هشتمین کنفرانس بین‌المللی مهندسی صنایع و سیستم‌ها
۱۶ و ۱۷ شهریور ۱۴۰۱ • دانشگاه فردوسی مشهد
8th International Conference on Industrial and Systems Engineering
7 & 8 Sep. 2022 • Ferdowsi University of Mashhad

پیاده‌سازی و مقایسه روش‌های یادگیری ماشین در ارزیابی ریسک اعتباری مشتریان موسسات اعتباری و مالی

زهرا فتحی اقدم[۱]، محمدرضا رسولی[۲]

[۱]کارشناسی، دانشکده مهندسی صنایع، داشگاه علم و صنعت ایران، تهران، ایران؛ zahraf.7811@gmail.com

[۲]استادیار، گروه مهندسی سیستم‌های هوشمند، دانشکده مهندسی صنایع، داشگاه علم و صنعت ایران، تهران، ایران؛ rasouli@iust.ac.ir

چکیده

عمده‌ترین ریسک در موسسات مالی، ریسک اعتباری[۱] است که عدم توانایی قرض‌گیرنده در بازپرداخت تسهیلات و وام در موعد مقرر معنا می‌شود. امروزه کارشناسان موسسات می‌توانند با تحلیل داده‌های مشتریان و استفاده از یادگیری ماشین، برای تخصیص اعتبارات به آن‌ها تصمیم‌گیری نمایند. بنابراین، اعطا و عدم اعطای وام، بدون قضاوت شخصی و بر مبنای ریاضیات انجام می‌گیرد. در تحقیق پیشرو، داده‌های مشتریان موسسه لندینگ‌کلاب[۲] از وبسایت کگل[۳] جمع‌آوری شده‌است و الگوریتم‌های مختلف یادگیری ماشین برای پیش‌بینی ریسک اعتباری بررسی شده‌است. چالش‌های مهم این پژوهش، انتخاب ویژگی‌ها و مدلسازی بر کلاس‌های نامتوازن بود. فلذا، نوآوری کلیدی این پژوهش استفاده از روش‌های انتخاب ویژگی شامل آزمون تحلیل واریانس[۴]، آزمون اطلاعات متقابل[۵] و روش حذف بازگشتی ویژگی[۶] و ترکیب آن‌ها با الگوریتم‌های یادگیری ماشین برای پیش‌بینی ریسک اعتباری می‌باشد. همچنین برای مدلسازی بر کلاس‌های نامتوازن[۷]، الگوریتم‌های حساس به هزینه[۸] مانند درخت تصمیم[۹]، ماشین بردار پشتیبان[۱۰] و الگوریتم‌های جمعی[۱۱] مثل جنگل تصادفی[۱۲] و تقویت گرادیان سبک[۱۳] استفاده شد. دیگر نوآوری پژوهش استفاده از الگوریتم‌های جدیدتر حوزه یادگیری ماشین مانند تقویت گرادیان سبک می‌باشد و نتایج نشان می‌دهد بهترین طبقه‌بندی توسط همین الگوریتم و بر ویژگی‌های منتخب از روش آزمون تحلیل واریانس انجام گرفته‌است. مساحت زیر منحنی مشخصه عملکرد[۱۴] این مدل برابر با ۰,۷۲۴ و معیار اف۰,۵ برابر با ۰,۷۸۸ است.

کلمات کلیدی: ریسک اعتباری، طبقه‌بندی، درخت تصمیم، ماشین بردار پشتیبان، جنگل تصادفی، تقویت گرادیان سبک

Implementation And Comparison Between Machine Learning Techniques For Assessing Credit Risk Of Customers Of Financial Institutions

Zahra Fathi Aghdam, Mohammadreza Rasouli

Bachelor, Faculty Of Industrial Engineering, Iran University Of Science And Technology, Tehran, Iran

Assistant Professor, Faculty Of Industrial Engineering, Iran University Of Science And Technology, Tehran, Iran

ABSTRACT

Profitability in financial institutions and banks always depends on granting loans and facilities. Therefore, one of the major risks in these organizations is credit risk, which means the borrower's inability to repay the loan on time. Data mining is one of the most efficient methods in credit risk management as helps relevant experts in institutions to analyze customer data and make appropriate decisions. As a result, granting or not granting loans is done without personal judgment and based on mathematics and statistics and using intelligent systems. In this research, various machine learning algorithms have been developed on Lending Club dataset, collected from Kaggle website, to assess credit risk. Feature selection and modeling on imbalanced classes were considerable challenges in this research. Therefore, innovation of the research is utilizing three different methods of feature selection named analysis of variance (ANOVA), mutual information and Recursive Feature Elimination (RFE) and combining them with machine learning algorithm to predict credit risk. On the other hand, cost-sensitive algorithms such as decision tree (DT) and support vector machine (SVM) and ensemble methods as random forest (RF) and light gradient boosting (LGBM) were implemented to face imbalanced classes. In spite of other research, newer algorithms like LGBM has been implemented to be consider as an innovation. Finally, the best classifier was LGBM built on features selected by ANOVA which resulted ROC AUC and F0.5 equal to 0.724 and 0.788, respectively.

