



Presenting an Optimal Model based on Artificial intelligence in reducing the credit risk of the SQEV

Farshad Parhizkar Miandehi

Ph.D. Student in Software Engineer, Azad Islamic University of Rasht
Email: farshad.parhizkar@gmail.com

Asadollah Shahbaharami

Ph.D. in Computer Engineering , Azad Islamic University of Rasht
Email: shahbahrami@guilan.ac.ir

Peyman Bayat

Ph.D. in Computer Engineering , Azad Islamic University of Rasht
Faculty member of Computer Department of Islamic Azad University, Rasht branch
Email: bayat@iaurasht.ac.ir

Imam Khomeini Relief Foundation
Empowerment and
Deprivation-elimination Journal
<https://pajooresh.emdad.ir>
Vol.1 | No.3
March 2024



| Received: 06, Jan. 2024 | Accepted: 04, Mar. 2024 |

Abstract

Empowering needy people through the payment of employment facilities is one of the challenges and strategies of developed countries. Paying facilities to needy people due to their special characteristics with common customers of financial institutions has a high credit risk. Common models of credit risk analysis such as credit score cards, decision trees, and gradient boosting have generally been developed for small samples and little data. They do not provide volume. On the other hand, collecting customer data is a difficult task due to the laws of maintaining the security of information. According to the above problem, in this research, using adversarial generative neural networks, a model has been presented that can perform credit risk analysis with high accuracy even for small samples. In this model, which is based on The theory of games is that new data are generated by the neural network until the Nash equilibrium point is reached, and then the data generated with the same distribution is combined with the real data. The results of the proposed model compared to the developed gradient boosting method show It is that while overcoming the problem of lack of successful samples, the accuracy of customer credit risk analysis has increased by 32.7%.

Keywords: Credit Risk, Data Mining, Generative Adversarial Network (GAN), Empowerment, Classification.



نشریه علمی توانمندسازی و محرومیت زدایی

کمیته امداد امام خمینی (ره)

<https://pajooresh.emdad.ir>

دوره ۱ | شماره ۳ | صص ۴۹-۷۲

زمستان ۱۴۰۲



ارائه یک مدل بهینه مبتنی بر هوش مصنوعی در کاهش ریسک اعتباری تسهیلات صندوق قرض الحسنه امداد ولایت

دانشجوی دکتری سیستم‌های نرم‌افزاری، دانشگاه آزاد اسلامی، واحد رشت.

farshad.parhizkar@gmail.com

دکتری کامپیوتر، دانشگاه آزاد اسلامی رشت.

shahbahrami@guilan.ac.ir

دکتری سیستم‌های نرم‌افزاری، دانشگاه آزاد اسلامی رشت؛

عضو هیئت علمی گروه کامپیوتر دانشگاه آزاد اسلامی، واحد رشت.

bayat@iaurasht.ac.ir

فرشاد پرهیزکار میاندهی

اسدالله شاه بهرامی

پیمان بیات

| دریافت: ۱۴۰۲/۱۰/۱۶ | پذیرش: ۱۴۰۲/۱۲/۱۴ |

چکیده: توانمندسازی افراد نیازمند از طریق پرداخت تسهیلات اشتغال یکی از چالش‌ها و راهبردهای کشورهای توسعه یافته می‌باشد. پرداخت تسهیلات به افراد نیازمند به دلیل ویژگی‌های خاص آنان با مشتریان رایج موسسات مالی دارای ریسک اعتباری بالایی می‌باشد. مدل‌های رایج تحلیل ریسک اعتباری مانند کارت امتیاز اعتباری، درخت تصمیم و گرادیان بوستینگ توسعه یافته عموماً نتیجه مطلوبی را برای نمونه‌های کوچک و داده‌های کم حجم ارائه نمی‌دهند. از سویی جمع‌آوری داده مشتریان به دلیل قوانین حفظ امنیت اطلاعات کاری دشوار می‌باشد. با توجه به مشکل فوق، در این تحقیق با استفاده از شبکه‌های عصبی مولد خصمانه و الگوریتم گرادیان بوستینگ تقویت شده مدلی ارائه گردیده است که بتواند برای نمونه‌های کوچک نیز تحلیل ریسک اعتباری را با دقت بالایی انجام دهد. در این مدل که بر پایه نظریه بازی‌ها می‌باشد داده‌های جدید توسط شبکه عصبی تا رسیدن به نقطه تعادل نش تولید می‌شوند و سپس داده‌های تولید شده با توزیع مشابه، با داده‌های واقعی ترکیب می‌گردند. نتایج حاصل از پیاده‌سازی مدل پیشنهادی در مقایسه با روش گرادیان بوستینگ توسعه یافته بیانگر آن است که ضمن غلبه بر مشکل کمبود نمونه موفق میزان دقت تحلیل ریسک اعتباری مشتریان بانک‌ها ۳۲.۷ درصد افزایش داشته است.

واژه‌های کلیدی: ریسک اعتباری، یادگیری عمیق، شبکه مولد خصمانه، گرادیان بوستینگ توسعه یافته.



مقدمه

در سالیان اخیر تخصیص اعتبار بانکی جهت توانمندسازی نیازمندان، به ویژه در کشورهای در حال توسعه مورد توجه رهبران قرار گرفته است [۱]. در این مسیر شناسایی افراد با توجه به ویژگی‌های فردی آنان و محدودیت منابع اعتباری چالش‌های فراوانی را برای موسسات مالی ایجاد نموده است. زیرا، هسته مدیریت ریسک اعتباری از داده‌های موجود در بانک اطلاعاتی مشتریان می‌باشد که این امر همواره عدم قطعیت در تصمیم‌گیری سرمایه‌گذاران را به همراه داشته است [۲].

در ایران نیز توانمندسازی نیازمندان با رویکرد ایجاد اشتغال پایدار و استفاده از تسهیلات بانکی مورد توجه بوده است که مراحل شناسایی، تخصیص و نظارت بر ایجاد این مشاغل از وظایف کمیته امداد امام خمینی (ره) می‌باشد.

لیکن، با توجه به جامعه هدف یعنی افراد نیازمند جامعه، ریسک اعتباری به شدت افزایش می‌یابد. از آنجا که ویژگی‌های این افراد خاص بوده و ویژگی‌های آنان با بنگاه‌های اقتصادی متفاوت است، ریسک اعتباری نیز به همان میزان افزایش می‌یابد.

حال سوال اساسی این است، چگونه با این ریسک مقابله نماییم؟

ابتدایی‌ترین روش، مدل کارت امتیاز اعتباری مشتریان می‌باشد. در این مدل از روش رگرسیون لجستیک جهت سنجش ریسک اعتباری استفاده می‌نمایند [۳]. تحلیلگران در این روش شاخص‌های متفاوتی را بر اساس متغیرهای موجود مانند: سطح درآمد، توانایی پرداخت تسهیلات و وضعیت اعتبار قبلی استخراج نموده و سپس شاخص‌های بدست آمده را جهت سنجش ریسک اعتباری در سطوح مختلف دسته‌بندی می‌نمایند [۴].

با توسعه هوش مصنوعی، روش‌های مبتنی بر داده کاوی و یادگیری ماشین [۵، ۶]، شبکه‌های عصبی [۷، ۸، ۹]، روش مونت کارلو [۱۰] و روش بوستینگ توسعه یافته [۱۳]، [۱۱، ۱۲] با بروز یادگیری عمیق و الگوریتم گرادیان کاهشی انقلابی در پیش‌بینی و کاهش ریسک اعتباری بانک‌ها ایجاد گردید [۱۴، ۱۵، ۱۶، ۱۷].

بزرگترین مشکل مدل رگرسیون لجستیک این است که هیچ‌گونه توزیع احتمالی را در نظر نمی‌گیرد و زمانی که نقاط داده از هم فاصله داشته باشند عملاً این روش کارا



ارائه یک مدل بهینه مبتنی بر هوش مصنوعی... | پرهیزکار میاندھی و دیگران

نمی‌باشد. درخت‌های تصمیم به ظاهر درک شهودی ساده‌ای دارند و قابل تحلیل می‌باشند [۱۹]. لیکنف زمانیکه با داده‌های حجیم مواجه می‌گردیم و ترکیب انفجاری داده‌ها رخ می‌دهد عملاً این گونه درخت‌ها بسیار پیچیده شده و عمل بیش برآزش رخ می‌دهد. در روش‌های مبتنی بر یادگیری ماشین نیز مشکلات متعددی وجود دارد که باعث می‌گردد تحلیل ریسک اعتباری با خطا همراه گردد، مواردی مانند: عدم ایجاد محدودیت در توزیع داده‌ها، عدم امکان استخراج ویژگی مناسب با توجه به نوع جامعه هدف و عدم وجود داده‌های آموزشی کافی [۲۰].

