterprise financial risk prediction method for unbalanced samples. *China Science and Technology Resources Guide*, *53*(5), 11–17.

Reychav, I., Zhu, L., Mchaney, R., Zhang, D., Shacham, Y., & Arbel, Y. (2019). Real-time survival prediction in emergency situations with unbalanced cardiac patient data. *Health and Technology*, *9*(3), 277–287.

Samatov, R. (2019). Application of the linear regression method to determine the effective organization of the transportation. *Acta of Turin Polytechnic University in Tashkent*, *9*(3), 4 pages.

Saunders, A., & Allen, L. (2002). *Credit risk measurement: New approaches to value at risk and other paradigms*. John Wiley & Sons.

Shoja, M., Zomordian, G., Pourzarandi, M. E., & Minouie, M. (2021). Investigating the effect of company helplessness risk on bank credit risk. *Journal of Management Accounting and Auditing Knowledge*, *10*(40), 309–321. [In Persian].

Silva, J. C. S., de Lima Silva, D. F., Ferreira, L. L. F., & de Almeida-Filho, A. T. (2021). A dominance-based rough set approach applied to evaluate the credit risk of sovereign bonds. *4OR–Quarterly Journal of Operations Research*, *20*, 139–164.

Thomas, L. C., Crook, J. N., & Edelman, D. B. (2017). *Credit scoring and its applications* (2nd ed.). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611974560

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Werth, J., & Sigman, M. S. (2021). Linear regression model development for analysis of asymmetric copper-bisoxazoline catalysis. *ACS Catalysis*, *11*(7), 3916–3922.

Yu, Y., Xiong, Z., Xiong, Y., & Li, W. (2019). Improved logistic regression algorithm based on kernel density estimation for multi-classification

with non-equilibrium samples. *Computers, Materials and Continua*, *61*(1), 103–118.

Zhang, G., Porikli, F., Sun, H., Sun, Q., Xia, G., & Zheng, Y. (2020). Cost-sensitive joint feature and dictionary learning for face recognition. *Neurocomputing*, *391*, 177–188.

Zhang, H., Shi, Y., Yang, X., & Zhou, R. (2021). A firefly algorithm modified support vector machine for the credit risk assessment of supply chain finance. *Research in International Business and Finance*, *58*, Article 101482.

Zhang, X., & Yu, L. (2024). Consumer credit risk assessment: A review from the state-of-the-art classification algorithm, data traits, and learning methods. *Expert Systems with Applications*, *237*, 121484.

Zhong, L., Yang, J., Xu, D., & Lai, X. (2020). Bandwidth-enhanced oversampling successive approximation readout technique for low-noise power-efficient MEMS capacitive accelerometer. *IEEE Journal of Solid-State Circuits*, *55*(9), 2529–2538.

استناد به این مقاله: نام خانوادگی نویسنده اول، نام. (سال). عنوان مقاله. *عنوان نشریه (ایتالیک)*، سال(شماره)، ص آغاز-ص پایان.