Keywords: Credit Risk, Classification, Decision Tree, Support Vector Machine, Random Forest, Light Gradient Boosting

۱- مقدمه

سودآوری در موسسات مالی و بانک‌ها همواره به اعطای تسهیلات وابسته است. بنابراین، وام‌ها مهم‌ترین منبع درامد بانک‌ها و موسسات اعتباری و همچنین بزرگترین ریسک برای این سازمان‌ها هستند [۵]. یکی از مهم‌ترین ریسک‌ها در این موسسات ریسک اعتباری نام دارد که عدم توانایی قرض‌گیرنده در بازپرداخت وام در موعد مقرر معنا می‌شود [۶]. در گذشته، موسسات اعتباری و بانک‌ها افرادی را برای ارزیابی پیشینه متقاضیان و تصمیم‌گیری در راستای اعطا و یا عدم اعطای تسهیلات استخدام می‌کردند [۷] ولی همچنان حجم تسهیلات اعطایی سوخت‌شده و معوقه بانک‌ها زیاد بود. این موضوع عدم وجود روش‌های مناسب اعتبارسنجی و سیستم‌های مدیریت ریسک در شبکه مالی را نشان می‌دهد [۱] که ضرورت بهره‌گیری از روش‌های نوین را ملموس می‌سازد. امروزه با توسعه بانکداری الکترونیکی روزانه داده‌های زیادی تولید می‌شود فلذا این موسسات می‌توانند بدون قضاوت شخصی، با تحلیل داده‌های مشتریان و بر مبنای علم ریاضیات و آمار و با بهره‌گیری از سیستم‌های هوشمند در خصوص اعطای تسهیلات به آن‌ها تصمیم‌گیری نمایند تا سرمایه سازمان از دست نرود و منابع به طور بهینه به وام‌های سودآور اختصاص‌یابد. در واقع، از دیدگاه یادگیری ماشین می‌توان این مسأله را با بهره‌گیری از الگوریتم‌های طبقه‌بندی دو کلاسه مدلسازی نمود [۸] که از اهداف این پژوهش است. علاوه بر این، شناسایی شاخص‌های مؤثر بر این ریسک نیز به موضوع قابل توجهی تبدیل شده‌است و موسسات اعتباری و بانک‌ها باید با توجه به پیچیدگی فعالیت‌ها و محیط اقتصادی پیرامونشان، معیارهای مناسبی برای ارزیابی ریسک اعتباری مشتریان انتخاب کنند [۲]. از این رو در این پژوهش به دنبال شناسایی تاثیرگذارترین ویژگی‌ها در قدرت بازپرداخت مشتریان خواهیم بود.

مروری بر تاریخچه ریسک اعتباری نشان می‌دهد که در دهه ۱۶۱۰ چندین بانک اروپایی و آمریکایی ورشکست شدند که علت اصلی آن، اعطای وام‌های با ریسک بالا تحت بحران‌های اقتصادی آن دوران معرفی شد. از این رو برای نخستین بار بیور در سال ۱۶۹۱ مدل "رگرسیون لجستیک " چند متغیره را برای تعیین ورشکستگی شرکت‌ها ارائه کرد [۳]. در پی آن مفاهیم و ایده‌هایی جدید برای تحلیل ریسک اعتباری در سال ۱۹۴۱ ظهور کرد [۹]. با گذر زمان ریسک اعتباری به یک مسأله‌ی مهم داده‌کاوی در حوزه‌ی مالی تبدیل شده و این مهم در سال ۱۹۵۰ معرفی شد [۱۰]. متدولوژی‌های متنوعی برای حل مسأله‌ی طبقه‌بندی ریسک اعتباری وجود دارد. این روش‌ها رگرسیون لجستیک، تحلیل نزدیک‌ترین همسایگی، شبکه بیزین، شبکه عصبی مصنوعی، درخت‌های تصمیم، جنگل تصادفی، الگوریتم ژنتیک، روش‌های تصمیم‌گیری چند معیاره، ماشین بردار پشتیبان و انواع دیگری را شامل می‌شود [۱۱]. جدول زیر الگوریتم‌های به کار گرفته شده در مقالات سال‌های اخیر را نشان می‌دهد.

جدول ۱: الگوریتم‌های یادگیری ماشین مورد استفاده در منابع اخیر

شماره مرجع	سال انتشار	رگرسیون لجستیک	ماشین بردار پشتیبان	درخت تصمیم	جنگل تصادفی	تقویت گرادیان	تقویت گرادیان شدید[۱۵]	تقویت انطباقی[۱۶]	شبکه عصبی
۱۳	۲۰۱۷		*	*					*
۱۴	۲۰۱۸	*			*	*			
۱۵	۲۰۱۹	*	*	*	*				
۱۶	۲۰۱۹						*		
۱۷	۲۰۲۰	*	*	*	*			*	*
۱۸	۲۰۲۰	*		*	*	*		*	*
۱۹	۲۰۲۱	*			*				*
۲۰	۲۰۲۲		*	*	*		*	*	
۲۱	۲۰۲۲	*	*	*	*			*	

بر اساس آنچه در در جدول ۱ مشاهده می‌شود، در این پژوهش نیز از سه الگوریتم درخت تصمیم، جنگل تصادفی و ماشین بردار پشتیبان به عنوان پرتکرارترین الگوریتم‌ها استفاده شده‌است. همچنین مدلسازی با الگوریتم تقویت گرادیان سبک نیز انجام گرفت زیرا این الگوریتم مناسب برای مدلسازی بر روی داده‌های حجم بالا است و در مقالات اخیر نیز استفاده نشده‌است.