در حال حاضر روش‌های مبتنی بر بوستینگ توسعه یافته به عنوان کارا ترین روش در دسته‌بندی الگوریتم‌ها شناخته می‌شوند که توانسته‌اند بر بسیاری از مشکلات غلبه نمایند. از سویی ساخت مدل، جهت کاهش ریسک اعتباری علاوه بر تشخیص متغیرهای تاثیر گزار نیازمند تعداد کافی نمونه استاندارد جهت آموزش شبکه می‌باشد. بطور کلی اندازه داده‌ها نباید خیلی کوچک و یا خیلی بزرگ باشد [۲۱]. برابر تحقیقات تجربی، برای اطمینان از پایداری مدل و تخمین مناسب از عملکرد آن لازم است تعداد نمونه‌های موفق و ناموفق (آموزشی) ۲۰ برابر تعداد متغیرهای مستقل باشد [۲۲]. لیکن، در عمل همواره مشکل تعداد نمونه و یا عدم تعادل در تعداد نمونه مورد نیاز به نسبت تعداد متغیرهای مستقل وجود دارد. زیرا، در شروع تخصیص اعتبار با هدف توانمندسازی افراد، عملاً تعداد نمونه‌های آموزشی (SMEs) بسیار ناچیز می‌باشد.

محور اصلی این تحقیق، متقاضیان و خانواده تحت پوشش کمیته امداد است و از آنجایی که تعداد متقاضیان این حوزه در مقایسه با میزان خدمت و منابع قابل ارائه معمولاً بسیار بیشتر است، دقت در توزیع خدمت به متقاضیان با اولویت بالاتر، بسیار حائز اهمیت است. زیرا همواره این احتمال وجود دارد که در روال توزیع اعتبارات، اولویت افراد با توجه به ویژگی‌های آنها مورد توجه قرار نگیرد. این امر به دلایل مختلفی می‌تواند رخ دهد. اولاً، با توجه به تعداد زیاد متقاضیان و بالا بودن حجم داده‌ها و اطلاعات هر متقاضی، این امکان وجود دارد که کارشناس سازمان نتواند همه داده‌ها را در نظر گرفته و پردازش کند. لذا ممکن است در انتخاب متقاضی با اولویت بالا برای ارائه خدمت، سهواً



اشتباهی رخ دهد. دوماً، این امکان وجود دارد که کارشناس، برخی مواقع اولویت یک متقاضی را برای ارائه خدمت، تنها براساس برخی ویژگی‌های فردی در نظر بگیرد و همه ویژگی‌های یک فرد را همزمان مد نظر قرار ندهد.

حال سؤال اساسی که این تحقیق تلاش در پاسخگویی به آن را دارد این است که چگونه می‌توان تعداد کم داده‌های آموزشی (SMEs) را حل نمود.

در این تحقیق، راه حلی ارائه می‌گردد، تا مشکلات مطرحه با استفاده از شبکه‌های مولد خصمانه (GAN) مرتفع گردد و سپس با تلفیق این شبکه و الگوریتم بوسستینگ توسعه یافته مدلی ارائه می‌دهد که مشکل کمبود داده در لحظه مرتفع شده و دسته‌بندی متقاضیان با دقت بالایی صورت پذیرد. قابل توجه است، که این تحقیق بر روی داده‌های حقیقی متقاضیان تسهیلات توانمندسازی کمیته امداد امام خمینی (ره) انجام پذیرفته است.

در ادامه در بخش دوم، شبکه GAN و XGBoost و مدل‌های ارزیابی روش‌های پیش بینی شامل معیار KS و AUC معرفی می‌گردند. بخش سوم تشریح راهکار پیشنهادی و متغیرهای تحقیق می‌باشد و بخش چهارم که شامل بحث و ارزیابی روش ارائه شده و نتایج خروجی است و در نهایت بخش پنجم نتیجه‌گیری و راهکارهای آینده را بیان می‌دارد.

پیشینه پژوهش

تکنیک‌های داده‌کاوی و یادگیری ماشین و بخصوص روش‌های طبقه‌بندی و به طور وسیع در حوزه‌های مختلف برای پیش‌بینی رفتارهای متفاوت استفاده شده است. از جمله حوزه‌های کاربردی تکنیک‌های داده‌کاوی می‌توان به ارزیابی ریسک اعتباری مشتریان بانکها اشاره نمود. بی‌شک، یکی از مهم‌ترین ریسک‌های موجود که می‌تواند باعث از بین رفتن منابع اعتباری سازمانها گردد ریسک پرداخت تسهیلات می‌باشد. در سال ۱۹۶۶، روش رگرسیون لجستیک توسط بی‌ور برای نخستین بار تعیین ورشکستگی شرکتها به کار گرفته شد و بعدها به منظور ریسک اعتباری اوراق قرضه مورد استفاده قرار گرفت [۹]. یکی از اصلی‌ترین متدهای اندازه‌گیری ریسک اعتباری توسط آلتمن در سال ۱۹۶۸ ارائه گردید و به مدل نمره Z آلتمن شهره شد، که در آن یک تحلیل ممیزی با استفاده از



ارائه یک مدل بهینه مبتنی بر هوش مصنوعی... | پرهیزکار میانه‌ی و دیگران

نسبت‌های مالی جهت متمایز نمودن شرکت‌های ورشکسته و موفق صورت می‌پذیرد. بعدها ساندرز و آلن از این متد برای تفکیک مشتریان اعتباری موسسات مالی و بانکها استفاده گردیده است. با گسترش هوش مصنوعی، داده‌کاوی و یادگیری ماشین، استفاده از روش‌های جدید طبقه‌بندی جهت تفکیک مشتریان مورد توجه قرار گرفت و تا کنون تحقیقات بسیاری در این حوزه صورت پذیرفته است [۲۳].

شبکه‌های مولد خصمانه^۱ (GAN)

GAN در واقع یک الگوی یادگیری عمیق^۲ است، که برای اولین بار توسط گودفلو^۳ و همکارانش ارائه شد [۲۴]. فرآیند بهینه‌سازی GAN در قالب تئوری بازی مابین دو بازیکن در حالت مینیماکس^۴ است برای دستیابی به تعادل نش^۵ است [۲۵]. طی سال‌های اخیر، GAN به عنوان الگوریتم کارا در دامنه وسیعی از پژوهش‌های علمی از جمله بینایی کامپیوتر^۶ مورد استفاده قرار گرفته است [۲۶، ۲۷]. با این حال، برخی مطالعات از GAN برای غلبه بر داده‌های آموزشی کم و تقویت داده^۷ استفاده کردند و مشکل ناکافی بودن داده را در پروژه‌هایی مانند پیش‌بینی مسیر، برق، تصاویر پزشکی و سایر صنایع را بطور کارآمد حل کردند [۲۹، ۳۰، ۳۱]. ساختار کلی شبکه GAN در شکل (۱) نشان داده شده است که شامل دو قسمت است: قسمت مولد^۸ (G) و متمایز کننده^۹ (D). مولد G برای به دست آوردن توزیع داده، و متمایز کننده D برای تخمین احتمال داده از نمونه‌های آموزشی استفاده می‌شود. G برای به حداکثر رساندن احتمال^{۱۰} خطای D آموزش دیده است. این مدل شبیه‌سازی رقابت در بین دوبازیکن می‌باشد، و تنها یک راه‌حل منحصر به فرد در فضای عملیاتی جواب^{۱۱} G و D وجود دارد. با توزیع داده‌های آموزشی به دست آمده

1. generative adversarial network
2. deep learning
3. Goodfellow
4. minimax two-player game
5. Nash equilibrium
6. computer vision
7. data amplification
8. Generative
9. discriminator
10. chance
11. functional solution space



توسط G ، احتمال خروجی D به تدریج به $1/2$ نزدیک می‌شود. هنگامی که D و G به عنوان پرسپترون‌های^۱ چندلایه تعریف شوند، سیستم با پس‌انتشار^۲ خطا آموزش می‌بیند. در روند این آموزش و تولید نمونه، به زنجیره مارکوف^۳ یا شبکه استنتاج تقریبی^۴ نیازی نیست.

مولد G : یک شبکه مولد است، که با دریافت نویز از پیش تعریف شده Z ، شبه داده $G(z)$ را تولید می‌کند.

