یک روش ترکیبی مؤثر برای پیش‌بینی ریزش و الگوپردازی مشتری

لادن فریدی[۱]، زهرا رضائی قهرودی[۲]

تاریخ دریافت: ۱۴۰۳/۱۰/۲۱
تاریخ پذیرش: ۱۴۰۳/۱۱/۲۳

چکیده:
یکی از نگرانی‌های عمدهٔ اقتصادی بسیاری از شرکت‌ها از جمله بانک‌ها ریزش مشتری است و بانک‌ها توجه خود را بر حفظ مشتری متمرکز کرده‌اند، زیرا هزینه‌های جذب یک مشتری جدید بسیار بیشتر از هزینه‌های نگهداری یک مشتری است؛ بنابراین، پیش‌بینی و الگوپردازی ریزش مشتریان دو دغدغه اقتصادی مهم برای بسیاری از شرکت‌هاست. روش‌های مختلف یادگیری ماشین، برای این اهداف پیشنهادشده‌اند، اما انتخاب بهترین مدل برای انجام این دو امر، به دلیل وابستگی زیاد به ویژگی‌های ذاتی داده‌های ریزش، کار ساده‌ای نیست. در این مقاله، چندین روش یادگیری ماشین با رویکردهای مختلف بازنمونه‌گیری برای متعادل‌سازی داده‌ها، روی داده‌های بانک پیاده‌سازی شده است. ارزیابی‌ها که بر اساس معیار سطح زیر منحنی و نرخ مثبت درست گزارش‌شده‌اند، تأثیر روش‌های متعادل‌سازی و عملکرد روش‌های مختلف یادگیری ماشین را بررسی می‌کند. در این مطالعه، مناسب‌ترین روش‌ها در زمینه ریزش به همراه یک فرآیند مؤثر مبتنی بر رویکرد ترکیبی و خوشه‌بندی معرفی‌شده است. این روش‌ها می‌تواند به خدمات بازاریابی یا منابع انسانی در درک الگوهای رفتاری مشتریان و احتمال ریزش آن‌ها کمک کند.

واژه‌های کلیدی: ریزش مشتری، الگوپردازی مشتری، روش‌های یادگیری ماشین، رویکرد ترکیبی، مشتری‌های بانک، معیار سطح زیر منحنی، متعادل‌سازی داده‌ها

۱ مقدمه

در فضای رقابتی امروز، مدیریت روابط مشتری (CRM) [۳]، به یکی از اولویت‌های اصلی بخش‌های مدیریت و بازاریابی تبدیل‌شده است. به‌ویژه، حفظ مشتریان توجه بسیاری را به خود جلب کرده است، زیرا تحقیقات نشان داده است که مشتریان وفادار می‌توانند با تبلیغ شفاهی مثبت، به رشد شرکت کمک کنند [۲۱]. چنین رفتاری می‌تواند در نهایت هزینه‌های بازاریابی برای جذب مشتریان جدید را کاهش دهد. علاوه بر این، بررسی‌ها نشان می‌دهد که هزینه‌های جذب مشتری جدید معمولاً بسیار بیشتر از هزینه‌های حفظ مشتری موجود است [۲۲]. بنابراین، جلوگیری از ریزش مشتریان می‌تواند برای شرکت‌هایی که مبتنی بر خدمات اشتراکی هستند و به درآمدهای ثابت و منظم عضویت، متکی هستند در صنایع گوناگون از جمله بیمه، بانکداری، بازی‌های ویدئویی آنلاین، پخش موسیقی، خدمات آنلاین یا ارتباطات از راه دور بسیار حیاتی باشد. به همین دلیل، پیش‌بینی دقیق مشتریانی که احتمال ریزش دارند در بسیاری از صنایع به یک الویت تبدیل‌شده است [۵].