۲- بیان مسأله

در حال حاضر به دلیل حجم بالای تسهیلات، ریسک باز پرداخت وام‌ها یک چالش بزرگ برای موسسات اعتباری و بانک‌ها می‌باشد فلذا استفاده از روش‌های اعتبارسنجی و در واقع ارزیابی ریسک اعطای تسهیلات به متقاضیان یکی از راه‌های کارآمد جهت مدیریت ریسک می‌باشد [۱۲]. علاوه بر این، وام‌های سوخت شده (وصول نشدنی) یکی از مسائل قابل توجه است که از طریق استفاده از مدل‌های اعتباری دقیق‌تر می‌توان تا حدودی بر این مسأله فائق آمد [۴]. هدف از این پژوهش پیاده‌سازی و مقایسه روش‌های مختلف یادگیری ماشین جهت تحلیل و پیش‌بینی ریسک اعتباری مشتریان و طبقه‌بندی آن‌ها است.

۳- روش حل

مطالعه منابع نشان می‌دهد روش‌های متنوعی نظیر کریسپ‌دی‌ام[۱۷]، سما[۱۸] و کادی‌دی[۱۹] برای تحلیل داده و حل پروژه‌های داده‌کاوی وجود دارد ولی به استناد انجمن تحقیقات داده‌کاوی، متداول‌ترین آن‌ها کریسپ‌دی‌ام است. برتری این متدولوژی در قابل فهم بودن آن است و همچنین این متدولوژی مبتنی بر تجارب علمی کسانی است که در دنیای واقعی پروژه‌های داده‌کاوی انجام می‌دهند. کریسپ‌دی‌ام از ۶ فاز اصلی تشکیل شده‌است که عبارت‌اند از: شناخت کسب‌وکار، شناخت داده، پیش‌پردازش داده، مدلسازی، ارزیابی و توسعه.

۳-۱- شناخت کسب‌وکار

در ابتدا باید اهداف، فرصت‌ها و عوامل موفقیت کلیدی کسب‌وکار موردنظر بررسی شوند، چرا که باعث افزایش اطلاعات و انجام بهتر کارها می‌شود. در واقع این گام به تعیین زمینه‌ها و عواملی می‌پردازد که داده‌ها می‌توانند ارزش افزوده ایجاد کنند. این پژوهش بر بررسی و پیش‌بینی ریسک اعتباری در موسسه اعتباری لندینگ کلاب متمرکز است. این سازمان بزرگترین موسسه اعطای وام به شکل آنلاین است که وام‌های شخصی و تجاری اعطا می‌کند. تصمیم‌گیری برای اعطای وام و تسهیلات این سازمان را متحمل ریسک اعتباری می‌کند و این ریسک به دو شکل در این کسب‌وکار تعریف می‌شود:

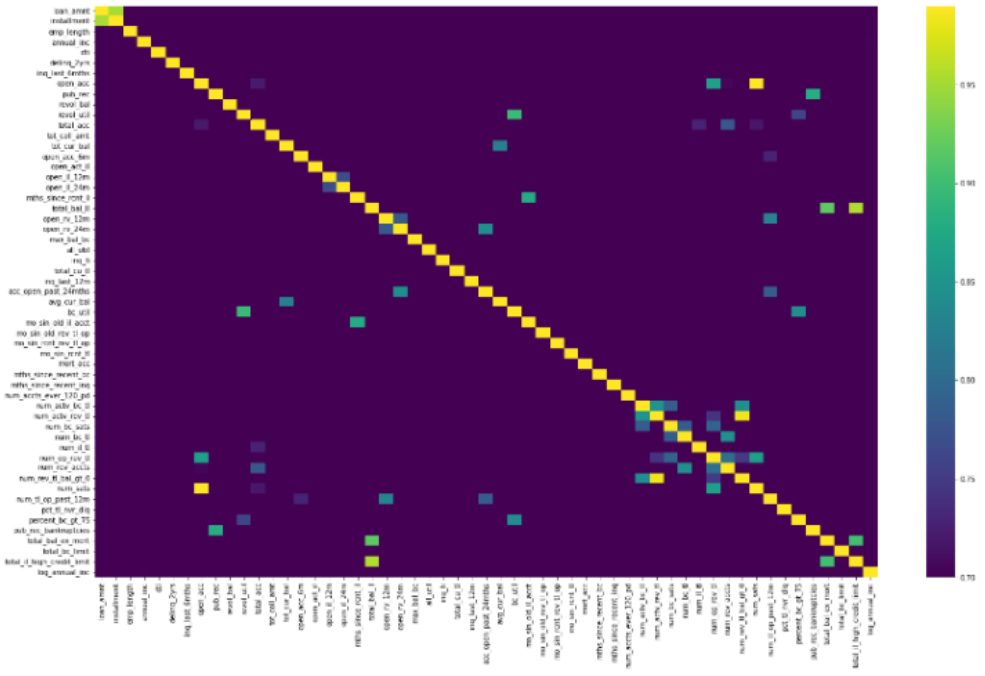
۱. ریسک اعطای وام‌های مشکوک به قصور در بازپرداخت که سبب از دست رفتن سرمایه‌ی سازمان و ضرردهی می‌شود.
۲. ریسک عدم اعطای وام‌هایی که احتمال قصور در بازپرداخت آن‌ها بسیار پایین است. این امر سبب فرصت از دست رفته و کاهش سود سازمان می‌شود.