تمایزکننده D : یک شبکه جداکننده (تمایزکننده) است، که با دریافت داده‌های ورودی نمونه x ، احتمال $D(x)$ اینکه داده‌های خروجی داده‌های درست هستند یا خیر، را تعیین می‌کند.

هدف تمایزکننده D این است که داده‌های درست و نادرست را تشخیص دهد. هرچه خروجی تابع $D(x)$ داده‌های نمونه اصلی x به 1 نزدیک‌تر باشد، بهتر است، و هرچه به 0 تابع $D(G(z))$ نزدیک‌تر باشد، بهتر است. هدف مولد G این است که از شبه داده $G(z)$ تولیدشده توسط مدل G برای فریب دادن تمایزکننده D استفاده کند، یعنی هرچه $D(G(z))$ بزرگتر کند بهتر است [۳۱]. بنابراین، آموزش D و G به مسئله مینیماکس مبدل می‌شود:

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

که $p_{data}(x)$ توزیع داده‌های حقیقی، و $p_z(z)$ توزیع شبه داده است. تابع هزینه مربوط به مولد G و تمایزکننده D عبارتند از:

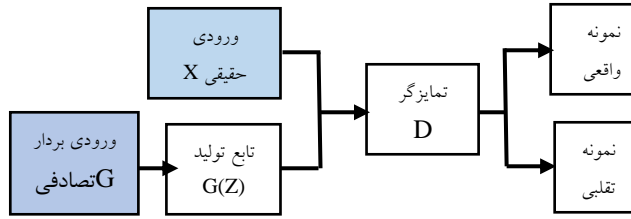
$$\min_G V(D, G) = -E_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

$$\max_D V(D, G) = -E_{x \sim p_{data}(x)} [\log D(x)] - E_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

-
1. perceptron
 2. back propagation
 3. Markov chain
 4. approximate inference network



ارائه یک مدل بهینه مبتنی بر هوش مصنوعی... | پرهیزکار میاندهی و دیگران



شکل ۱. ساختار استاندارد شبکه GAN

الگوریتم گرادینان تقویتی^۱ (XGBoost)

XGBoost الگوریتم گرادینان تقویت شده^۲، توسعه‌ای از درخت تصمیم تقویت گرادینان^۳ استاندارد (GBDT) است (فریدمن ۲۰۰۱)^۴، که در سال ۲۰۱۶ توسط چن و گاسترین^۵ پیشنهاد شد. ایده اصلی XGBoost این است که برای رشد درختان، ویژگی‌ها تقسیم شوند. برای برآزش با باقیمانده پیش‌بینی شده در آخرین تکرار، یک درخت به تکرار محاسباتی اضافه می‌شود تا تابع جدید را آموزش ببیند.

برای یک مجموعه داده با n نمونه و m ویژگی، مدل جمعی درخت^۶ از K تابع جمع‌پذیر^۷ برای پیش‌بینی خروجی استفاده می‌کند.

$$\hat{y}_i = \phi(X_i) = \sum_{k=1}^K f_k(X_i), f_k \in F$$

که $F = \{f(x) = \omega_{q(x)}\} (q: R^m \rightarrow T, \omega \in R^T)$ فضای درختان رگرسیون، q ساختار هر درخت است که یک نمونه را به شاخص برگ^۸ مربوطه نگاشت می‌کند، T تعداد برگ‌های درخت، و f_k مربوط به ساختار درختی مستقل q و وزن‌های برگ ω است. برخلاف درختان تصمیم، هر درخت رگرسیون شامل یک نمره^۹ پیوسته در هر برگ است، ω_i نمره در برگ i ام است. تابع هدف آن عبارت است از:

1. Extreme Gradient Boosting
2. Tree Boosting algorithm
3. Gradient Boosting Decision Tree
4. (Friedman 2001)
5. Chen and Guestrin
6. tree ensemble model
7. additive function
8. leaf
9. score



$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

که $\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2$ است، و L تابع زیان محدب^۱ مشتق پذیر است که نشان دهنده خطای بین نتیجه پیش بینی شده \hat{y}_i و نتیجه واقعی y_i است. در عین حال، برای جلوگیری از برآزش بیش از حد مدل، عبارت منظم ساز^۲ Ω اضافه می شود.

در مقایسه با XGBoost، GBDT می تواند به طور خودکار از قابلیت چندرشته ای^۳ CPU برای موازی سازی استفاده کند و دقت الگوریتم را بهبود بخشد. علاوه بر این، XGBoost از توابع هزینه سفارشی پشتیبانی می کند، زیرا تابع می تواند مشتق مرتبه اول و مرتبه دوم باشد. XGBoost برای کنترل پیچیدگی مدل، عبارات منظم ساز را به تابع هزینه اضافه می کند. عبارت منظم ساز شامل تعدادی گره برگ درخت و مجموع مربعات ماژول های L2 نسبت های^۴ خروجی هر گره برگ است. عبارت منظم ساز واریانس مدل را کاهش می دهد، مدل آموخته شده را ساده تر می کند و از برآزش بیش از حد جلوگیری می کند.

شاخص های ارزیابی تحقیق

رایج ترین شاخص های ارزیابی ریسک مالی اعتباری ماتریس درهم ریختگی، مقدار کولموگروف-اسمیرنوف (KS)، مقدار ناحیه زیر منحنی (AUC) است. در این مقاله KS و AUC برای ارزیابی مدل انتخاب شده اند.

همانطور که در جدول (۱) مشاهده کردید، TP نمونه هایی که به درستی به عنوان مشتریان خوب ارزیابی می شوند را نشان می دهد، TP+FN نشان دهنده همه نمونه های مشتریان مطلوب است. FP نمونه هایی را نشان می دهد که به اشتباه به عنوان مشتری خوب ارزیابی شده و FP + TN نشان دهنده همه نمونه هایی که در واقع مشتریان بد هستند. $TPR = TP / (TP + FN)$ ، نرخ ضربه یا حساسیت نامیده می شود، که به معنای درصد تمام نمونه هایی است که در واقع مشتریان خوبی هستند و به درستی به عنوان مشتریان خوب

1. convex loss function

2. regular term

3. multithreading

4. fraction

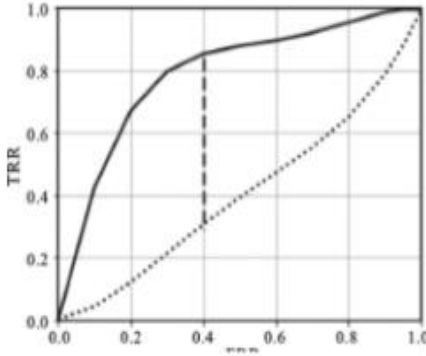


ارائه یک مدل بهینه مبتنی بر هوش مصنوعی... | پرهیزکار میانه‌ی و دیگران

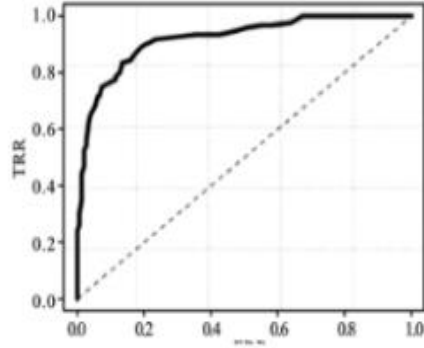
ارزیابی می‌شوند. $FPR = FP/(FP+TN)$ ، نرخ مثبت کاذب نامیده می‌شود، که به معنای درصدی از تمام نمونه‌هایی است که در واقع مشتریان بدی هستند که به اشتباه به عنوان مشتریان خوب قضاوت می‌شوند.

جدول ۱. ماتریس درهم ریختگی

مجموع	خروجی منفی تست	خروجی مثبت تست	
TP+FN	منفی کاذب FN	مثبت واقعی TP	شرایط مثبت
FP+TN	منفی واقعی TN	مثبت کاذب FP	شرایط منفی
TP+FP+FN+TN	FN+TN	TP+FP	مجموع



(a)



(b)

شکل ۲. ارزیابی مدل: نمودار KS و نمودار ROC

منحنی KS در شکل ۲ (a) نشان داده شده است. جمعیت را به n (معمولاً ۱۰) قسمت مساوی تقسیم کرده و با توجه به احتمال پیش فرض به ترتیب نزولی مرتب می‌کند، توزیع تجمعی پیش فرض و درصد نرمال را در هر قسمت مساوی محاسبه می‌کنیم، تفاوت بین آنها را رسم می‌کنیم و سپس منحنی KS را بدست می‌آوریم. حداکثر مقدار در منحنی KS، مقدار KS نامیده شده و بین ۰ و ۱ است. هر چه مقدار KS بالاتر باشد، توانایی رتبه‌بندی و پیش‌بینی مدل، قوی‌تر است. هنگام نمونه‌گیری تصادفی، مقدار KS صفر و در طبقه‌بندی بهینه، مقدار آن یک است.

منحنی عملیاتی گیرنده (ROC) بر اساس ماتریس درهم ریختگی است. همانطور که در



شکل ۲ (b) نشان داده شده است، FPR به عنوان محور X و TPR به عنوان محور Y تعریف می شود. با توجه به حد آستانه، یک نقطه مختصات ($y = TPR, x = FPR$) را می توان از مقادیر واقعی و پیش بینی شده نمونه ها محاسبه کرد. منحنی ROC با تغییر آستانه از ۰ به حداکثر مجموعه نقاط مختصات رسم می شود. ناحیه زیر منحنی ROC مقدار AUC نامیده شده که نشانگر مدل نیز می باشد و همچنین نشان دهنده احتمال رتبه بندی مورد مثبت پیش بینی شده از حالت منفی است. هر چه مقدار AUC بیشتر باشد، اثر پیش بینی مدل بهتر است.

روش پژوهش و مدل پیشنهادی

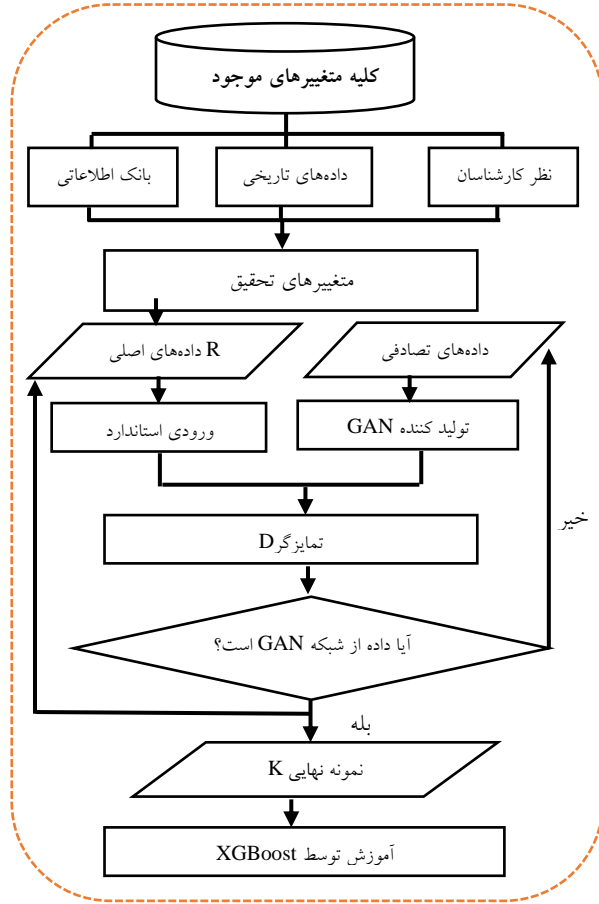
راه حل پیشنهادی برای تخصیص اعتبارات بر مبنای ویژگی متقاضیان در شکل (۱) نشان داده شده است. در این تحقیق از شبکه GAN جهت غلبه بر کمبود نمونه های موفق و از شبکه XGBoost جهت پیش بینی و دسته بندی مشتریان استفاده گردیده است که برای درک بهتر مدل آن در شکل (۲) نمایش داده شده است.

شکل ۱. روند اجرای الگوریتم طبقه بندی بر روی داده های موجود در بانک اطلاعاتی کمیته امداد





ارائه یک مدل بهینه مبتنی بر هوش مصنوعی... | پرهیزکار میاندهی و دیگران
در شکل (۲) معماری مورد نظر تحقیق جهت طبقه بندی مشتریان و کاهش ریسک اعتباری ارائه گردیده است.



شکل ۲. مدل پیشنهادی طبقه‌بندی مشتریان تسهیلات حمایت

در این مرحله پژوهش، تعداد کل نمونه‌های مورد استفاده در تحقیق ۳۴۹۴۵۰ نمونه که به سه گروه کشاورزی و دامپروری به تعداد ۸۸۷۵۰ نمونه، تولیدی ۵۵۲۰۰ نمونه، اصناف و خدمات ۲۰۵۵۰۰ نمونه تخصیص یافته است که از این تعداد ۲۳۸۷ نمونه موفق و نا موفق برچسب گذاری شده موجود است. داده‌های انتخاب شده براساس بانک اطلاعاتی موجود در کمیته امداد استخراج گردید.



از آنجایی که عوامل متعددی بر ریسک اعتباری تأثیر می‌گذارد و ابعاد عوامل مختلف به‌طور قابل ملاحظه‌ای متفاوت است، استفاده مستقیم از داده‌ها دشوار است. بنابراین، در آزمایش، داده‌های اصلی در محدوده $[-1, 1]$ نرمال می‌شوند. فرمول تبدیل بین هر گروه از عوامل تأثیرگذار در زیر آمده است.

$$x_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \times 2 - 1$$

GAN توسط داده‌های عادی، نویز تصادفی و داده‌های نمونه واقعی آموزش داده می‌شود، که در آن بازی بین مولد G و متمایزکننده D تا زمانی که به تعادل نش برسند، ادامه می‌یابد. در نهایت، شبکه‌های قدرتمند G و D که می‌توانند داده‌های اصلی را شبیه‌سازی کنند، به دست می‌آیند.

مولد G شبه داده $G(z)$ را با دریافت نویز تصادفی z تولید می‌کند. با فرض اینکه توزیع نمونه‌های آموزشی $p(x)$ مشخص باشد، می‌توان نمونه‌های جدید را به‌طور تصادفی نمونه برداری کرد. مولد برای بدست آوردن رابطه بین نویز z و نمونه آموزشی x در حال کار است. در آزمایش، نویز تصادفی از توزیع گاوسی $N(0, 1)$ پیروی می‌کند.

متمایزکننده D با دریافت ورودی x ، احتمال $D(x)$ درست یا $G(z)$ بودن داده را تعیین می‌کند. از طریق بازی و تمرین متقابل G و D، عملکرد netها بهبود یافته و در نهایت به حالت تعادل نش می‌رسد.

در GAN، می‌توانیم تفکیک‌کننده را با افزایش و مولد را با کاهش گرادیان تصادفی به‌روزرسانی کنیم. حداکثر تعداد تکرار را با T نشان می‌دهند. تعداد مراحل آموزش k یک هایپر پارامتر بوده، $k = \text{Round}(N/\text{batch})$ ، که در آن N تابع گرد کردن، N اندازه نمونه و batch تعداد نمونه‌هایی است که هر بار گرفته می‌شود.

دسته‌ای از m نمونه $\{z_1, z_2, \dots, z_m\}$ را از شبه داده‌ها $P_z(z)$ و دسته‌ای از m نمونه $\{x_1, x_2, \dots, x_m\}$ را از داده‌های واقعی $P_{\text{data}}(x)$ انتخاب کرده، سپس توابع cost آنها با استفاده از معادله ۳ و به‌روزرسانی تمایزگر با افزایش گرادیان تصادفی آن محاسبه شود.

$$\nabla_{\theta_i} \frac{1}{m} \sum_{i=1}^m [\log D(x_i) + \log(1 - D(G(z_i)))]$$



ارائه یک مدل بهینه مبتنی بر هوش مصنوعی... | پرهیزکار میاندهی و دیگران

پس از اینکه تمایزگر با k بار آموزش به روز شد، دسته‌ای از m نمونه $\{z_1, z_2, \dots, z_m\}$ را از $P_z(z)$ انتخاب کنید. سپس توابع هزینه آنها با استفاده از معادله ۲ برای به روزرسانی مولد با کاهش گرادیان تصادفی آن محاسبه شود:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^m \log(1 - D(G(z_i)))$$

فرض کنیم که مجموعه‌ای از داده‌های بردار داشته و $P_{\text{data}}(x)$ توزیع داده‌های واقعی باشد. در طول فرآیند آموزش، $P_z(z)$ به تدریج به $P_{\text{data}}(x)$ همگرا خواهد شد. گزاره به صورت زیر نشان داده شده است:

گزاره ۱: برای G ثابت، تفکیک‌کننده بهینه D عبارت است از:

$$D^*(x) = \frac{P_{\text{data}}(x)}{P_{\text{data}}(x) + p_z(x)}$$

گزاره ۲: اگر G و D ظرفیت کافی داشته باشند و در هر مرحله از الگوریتم، تمایزکننده

D مجاز باشد به G بهینه خود برسد، و $P_z(x)$ به روزرسانی شود، معیار بهبود می‌یابد:

$$E_{x \sim P_{\text{data}}(x)} [\log D_G^*(x)] + E_{z \sim P_z(x)} [\log(1 - D_G^*(x))]$$

اگر فرمول به مقدار بهینه برسد، آنگاه $D^*(x) = 1/2$ و $P_z(x) = P_{\text{data}}(x)$ خواهد بود.