اگر بتوان درخواست‌های پر ریسک را شناسایی کرده و از اعطای تسهیلات به آن‌ها سر باز زد، شرایط سخت‌گیرانه برای آن‌ها اعمال کرد و یا نرخ بهره‌ی بالاتری برای آن‌ها در نظر گرفت، می‌توان تا حد خوبی ضرر را در این موسسه و سازمان‌های مشابه کاهش داد. همچنین با شناسایی وام‌های کم ریسک و اعطای آن‌ها سود سازمان افزایش میابد.

۳-۲- شناخت داده

این مرحله به جمع‌آوری، توصیف اولیه، کاوش و تایید کیفیت داده اختصاص دارد. داده‌های مورد استفاده در این تحقیق، از طریق وبسایت معتبر کگل جمع‌آوری شده‌است و مربوط به مشتریان لندینگ کلاب در سال‌های ۲۰۰۷ تا ۲۰۱۸ است. این مجموعه داده شامل حدود ۲۵۰۰۰۰۰ موجودیت و ۱۵۱ ویژگی است. ویژگی‌های جمع‌آوری شده برای هر موجودیت در سه دسته شامل ویژگی‌های فرد وام‌گیرنده مانند شغل و سابقه کاری، ویژگی‌های وام نظیر مبلغ وام و تعداد اقساط و نهایتا گزارش و سوابق اعتباری فرد مانند تعداد دفعات قصور در بازپرداخت خلاصه می‌شود و ویژگی هدف که وضعیت وام را شرح می‌دهد از هفت برچسب تشکیل شده‌است.

در این گام نمودارهای متعددی ترسیم شده‌اند که روابط هر ویژگی را با ویژگی هدف و با سایر ویژگی‌ها نشان می‌دهند. به عنوان مثال تصویر ۱ همبستگی ویژگی‌های عددی را با یکدیگر به نمایش می‌گذارد که نشان می‌دهد برخی زوج ویژگی‌ها بسیار به یکدیگر همبسته هستند و باید یکی از آن‌ها حذف شوند.



تصویر ۱: همبستگی ویژگی‌ها با یکدیگر

۳-۳- پیش‌پردازش داده

پیش‌پردازش داده‌ها به جهت بهبود کیفیت داده‌های واقعی برای داده‌کاوی لازم است. بنابراین در ابتدا به جهت نمونه‌گیری از داده‌ها، بر اطلاعات سال ۲۰۱۸ تمرکز شده و همچنین ویژگی هدف به دو حالت "قصور در بازپرداخت" و "عدم قصور در بازپرداخت" و یا به طور خلاصه وام‌های "خوب" و "بد" خلاصه شده‌است. وام‌های بدون قصور ۷۱ درصد حجم موجودیت‌ها را تشکیل می‌دهند که نشان‌دهنده نامتوازن بودن کلاس‌ها می‌باشد. جهت مدیریت داده‌های مفقود، در ویژگی میزان سابقه کار متقاضی، داده‌های مفقود با مقدار صفر جایگزین شده و در سایر ویژگی‌ها با حداکثر مقدار در آن ویژگی جایگزین شده‌اند. انتخاب ویژگی‌ها از اصلی‌ترین چالش‌ها در این پژوهش است فلذا در ابتدا ویژگی‌هایی که بیش از ۲۰ درصد موجودیت‌ها در آن‌ها مفقود هستند حذف شده و همچنین اطلاعات مربوط به تعدادی از ویژگی‌ها در زمان تصمیم‌گیری اولیه موجود نیستند و پس از اعطای وام ایجاد می‌شوند، بنابراین این دسته ویژگی‌ها نیز حذف شده‌اند تا از نشتی اطلاعات[۲۰] و بیش‌برازش جلوگیری شود. در ادامه ویژگی‌های اسمی که تنوع بسیار بالایی دارند و ویژگی‌های عددی که تنها یک مقدار ثابت دارند یا واریانس آن‌ها بسیار کم است نادیده گرفته شده‌اند. علاوه بر آن، گام شناخت داده نشان می‌دهد که نرخ وام‌های بد در مقادیر مختلفی از بعضی ویژگی‌ها مقدار نسبتا ثابتی است و می‌توان نتیجه گرفت آن ویژگی‌ها تاثیر ویژه‌ای بر ویژگی هدف ندارند پس می‌توان آن‌ها را حذف نمود. نهایتا ۵۸۰۳۸ موجودیت و ۵۸ ویژگی باقی ماند که با استفاده از سه روش آزمون تحلیل واریانس، آزمون اطلاعات متقابل و حذف بازگشتی ویژگی، ویژگی‌های موثرتر از میان آن ۵۸ ویژگی انتخاب شده‌است. در گام بعدی مدلسازی بر روی ویژگی‌های منتخب هر روش به تفکیک شرح داده خواهد شد.