با توجه به گزاره ۱ و ۲، پس از رسیدن کل مدل به تعادل نش، مولد توزیع، تخمینی از

توزیع داده اصلی $P_z(z) = P_{\text{data}}(x)$ به دست می‌آورد.

پس از آموزش شبکه، شبه داده F توسط مولد آموزش دیده، نمونه تقویت شده K که

از داده‌های نمونه اصلی R و داده‌های شبه F تشکیل شده، تولید می‌شوند. در مقایسه با

داده‌های اصلی R ، توزیع نمونه‌های تقویت شده K اساساً یکسان است، در حالی که مقدار

داده‌ها به طور قابل توجهی بهبود یافته است. حالا، $XGBoost$ را می‌توان با نمونه تقویت

شده K برای پیش‌بینی ریسک آموزش داد.



متغیرهای تحقیق

اطلاعات اولیه درخواست کننده	نوع تخصص، سابقه فعالیت و اشتغال، وضعیت فعالیت مورد نظر در آینده، سرمایه فعلی و غیره
اطلاعات قضایی درخواست کننده	دعوی قضایی و سوابق بد، مانند نقض قوانین، اخطار دادگاه، اسناد قضایی و سوء سابقه فردی
اطلاعات اشخاص مرتبط با کسب و کار	تعداد اشخاص مرتبط به صورت مستقیم، نوع فعالیت اشخاص مرتبط و تعداد افراد که در آینده می‌توانند مشغول به کار شوند
اطلاعات کسب و کار	اطلاعات مالی، مانند، حاشیه سود، نحوه بازاریابی، گردش حساب‌های دریافتی، بازده مورد نظر و غیره
اطلاعات اعتباری درخواست کننده	سابقه اعتباری، بدهی‌های جاری، بدهی‌های احتمالی، تقاضای اعتبار، وضعیت ملک و معوقه و غیره
اطلاعات اولیه شخص درخواست کننده	سن، تحصیلات، جنسیت، وضعیت تأهل، سواد، وضعیت جسمانی، و غیره
اطلاعات دارایی درخواست کننده	سپرده‌ها و سایر منابع مالی و غیره
اطلاعات بدهی درخواست کننده	نوع وام، تاریخ وام، مبلغ وام و غیره

تجزیه و تحلیل

مجموعه‌ای از ۲۳۸۷ نمونه مشتری ابرچسب‌گذاری شده شامل ۱۲۸ متغیر که از این تعداد ۴۲۱ نمونه شکست خورده ($y = 1$) و ۱۹۶۶ درخواست کننده موفق ($y = 0$) هستند، جمع آوری شده است. در آزمایش، ۵۰۰ نمونه به‌طور تصادفی به‌عنوان مجموعه تأیید انتخاب شدند که نسبت مشتریان خوب و بد در مجموعه تأیید با داده‌های اصلی یکسان است. سپس از ۱۸۸۷ نمونه باقی مانده به‌عنوان مجموعه آموزشی برای



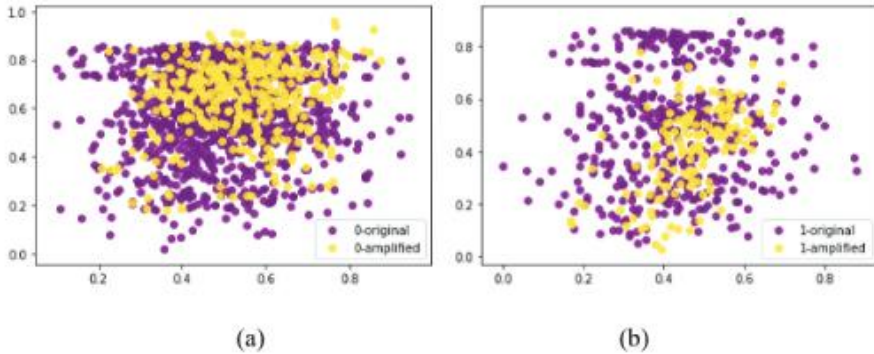
ارائه یک مدل بهینه مبتنی بر هوش مصنوعی... | پرهیزکار میاندهی و دیگران

XGBoost و G-XGBoost استفاده می‌شود که در آن GAN در مدل G-XGBoost به ترتیب با نمونه‌های خوب و بد آموزش داده می‌شود.

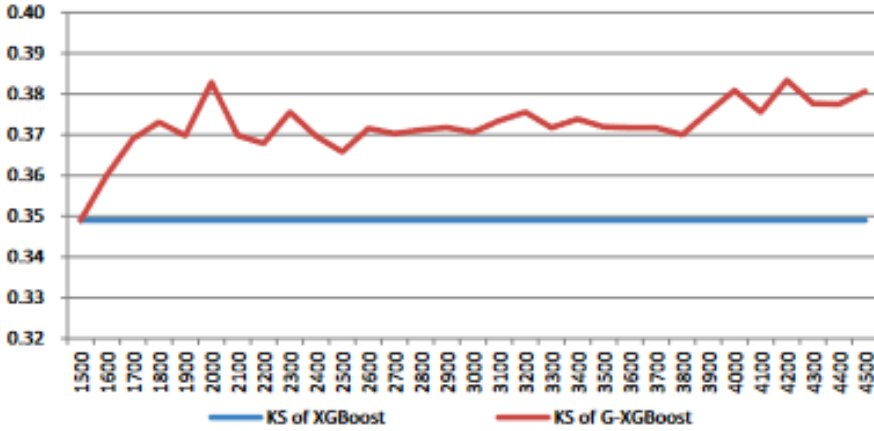
پس از آموزش شبکه، شبه داده توسط مولد آموزش دیده تولید شده و نمونه تقویت شده از داده‌های نمونه اصلی و شبه داده‌ها تشکیل می‌شود. همانطور که در شکل (۴) می‌بینید با اینکه مقدار داده بهبود یافته است باز هم در مقایسه با داده‌های اصلی، توزیع نمونه‌های تقویت شده اساساً یکسان است.

به منظور بررسی تأثیر تعداد نمونه‌های تقویت شده بر نتایج آموزش، هر بار مجموعه نمونه‌ها با ۱۰۰ نمونه تولید شده دیگر با همان نسبت داده‌های اصلی تقویت می‌شود.

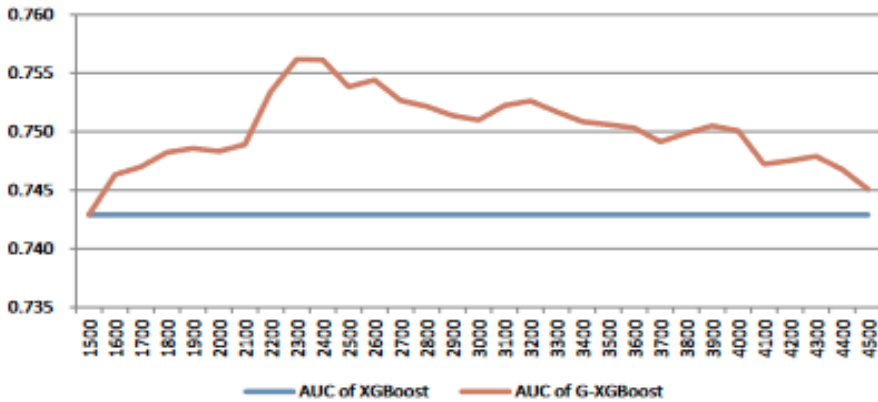
همانطور که در شکل ۵ و ۶ می‌بینید آموزش مدل G-XGBoost با استفاده از داده‌های تقویت شده و محاسبه میانگین و منحنی تغییر KS و AUC، با افزایش حجم نمونه به دست می‌آیند.