در انتها و پیش از مدلسازی و ارائه طبقه‌بندها، ۷۰ درصد داده‌ها برای آموزش و ارزیابی از طریق روش کراس‌ولیدشن[۲۱] جداسازی شده و ۳۰ درصد باقی مانده به عنوان داده‌ی آزمون و نماینده‌ی دنیای واقعی در نظر گرفته شده‌است. همچنین داده‌های عددی نرمالسازی شده و داده‌های اسمی به داده‌های موهومی[۲۲] تبدیل شده‌اند.

۴-۳- ایجاد و پیاده‌سازی طبقه‌بندها

در این مرحله به انتخاب الگوریتم مدلسازی و اجرای آن پرداخته می‌شود. در ادامه‌ی مراحل قبلی، طبقه‌بندی کلاس‌های نامتوازن دومین چالش در این پژوهش است. دو روش موثر بر این چالش، طبقه‌بندی با الگوریتم‌های حساس به هزینه و همچنین الگوریتم‌های جمعی است. بنابراین دو الگوریتم درخت تصمیم و ماشین بردار پشتیبان به عنوان طبقه‌بندهای حساس به هزینه و همچنین جنگل تصادفی و تقویت گرادیان سبک به عنوان نماینده‌ی الگوریتم‌های جمعی مورد استفاده قرار گرفته‌اند. همانطور که پیش‌تر اشاره شد از سه روش برای انتخاب ویژگی استفاده شده‌است و حال بر ویژگی‌های منتخب از هر روش، هر چهار الگوریتم مورد بحث پیاده‌سازی شده‌اند و در مجموع ۱۲ طبقه‌بند حاصل از حالات مختلف انتخاب ویژگی و الگوریتم پیش‌بینی موجود است.

مدلسازی با مقادیر پیش فرض در هر الگوریتم موجب بیش‌برازش شد فلذا تنظیم پارامترهای هر الگوریتم انجام گرفت تا از پیچیدگی بیش از حد مدل‌ها جلوگیری شود. از روش گریدسرچ‌سی‌وی[۲۳] تحت حالت ۵ فولد برای انتخاب بهترین مقادیر پارامترها استفاده شده‌است. همچنین پارامترها به گونه‌ای انتخاب شده‌اند تا مدلی پایدار حاصل شود و نه مدلی با مقدار بسیار بالا برای معیار صحت[۲۴] و بسیار پایین برای معیار حساسیت[۲۵] و بلعکس. زیرا بالا بودن هر دوی آن‌ها مطلوب مسأله است. جدول ۲ پارامترهای بهینه معرفی شده توسط گریدسرچ سی‌وی در هر الگوریتم را بیان می‌کند. نهایتا نتایج مدلسازی با پارامترهای منتخب در بخش بعدی مطرح خواهد شد.

جدول ۲: پارامترهای منتخب برای هر الگوریتم حاصل از روش گریدسرچ سی‌وی

الگوریتم / انتخاب ویژگی	درخت تصمیم	ماشین بردار پشتیبان	جنگل تصادفی	تقویت گرادیان سبک
آزمون تحلیل واریانس	عمق درخت = ۲ حداقل تعداد نمونه برای اشعاب=۲ معیار: آنتروپی	سی= ۱ گاما= ۰,۱	عمق درخت‌ها = ۸ حداقل تعداد نمونه برای اشعاب= ۸ تعداد درخت‌ها= ۲۰	عمق درخت‌ها= ۶ نرخ‌یادگیری= ۰,۱
آزمون اطلاعات متقابل	عمق درخت = ۲ حداقل تعداد نمونه برای اشعاب=۲ معیار: آنتروپی	سی= ۱ گاما= ۰,۱	عمق درخت‌ها = ۸ حداقل تعداد نمونه برای اشعاب= ۸ تعداد درخت‌ها= ۲۵	عمق درخت‌ها= ۶ نرخ‌یادگیری= ۰,۱
حذف بازگشتی ویژگی	عمق درخت = ۲ حداقل تعداد نمونه برای اشعاب=۲ معیار: آنتروپی	سی= ۱ گاما= ۰,۱	عمق درخت‌ها = ۸ حداقل تعداد نمونه برای اشعاب= ۲ تعداد درخت‌ها= ۲۰	عمق درخت‌ها= ۶ نرخ‌یادگیری= ۰,۰۷

۵-۳- ارزیابی

معیارهای متنوعی برای انتخاب بهترین طبقه‌بند وجود دارد. از آنجایی که در این پژوهش کلاس‌ها نامتوازن هستند، معیار دقت مناسب نیست حال آن که دو معیار امتیاز اف و مساحت زیر نمودار منحنی مشخصه عملکرد توانایی و عملکرد مدل را به خوبی بازتاب می‌دهند.