شکل ۴. توزیع داده‌های تقویت شده با داده‌های اصلی مطابقت دارد



شکل ۵. KS تحت آستانه‌های نمونه تقویت شده مختلف



شکل ۶. AUC تحت آستانه‌های نمونه تقویت شده مختلف

همانطور که در شکل ۵ می‌بینید، محور X تعداد نمونه تقویت شده (اندازه نمونه اصلی ۵۰۰ است که هر بار ۱۰۰ شبه داده افزایش می‌یابد) و محور Y نشان‌دهنده میانگین KS است. هنگامی که حجم نمونه تقویت شده ۱۵۰۰ تا ۲۰۰۰ باشد، مقدار KS با افزایش حجم نمونه روند صعودی مشخصی را نشان می‌دهد. از ۲۰۰۰ تا ۳۰۰۰، با افزایش نمونه‌ها، مقدار KS اندکی نوسان کرد، اما همچنان به طور قابل توجهی بهتر از نمونه اصلی و غیرتقویت شده بود. سپس مقدار KS به تدریج افزایش یافت و تثبیت شد. به طور کلی، مدل G-XGBoost از نظر مقدار KS بهتر از مدل XGBoost آموزش داده شده توسط نمونه‌های اصلی است.



ارائه یک مدل بهینه مبتنی بر هوش مصنوعی... | پرهیزکار میاندهی و دیگران

همانطور که در شکل ۶ می‌بینید، محور x تعداد نمونه تقویت شده (اندازه نمونه اصلی ۱۵۰۰ است که هر بار ۱۰۰ شبه داده افزایش می‌یابد) و محور y نشان‌دهنده میانگین AUC است. وقتی حجم نمونه تقویت شده ۱۵۰۰ تا ۲۳۰۰ باشد، مقدار AUC افزایش یافته و زمانی که حجم نمونه حدود ۲۳۰۰ باشد به حداکثر مقدار می‌رسد. سپس، مقدار AUC یک نوسان کوچک با روند نزولی کلی را نشان می‌دهد.

در ترکیب با شکل ۵ و ۶، افزایش تعداد نمونه‌ها در مرحله اولیه، اثر مدل را بهبود می‌بخشد. پس از رسیدن به یک آستانه خاص، اثر مدل بهینه خواهد بود. همانطور که در بخش ۱ توضیح داده شد، یک حجم نمونه بهینه وجود دارد که مدل را بهینه می‌کند. پس از آن، افزایش حجم نمونه تاثیر محدود یا حتی منفی بر اثر مدل خواهد داشت. این پدیده ممکن است به انحراف نمونه‌های تولید شده GAN از توزیع نمونه‌های واقعی نسبت داده شود. علاوه بر این، زمانی که مقدار داده‌های تقویت شده خیلی زیاد باشد، ممکن است مستعد خطا باشد. افزودن بسیاری از نمونه‌های منحرف و در نتیجه اشتباه به مجموعه آموزشی اثرات انباشته‌ای در کاهش دقت مدل و کاهش عملکرد خواهد داشت.

به‌طور کلی، مدل G-XGBoost بهتر از مدل فعلی ریسک اعتباری XGBoost که نیاز به انتخاب اندازه نمونه دارد، می‌باشد.

به منظور بررسی بیشتر اثربخشی مدل G-XGBoost، اعتبار سنجی متقابل در آزمایش انجام شده است. نمونه‌های اصلی به طور تصادفی به چهار گروه تقسیم شده که یک گروه مختلف (۵۰۰ نمونه) به عنوان مجموعه تایید و سه گروه دیگر به عنوان مجموعه‌های آموزشی برای آموزش XGBoost و G-XGBoost انتخاب می‌شوند. علاوه بر این، در مدل G-XGBoost از یک گروه آموزشی دیگر متشکل از ۵۰۰ نمونه GAN استفاده شده است. این عمل چهار بار تکرار می‌شود. همانطور که در جدول ۳ می‌بینید، مقادیر KS و AUC در مجموعه تأیید مقایسه می‌شوند.

جدول ۳ ارزش KS، G-XGBoost را که (۵۰۰ نمونه تولید شده) به طور قابل توجهی بالاتر از XGBoost از یک گروه یا میانگین است، نشان داده و می‌بینیم که مدل G-XGBoost واقعاً توانایی تشخیص ریسک را بهبود بخشیده است. جدول ۳ نشان می‌دهد



که دو مدل از نظر دقت پیش‌بینی به‌جز گروه دوم که G-XGBoost (۵۰۰ نمونه تولید شده) کمی دقیق‌تر از XGBoost است، بسیار نزدیک هستند. برای وضوح، مقایسه دقیق‌تری بین دو مدل با استفاده از گروه ۱ همانطور که در جداول ۳ و ۴ نشان داده شده است، انجام شده است.

جدول ۳. مقدار KS و AUC مدل‌های مختلف.

	KS-XGBoost	AUC-XGBoost	KS-GXGBoost	AUC-GXGBoost
گروه ۱	0/2491	0/6241	0/2521	0/6892
گروه ۲	0/2723	0/6315	0/2994	0/6542
گروه ۳	0/2943	0/6254	0/3215	0/6411
گروه ۴	0/2854	0/6551	0/3012	0/6671
میانگین	0/275275	0/634025	0/29355	0/6629

جدول ۴. آمار مجموعه‌های اعتبارسنجی مدل ۱

مرحله	مقادیر پیش‌بینی شده	خوب	بد	درصد خوب	درصد بد	نسبت بد	KS
1	>0/612	68	96	3.7%	23.1%	66%	18/20 %
2	0/522-0/612	112	87	6.9%	19.7%	57%	28.2%
3	0/499-0/522	165	54	7.8%	12/00%	38%	31.3%
4	0/452-0/499	193	62	9.3%	12.3%	32%	33.7%
5	0/398-0/452	201	51	9.6%	10.1%	28%	29.4%
6	0/344-0/398	252	41	11.1%	8.7%	21%	34.7%
7	0/284-0/344	273	18	12.3%	6.6%	18%	35.1%
8	0/201-0/284	312	8	13.4%	4.7%	19%	11.2%
9	0/156-0/201	320	3	14.7%	3.1%	2%	7.1%
10	<0/156	70	1	11.2%	0.7%	6%	0.03%
جمع		1966	42 1	100/00%	100/00 %	28/70 %	

مدل ۱: از مجموعه‌های آموزشی اصلی (۲۳۸۷ نمونه) برای آموزش XGBoost و پیش‌بینی مجموعه‌های اعتبارسنجی استفاده کنید. مقادیر پیش‌بینی شده به ترتیب نزولی

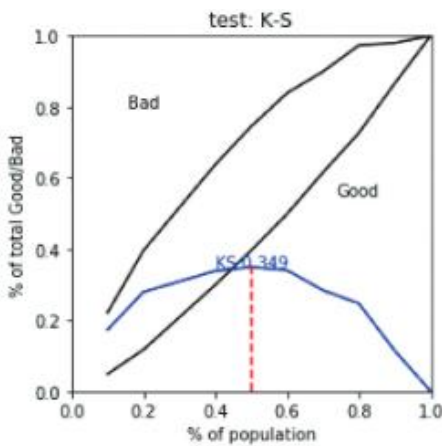


ارائه یک مدل بهینه مبتنی بر هوش مصنوعی... | پرهیزکار میاندهی و دیگران

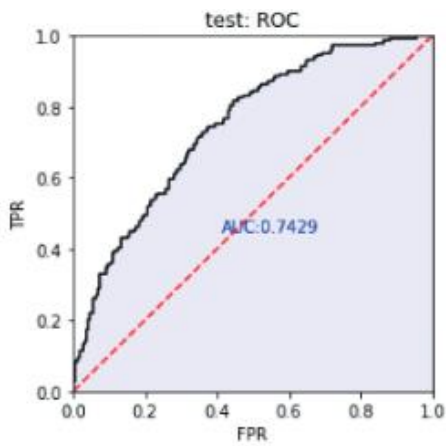
مرتب شده و به ده قسمت تقسیم می شوند. تعداد نمونه‌های خوب و بد مربوط به هر بازه شمارش و نسبت آنها محاسبه می شود. نسبت بدهی بد و ارزش KS نیز، همانطور که در جدول ۴ نشان داده شده، محاسبه می شوند. مشاهده می شود که با کاهش ارزش پیش‌بینی شده نمونه‌ها، نرخ بدهی بد متناظر یک روند نزولی نشان می دهد. نسبت بدهی بد سطح ۱ (گروه با بیشترین ریسک) ۶۶ درصد است که یعنی حدود ۲.۲ برابر مقدار متوسط، بهتر می تواند نمونه‌های بد را تشخیص دهد. اما برای نمونه‌های خوب، سطح 10-7 از کار افتاده و توانایی مرتب‌سازی نسبت بد، ضعیف است.

شکل ۸ بخش (a)، KS مدل ۱ را به طور مستقیم و بخش (b) مقدار AUC مدل XGBoost را نشان می دهد.