همانطور که پیش‌تر اشاره شد بالا بودن و نطدیک بودن مقدرا هر دو معیار صحت و حساسیت به طور همزمان مطلوب مسئله است ولی در نگاهی دقیق‌تر مثبت کاذب از اهمیت نسبی بیشتری نسبت به منفی کاذب برخوردار است، زیرا از اعطای وام‌های مشکوک به قصور در بازپرداخت جلوگیری می‌کند و مانع از دست رفتن سرمایه سازمان می‌شود حال آن که زیاد بودن منفی کاذب سبب فرصت از دست رفته برای سازمان است. بنابراین امتیاز اف۰٫۵ مناسب‌تر از اف۱ است زیرا بزرگ بودن صحت را بر حساسیت اولویت می‌دهد. عبارت ۱ معیار اف۰٫۵ را معرفی می‌کند.

$$\frac{(1+0.5^2)*\text{صحت}*\text{حساسیت}}{(0.5^2*\text{صحت})+\text{حساسیت}} \quad (۱)$$

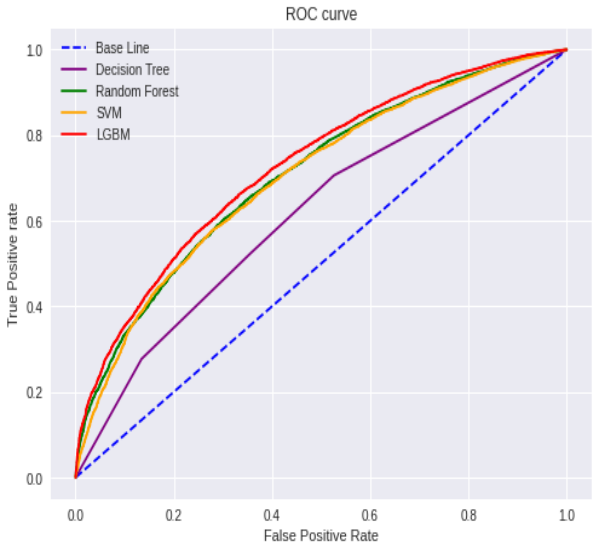
جدول ۳: نتایج حاصل از طبقه‌بندی بر ویژگی‌های استخراجی از سه روش مورد بحث، (بهترین طبقه بند، پررنگ نگارش شده‌است)

روش انتخاب ویژگی	الگوریتم یادگیری ماشین	معیار اف۰٫۵	AUC-ROC
آزمون تحلیل واریانس	درخت تصمیم	۰٫۷۵۷۲	۰٫۶۱۵۹
	ماشین بردار پشتیبان	۰٫۷۷۸۷	۰٫۶۹۸۲
	جنگل تصادفی	۰٫۷۸۰۴	۰٫۷۰۶۷
	تقویت گرادیان سبک	**۰٫۷۸۸۷**	**۰٫۷۲۴۶**
آزمون اطلاعات متقابل	درخت تصمیم	۰٫۷۵۷۲	۰٫۶۱۵۹
	ماشین بردار پشتیبان	۰٫۷۷۶۱	۰٫۶۹۱۴
	جنگل تصادفی	۰٫۷۷۸۲	۰٫۷۰۲۳
	تقویت گرادیان سبک	۰٫۷۸۴۴	۰٫۷۱۷۷
حذف بازگشتی ویژگی	درخت تصمیم	۰٫۷۵۷۲	۰٫۶۱۵۹
	ماشین بردار پشتیبان	۰٫۷۷۵۹	۰٫۶۸۹۹
	جنگل تصادفی	۰٫۷۷۸۹	۰٫۷۰۴۷
	تقویت گرادیان سبک	۰٫۷۸۸۵	۰٫۷۲۳۹

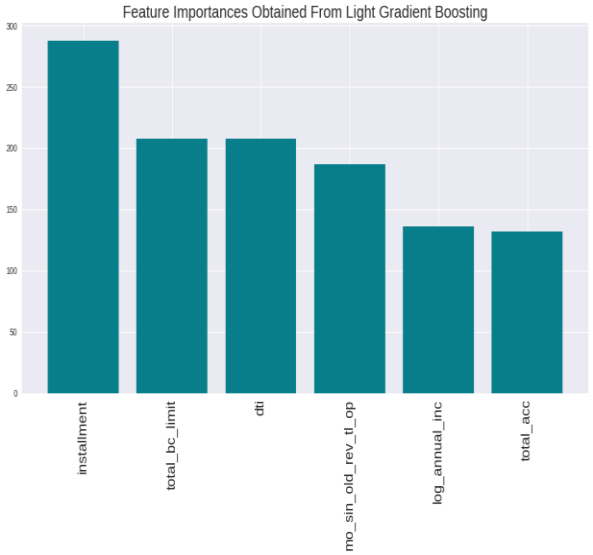
بر اساس نتایج، به طور کلی هر چهار الگویتم زمانی عملکرد بهتری دارند که ویژگی‌ها با روش آزمون تحلیل واریانس انتخاب شده‌اند و نتایج مدلسازی تحت این حالت، نشان می‌دهد که الگوریتم تقویت گرادیان سبک عملکرد بهتری نسبت به سایرین دارد.