مدل ۲: از نمونه‌های تقویت شده (۲۳۸۷ نمونه اصلی + ۵۰۰ نمونه تولید شده) برای آموزش G-XGBoost و پیش‌بینی مجموعه‌های اعتبار سنجی استفاده کنید. مقادیر پیش‌بینی شده به ترتیب نزولی مرتب شده و به ده قسمت مساوی تقسیم می شوند. جدول ۵ مانند جدول ۴ طراحی شده است. در مقایسه با مدل ۱، سطح ۳ بهتر می تواند نمونه‌های بد را تشخیص دهد (نسبت بد بالاتر از مدل ۱ است)، و سطح 10-7 توانایی مرتب‌سازی بهتری نسبت به مدل ۱ دارد که می تواند نمونه‌های خوب را بهتر طبقه‌بندی کند.



(a)



(b)



شکل ۷. نتیجه پیش‌بینی مجموعه‌های اعتبارسنجی مدل ۱

جدول ۵. آمار مجموعه‌های اعتبارسنجی مدل ۲

مرحله	مقادیر پیش‌بینی شده	خوب	بد	درصد خوب	درصد بد	نسبت بد	KS
1	>0/590	68	96	4.8%	23.1%	66%	18/20 %
2	0/500-0/590	112	85	6.8%	19.7%	57%	28.2%
3	0/4329-0/5002	152	54	7.8%	14/00%	44%	34.2%
4	0/397-0/432	163	48	8.3%	11.3%	32%	36.7%
5	0/348-0/397	231	51	9.1%	10.1%	28%	35.7%
6	0/266-0/348	245	53	10.1%	9.6%	16%	41.2%
7	0/201-0/266	241	21	12.3%	5.2%	17%	35.1%
8	0/181-0/201	320	8	14.2%	4/00%	19%	11.2%
9	0/136-0/181	321	4	13.8%	2.3%	2%	7.1%
10	<0/136	113	1	12.8%	0.7%	6%	0.03%
جمع		1966	42 1	100/00%	100/00 %	28/70 %	

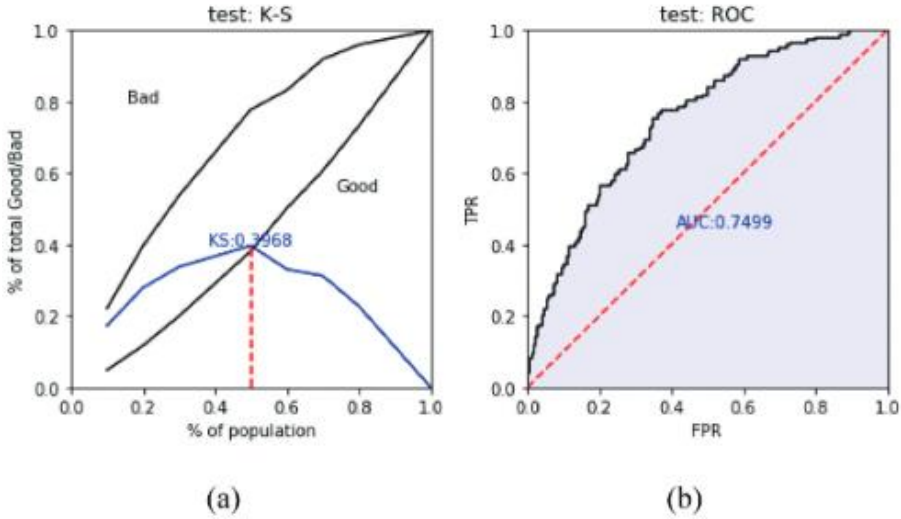
شکل ۸ بخش (a)، KS مدل ۲ و بخش (b) مقدار AUC مدل G-XGBoost را نشان می‌دهد.

با مقایسه شکل ۷ و شکل ۸ می‌توان دریافت که مدل G-XGBoost نه تنها در مقادیر KS نسبت به مدل XGBoost برتری دارد، بلکه در مقادیر AUC نیز بالاتر است. در نتیجه می‌بینیم که مدل G-XGBoost از XGBoost هم در تشخیص ریسک اعتباری و هم در دقت پیش‌بینی بهتر عمل می‌کند.

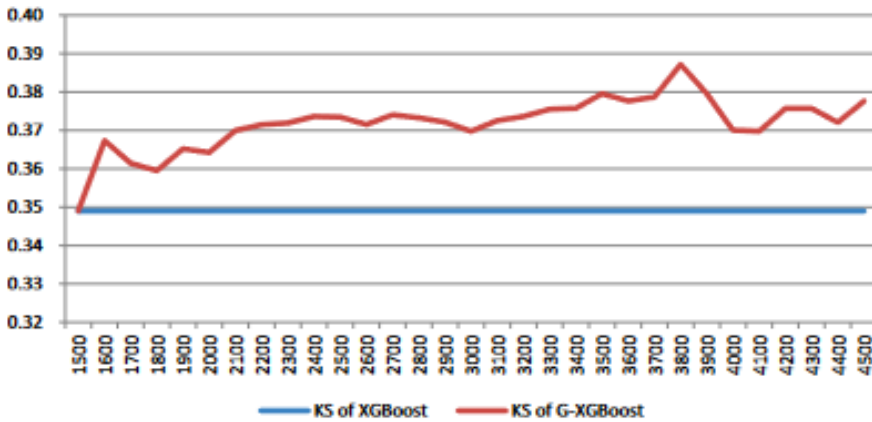
علاوه بر این، برای تقویت نمونه‌های بد آزمایش می‌کنیم. در شکل ۹ منحنی میانگین مقدار KS را با مجموعه نمونه تقویت شده تنها با افزودن نمونه‌های بد می‌بینید. مشاهده می‌شود که مقدار KS در ابتدا به شدت افزایش می‌یابد، سپس به آرامی به رشد ادامه داده و در حالت پایدار با نوسانات جزئی باقی می‌ماند. در مقایسه با شکل ۵، می‌توان دریافت که نتایج تجربی دو روش تقویت مشابه هستند. مقادیر KS در روش شکل ۵ پایدارتر و در روش تقویتی شکل ۹ در مرحله اولیه KS، سریعتر رشد می‌کند.



ارائه یک مدل بهینه مبتنی بر هوش مصنوعی... | پرهیزکار میاندهی و دیگران



شکل ۸. اثر پیش‌بینی مجموعه‌های اعتبارسنجی مدل ۲



شکل ۹. KS فقط از نمونه‌های بد برای تقویت مجموعه نمونه استفاده می‌کند.

در عمل، مؤسسات مالی، قوانین و استراتژی‌هایی را بر اساس توزیع SMEها با احتمالات مختلف مانند پیش‌بینی ریسک و نسبت شرکت‌های خوب و بد تدوین می‌کنند. شرکت‌هایی با احتمال پیش‌بینی بسیار بد، مستقیماً رد شوند و شرکت‌های با کیفیت بالا با ریسک پایین را بر اساس نتایج پیش‌بینی شده درجه‌بندی کرد. SMEهای مختلف دارای مقادیر متفاوتی که می‌تواند نرخ عبور کلی را بهبود بخشد و ریسک را کاهش دهد می‌باشند.



نتیجه گیری

در این تحقیق، از شبکه GAN جهت تولید شبه داده‌ها و غلبه بر تعداد نمونه‌های کم آموزشی استفاده گردید و سپس از طریق طیف گسترده‌ای از آزمایش‌ها مشخص گردید که نمونه تقویت شده می‌تواند عملکرد شبکه XGBoost را بهینه نماید.

نتایج خروجی از KS و AU، بر روی مدل پیشنهادی بیانگر آن است که دقت پیش بینی ریسک اعتباری با توجه به داده‌های کم مشتریان به دلیل امنیت اطلاعات آنها نسبت به مدل پایه عملکرد بسیار بهتری داشته است. همچنین، هیچ نسبت کاملی را در خصوص میزان شبه داده‌های تولید شده خوب و بد نمی‌توان بدست آورد و تولید شبه داده‌ها از هر دسته می‌تواند عملکرد مدل پیشنهادی را بهینه سازد. نتایج خروجی حاصل از این مدل نسبت به مدل پایه بیانگر بهبود ۳۲.۷ درصدی دقت پیش بینی تحلیل ریسک اعتباری مشتریان می‌باشد.