تصویر۲ مقادیر امتیاز اف۰٫۵را در ترکیب‌های مختلف الگوریتم‌ها و روش‌های انتخاب ویژگی با یکدیگر مقایسه می‌کند. تصویر۳ منحنی‌های مشخصه عملکرد حاصل از مدلسازی با چهار الگوریتم بر روی ویژگی‌های منتخب از روش آزمون تحلیل واریانس را نشان می‌دهد. منحنی تقویت گرادیان سبک بهترین عملکرد و بیشترین مساحت زیرنمودار را نسبت به سایرین دارد و نسبت به خط مبنا بهبود خوبی یافته است. همچنین تصویر۴ تاثیرگذارترین ویژگی‌ها را در عملکرد متقاضیان برای بازپرداخت وام به نمایش میگذارد. این نمودار نشان می‌دهد هر سه دسته ویژگی اشاره شده در قبل، تاثیر به سزایی بر این موضوع دارند. میزان اقساط وام از جمله ویژگی‌های مربوط به وام، درآمد شخص از جمله ویژگی‌های مربوط به شخص و نهایتا چهار ویژگی دیگر از سوابق اعتباری فرد استخراج می‌شوند.

تصویر ۲: مقادیر اف۰٫۵ حاصل از چهار روش طبقه‌بندی بر روی ویژگی‌های منتخب استخراجی از روش‌های انتخاب ویژگی مذکور



تصویر ۳: منحنی مشخصه عملکرد چهار طبقه‌بند بر ویژگی‌های منتخب از روش آزمون تحلیل واریانس



تصویر ۴: مهم‌ترین ویژگی‌ها در مدل حاصل از ترکیب آزمون تحلیل واریانس و الگوریتم تقویت گرادیان سبک

۶-۳- توسعه

لیندینگ کلاب می‌تواند این برترین طبقه‌بند را به عنوان ابزاری جهت تصمیم‌گیری برای اعطای تسهیلات به متقاضیان مورد استفاده قرار دهد تا عملکرد کارشناسان مربوطه بهبود و خطای آن‌ها کاهش دهد. در نتیجه با مدیریت درست ریسک اعتباری ضررهای سازمان کاهش میابد.

۴- جمع‌بندی و نتیجه گیری

بر اساس مروری بر منابع، داده‌کاوی و پیاده‌سازی الگوریتم‌های یادگیری ماشین سبب مدیریت ریسک اعتباری در موسسات اعتباری و بانک‌ها می‌شود. این پژوهش نیز با مطالعه موردی بر داده‌های لندینگ‌کلاب نشان داد با تکیه بر متدولوژی کریسپ و توسعه مدل‌ها می‌توان در خصوص متقاضیان هوشمندانه تصمیم‌گیری نمود و ریسک اعتباری را کاهش داد. عملکرد خوب این مدل می‌تواند ضررهای ناشی از اعطای وام‌های مشکوک به قصور در بازپرداخت را در این موسسه اعتباری کاهش دهد. همچنین فرصت‌های از دست رفته به شکل عدم اعطای وام‌های بدون قصور در بازپرداخت را تعدیل نماید.

نوآوری کلیدی در این پژوهش، بر استفاده از روش‌های انتخاب ویژگی و ترکیب آن‌ها با الگوریتم‌های مختلف یادگیری ماشین به جهت مدیریت ریسک اعتباری متمرکز می‌باشد. سه روش آزمون تحلیل واریانس، آزمون اطلاعات متقابل و روش حذف بازگشتی ویژگی برای انتخاب ویژگی‌ها مورد استفاده قرار گرفتند و پیش‌بینی ریسک اعتباری بر روی خروجی‌های روش‌های مذکور به صورت تفکیکی انجام گرفت. همچنین سعی شده از الگوریتم‌های جدیدتر حوزه یادگیری ماشین مانند الگوریتم تقویت گرادیان سبک برای پیش‌بینی ریسک اعتباری استفاده شود.

آنچنان که پیش‌تر اشاره شد دو چالش مهم در این پژوهش انتخاب ویژگی‌ها و مدلسازی بر کلاس‌های نامتوازن بود. نتایج نشان می‌دهد بهترین روش برای انتخاب ویژگی در این مجموعه داده، آزمون تحلیل واریانس و بهترین روش برای مقابله با کلاس‌های نامتوازن استفاده از یک الگوریتم جمعی مثل تقویت گرادیان سبک است زیرا الگوریتم‌های پرتکرار در این حوزه در این مطالعه موردی آنچنان موفق نبودند. در نهایت پیشنهاد می‌شود در پژوهش‌های آتی از دیگر الگوریتم‌های یادگیری ماشین و یا دیگر روش‌های انتخاب ویژگی استفاده شود. همچنین استفاده از روش‌های تولید موجودیت[۲۶] و یا کاهش موجودیت[۲۷] در راستای متوازن نمودن کلاس‌ها نیز می‌تواند مفید باشد.

مراجع

[۱] کیقبادی، ا.، و خدامی، ۱۳۹۲، داده‌کاوی صورت‌های مالی جهت اعطای تسهیلات مالی. پژوهش‌های حسابداری مالی و حسابرسی (پژوهشنامه حسابداری مالی و حسابرسی)، ۵(۱۷)، ۲۱۱-۱۷۹.

[۲] عبدالهی، علی و فرزی زاده، محمد، ۱۳۹۶، ارایه‌یک مدل دسته‌بندی به کمک داده‌کاوی جهت امتیازدهی مشتریان بانک، چهارمین کنفرانس ملی فناوری اطلاعات، کامپیوتر و مخابرات، مشهد.

[۳] صفری، زهره و خسروی، حمید، ۱۳۹۵، ارائه مدلی برای اعتبارسنجی مشتریان بانک‌ها با استفاده از الگوریتم‌های داده‌کاوی،دومین کنفرانس ملی رویکردهای نوین در مهندسی کامپیوتر و برق.

[۴] طلوعی اشلقی، عباس و مقدوری شربیانی، فرناز و دانشگر، فرید، ۱۳۸۸، امتیاز دهی اعتباری متقاضیان کارتهای اعتباری بانکها با استفاده از تکنیک ماشین بردار پشتیبان،دومین کنفرانس بین المللی شهر الکترونیک،تهران.

[۵] Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. In IOP Conference Series: Materials Science and Engineering (Vol. 1022, No. 1, p. 012042).

[۶] Aslam, U., Tariq Aziz, H. I., Sohail, A., & Batcha, N. K. (2019). An empirical study on loan default prediction models. Journal of Computational and Theoretical Nanoscience, 16(8), 3483-3488

[۷] Ahmed, M. I., & Rajaleximi, P. R. (2019). An empirical study on credit scoring and credit scorecard for financial institutions. Int. Journal of Advanced Research in Computer Engineering & Technol. (IJARCET), 8, 275-9

[۸] Zhou, L., & Wang, H. (2012). Loan default prediction on large imbalanced data using random forests. TELKOMNIKA Indonesian Journal of Electrical Engineering, 10(6), 1519-1525.

[۹] Durand, D. (1941). Risk elements in consumer installment financing. National Bureau of Economic Research, New York.

[۱۰] Luo, S., Kong, X., & Nie, T. (2016). Spline based survival model for credit risk modeling. European Journal of Operational Research, 253(3), 869-879.

[۱۱] Zhang, Z., Gao, G., & Shi, Y. (2014). Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors. European Journal of Operational Research, 237(1), 335-348.

[۱۲] Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. Expert systems with applications, 36(2), 3302-3308.

[۱۳] Pandey, T. N., Jagadev, A. K., Mohapatra, S. K., & Dehuri, S. (2017, August). Credit risk analysis using machine learning classifiers. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 1850-1854). IEEE.

[۱۴] Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. Risks, 6(2).

[۱۵] Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. Procedia Computer Science, 162, 503-513.

[۱۶] Qiu, W. (2019, July). Credit risk prediction in an imbalanced social lending environment based on XGBoost. In 2019 5th International Conference on Big Data and Information Analytics (BigDIA) (pp. 150-156). IEEE.

[۱۷] Tian, Z., Xiao, J., Feng, H., & Wei, Y. (2020). Credit risk assessment based on gradient boosting decision tree. Procedia Computer Science, 174, 150-160.

[۱۸] Song, Y., Wang, Y., Ye, X., Wang, D., Yin, Y., & Wang, Y. (2020). Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending. Information Sciences, 525, 182-204.

[۱۹] Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. Expert Systems with Applications, 165, 113986.

[۲۰] Dornigg, T. (2022). Credit risk modeling-predicting customer loan defaults with machine learning models (Doctoral dissertation).

[۲۱] Laulajainen, M. (2022). Case Study on Utilizing Machine Learning in Corporate Default Risk Prediction: A practical Implementation to Credit Risk Management Process.

---

۱ Credit Risk
۲ Lending Club
۳ www.kaggle.com
۴ Analysis of Variance (ANOVA)
۵ Mutual Information
۶ Recursive Feature Elimination (RFE)
۷ Imbalanced Classes
۸ Cost-Sensitive
۹ Decision Tree (DT)
۱۰ Support Vector Machine (SVM)
۱۱ Ensemble Models
۱۲ Random Forest (RF)
۱۳ Light Gradient Boosting (LGBM)
۱۴ Area Under The Receiver Operating Characteristic (AUC-ROC)
15 Extreme Gradient Boosting (XGBoost)
۱۶ Adaptive Boosting (Adaboost)
۱۷ Crisp-DM (Cross-industry standard process for data)
۱۸ SEMMA (Sample, Explore, Modify, Model, Assess)
۱۹ KDD (Knowledge Discovery in Databases)
۲۰ Data Leakage
۲۱ Cross Validation

[۲۲] Dummy Variables
[۲۳] Grid Search CV
[۲۴] Precision
[۲۵] Recall
[۲۶] Oversampling
[۲۷] Undersampling