در نهایت پیشنهاد می‌گردد، مجموعه نمونه‌های تقویت شده براساس مدل پیشنهادی جهت دسته‌بندی به سایر روش‌ها مانند کارت امتیاز، شبکه عصبی و... اعمال و نتیجه مقایسه گردد

فهرست منابع

- باقرزاده، عنایت، و میرزازاده، حجت. ۱۳۸۸. مدل‌سازی پیش‌بینی تورم با استفاده از مدل تلفیقی شبکه‌های عصبی مصنوعی با PCA. هفتمین کنفرانس بین‌المللی مدیریت. تهران.
- بزازان، فاطمه، شیرین بخش ماسوله، شمس‌اله، و صفری، سولماز. ۱۳۹۱. بررسی اثرات روزهای هفته بر بازده سهام رویکرد رگرسیون فازی گارچ بوت استرپ. دانش مالی تحلیل اوراق بهادار (مطالعات مالی). ۵(۱۳): ۹۹-۱۱۰.
- تهرانی، رضا، و مرادپور، سعید. ۱۳۹۱. پیش‌بینی بازده شاخص بورس اوراق بهادار با استفاده از مدل‌های شبکه‌ها عصبی مصنوعی شعاع پایه. مهندسی مالی و مدیریت اوراق بهادار (مدیریت پرتفوی). ۳(۱۰): ۷۵-۹۹.
- سینایی، حسنعلی، مرتضوی، سعیداله، و تیموری اصل، یاسر. ۱۳۸۴. پیش‌بینی شاخص بورس اوراق بهادار تهران با استفاده از شبکه‌های عصبی مصنوعی. بررسی‌های حسابداری و حسابرسی. ۱۲(۴۱): ۵۹-۸۳.



ارائه یک مدل بهینه مبتنی بر هوش مصنوعی... | پرهیزکار میاندهی و دیگران

شهرابی، جمال. ۱۳۹۲. داده کاوی. انتشارات جهاد دانشگاهی واحد صنعتی امیرکبیر. تهران.
شیرین بخش ماسوله، شمس‌اله، و صفری، سولماز. ۱۳۹۲. بررسی اثرات روزهای هفته بر بازده
سهام الگوریتم رگرسیون حداقل میانگین مربعات (LMS). مهندسی مالی و مدیریت اوراق
بهادار (مدیریت پرتفوی). ۴(۱۷): ۴۵-۵۹.

راعی، رضا، و شیرزادی، سعید. ۱۳۸۷. بی‌قاعدگی‌های تقویمی و غیر تقویمی در بازارهای مالی.
بورس اوراق بهادار. ۱(۱): ۱۰۱-۱۳۲.

راعی، رضا، نیک‌عهد قصیرائی، علی، و حبیبی، مصطفی. ۱۳۹۵. پیش‌بینی شاخص بورس اوراق
بهادار تهران با ترکیب روش‌های آنالیز مولفه‌های اصلی، رگرسیون بردار پشتیبان و حرکت
تجمعی ذرات. راهبرد مدیریت مالی. ۴(۱۵): ۱-۲۳.

سرائی، محمدحسین، فرکی، مسعود، و کیخایی، شهرام. ۱۳۸۶. یک مورد مطالعه موردی
پارامترهای مدل پیش‌بینی قیمت سهام با استفاده از الگوریتم انتشار بازگشتی. اولین کنفرانس
داده کاوی ایران. تهران.

محدث، فخری. ۱۳۹۰. روش تحلیل مولفه‌های اساسی و بررسی عوامل مطالعه موردی: استخراج
شاخص قیمت دارایی‌ها و بررسی اثر آن بر تورم. مجموعه پژوهش‌های اقتصادی. ۴۱(۴): ۱-۴۲
نظری، محسن، و فرزادگان، الهام. ۱۳۹۰. بی‌قاعدگی‌های دوره‌ای در بازدهی سهام عادی بورس
اوراق بهادار تهران (روش باز نمونه‌گیری بوت استرپ ناپارامتریک). تحقیقات مالی. ۱۳(۳۱):
۱۶۷-۱۴۷.

یحیی‌زاده‌فر، محمود، ابونوری، اسماعیل، و شبابی، هومن. ۱۳۸۴. بررسی اثر روزهای هفته بر بازده
سهام بورس اوراق بهادار تهران و مقایسه آن با سایر بازارهای نوظهور (تحلیل تجربی). علوم
اجتماعی و انسانی دانشگاه شیراز. ۲۲(۲): ۱۷۸-۱۹۵.

Boz, Z., D. Gunec, S. I. Birbil, and M. K. Öztürk. 2018. Reassessment and monitoring of loan applications with machine learning. *Applied Artificial Intelligence* 32 (9–10):939–55. doi:10.1080/08839514.2018.1525517.

Chen, Y. Z., Y. S. Wang, D. Kirschen, and B. Zhang. 2018. Model-free renewable scenario generation using generative adversarial networks. *IEEE Transactions on Power Systems* 33 (3):3265–75. doi:10.1109/TPWRS.2018.2794541.

Feng, G. Q., D. L. Cui, K. Q. Zhu, and Q. Zhang. 2019. Research of modeling with small sample for complex problem. *Control Engineering of China* 26 (11):2013–18.

Frid-Adar, M., E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. 2018. Synthetic data augmentation using GAN for improved liver lesion classification. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 289–93. doi: 10.1109/ISBI.2018.8363576.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29 (5):1189-232. doi:10.1214/aos/1013203451.



- Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative Adversarial Networks, *Advances in Neural Information Processing Systems*, 3:2672–80. [online] Available: <https://arxiv.org/abs/1406.2661>.
- Lee, Y. H., and S. Z. Cho. 2020. Design of semantic-based colorization of graphical user interface through conditional generative adversarial nets. *International Journal of Human-Computer Interaction* 36 (8):699–708. doi:10.1080/10447318.2019.1680921.
- Li, C. Y. 2021a. The influence of sample size change on the prediction accuracy of shanghai stock index. *Henan Science and Technology* 28:8–10.
- Li, Y. 2019b. Credit risk prediction based on machine learning methods. *2019 14th International Conference on Computer Science & Education (ICCSE)*, Toronto, Canada, 10 11–13.
- Li, Z. J., F. Ju, C. B. Xiu, and G. H. Qiao. 2016. The construction of bank credit risk small sample rating model. *Statistics and Decision* 453 (9):41–45. *of Artificial Intelligence* 65 (32). doi: 10.1016/j. engappai.2016.12.002.
- Qiu, W. Y. 2019. Credit risk prediction in an imbalanced social lending environment based on XGBoost. *2018 5th International Conference on Big Data and Information Analytics (BigDIA)*, 150–56. Kunming, China. doi: 10.1109/ BigDIA.2019.8802747.
- Rahbar, M., M. Mahdavejad, M. Bemanian, A. H. Davaie Markazi, and L. Hovestadt. 2019. Generating synthetic space allocation probability layouts based on trained conditional-GANs. *Applied Artificial Intelligence* 33 (8):689–705. doi:10.1080/ 08839514.2019.1592919.
- Wang, B., Y. Kong, T. T. Zhang, D. P. Liu, and L. J. Ning. 2019a. Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications* 128: 15–301. doi:10.1016/ j.eswa.2019.02.033.
- Wang, C. Y., C. Xu, X. Yao, and D. C. Tao. 2019b. Evolutionary generative adversarial networks. *IEEE Transactions on Evolutionary Computation*. [online] Available: <https://arxiv.org/abs/1803.00657> Wang, G. X., W. X. Kang, Q. X. Wu, Z. Y.
- Wang, and J. B. Gao. 2018. Generative adversarial network (GAN) based data augmentation for palmprint recognition. *2018 Digital Image Computing: Techniques and Applications (DICTA)*, 1–7. Canberra, Australia. doi: 10.1109/ DICTA.2018.8615782.
- Wang, H. X., J. D. Zhong, D. F. Zhang, and X. Y. Zou. 2017. A new classification algorithm for the bank customer credit rating. *2017 Ninth International Conference on Advanced Computational Intelligence (ICACI) IEEE*, Doha, Qatar, 143–48. doi: 10.1109/ ICACI.2017.7974499.
- Wiese, M., R. Knobloch, R. Korn, and P. Kretschmer. 2020. Quant GANs: Deep generation of financial time series. *Quantitative Finance* 1–22. doi:10.1080/14697688.2020.1730426.
- Wiginton, J. 1980. A note on the comparison of logit and discriminant models of consumer credit behavior. *The Journal of Financial and Quantitative Analysis* 15 (3):757–70. doi:10.2307/2330408.
- Zheng, L. C. 2014. Internet finance in China: Models, impact, nature and the risks. *International Economic Review* 5:18-103.