امروزه علاوه بر پیش‌بینی ریزش مشتری، بررسی رابطهٔ بین رضایت مشتری، کیفیت خدمات و رفتار مشتری - از جمله وفاداری یا ریزش مشتری- حوزهٔ مهمی از تحقیقات است. در واقع، درک بهتر از تجربه مشتری، اطلاعات ارزشمندی را برای بازاریابان فراهم می‌کند. به‌عنوان‌مثال، مشتریان راضی نسبت به افزایش قیمت‌ها تحمل بیشتری نشان می‌دهند که این امر می‌تواند سود بیشتری به همراه داشته باشد. بااین‌حال، گروه‌های خاصی از مشتریان ممکن است دیدگاه‌های متفاوتی نسبت به ارائه‌دهندگان خدمات داشته باشند. برای مثال، بسیاری از مطالعات نشان می‌دهند که رضایت مشتری ترکیبی از عواملی مانند تصویر شرکت، سازمان داخلی، محیط فیزیکی، خدمات کارکنان و تعامل شخصی با مشتری است. در صنعت بانکداری، لاروش و همکاران [۱۵] نشان داده‌اند که عواملی مانند سرعت خدمات، راحتی دسترسی، شایستگی کارکنان و دوستانه بودن محیط بانک بر رضایت مشتری تأثیرگذارند. از سوی دیگر، تقسیم‌بندی مشتریان به گروه‌های مختلف، متناسب با انتظارات و رفتار آن‌ها، راهکاری مؤثر برای

[۱] دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه تهران.
[۲] دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه تهران. (نویسنده مسئول: z.rezaeigh@ut.ac.ir)

[3] customer relationship management

مدیریت بهینه رفتار ریزش است.

ازآنجاکه اثرات منفی ریزش مشتری - مانند کاهش درآمدها یا هزینه‌های اضافی جذب مشتریان جدید - به‌راحتی قابل‌مشاهده هستند، دلایل ریزش به‌طور مستمر موردمطالعه قرارگرفته‌اند. در خصوص توصیف الگوپردازی رضایت مشتریان، آتناسپولوس [۱۴] به پنج بُعد برای توصیف الگوپردازی‌های مختلف رضایت مشتری در خدمات بانکداری شامل خدمات کارکنان، الگوپردازی کسب‌وکار، نوآوری، راحتی و قیمت اشاره‌کرده‌اند. او همچنین به گروه‌بندی مشتریان بر اساس انتظارات خاص آن‌ها تأکید کرده است. هدف از گروه‌بندی مشتریان که یکی از روش‌های پرکاربرد در مطالعات بازاریابی است، این است که مشتریان مناسب برای هر طرح بازاریابی انتخاب شوند. این کار معمولاً باعث افزایش سودآوری از طریق هدف‌گیری دقیق مشتریان می‌شود. هرسال روش‌های جدیدی برای تقسیم‌بندی مشتریان توسعه‌یافته است [۳] که مقایسه جامع بین آن‌ها را دشوار می‌کند.

هدف این مقاله، بررسی عملکرد چندین الگوریتم یادگیری ماشین برای پیش‌بینی و الگوپردازی ریزش مشتریان است که در کار کاربردی، ریزش مشتریان بانک در نظر گرفته‌شده است. متغیر هدف در این مطالعه یک متغیر دودویی است که وضعیت ریزش مشتریان بانک را نشان می‌دهد. ریزش بدان معنا است که مشتری حساب بانکی خود را می‌بندد و بانک را ترک می‌کند و غیر ریزش بدان معنا است که مشتری نسبت به بانک وفادار و همچنان مشتری بانک باقی می‌ماند. پیش‌بینی ریزش به‌عنوان یک مسئله رده‌بندی دودویی مدل‌سازی می‌شود که هدف آن برآورد احتمال تعلق هر نمونه به یکی از این دو رده است.

روش‌های یادگیری ماشین متعددی در حوزهٔ ریزش مشتری استفاده می‌شوند. این روش‌ها شامل رویکردهای نظارتی و غیرنظارتی است که به پیش‌بینی ریزش مشتری یا الگوپردازی رفتار مشتری می‌پردازد. روش‌های k-نزدیک‌ترین همسایه، رده‌بندهای بیز ساده، رگرسیون خطی، رگرسیون لوژستیک، تحلیل تشخیصی خطی، یادگیری درخت تصمیم و ماشین بردار پشتیبان از جمله الگوریتم‌های نظارتی پرکاربرد در زمینه پیش‌بینی ریزش مشتری هستند [۶]. علاوه بر این، استفاده از رویکردهای ترکیبی مانند جنگل تصادفی، روش تقویت سازوار، تقویت گرادیان یا تقویت گرادیان شدید برای پیش‌بینی ریزش مشتری مطرح‌شده است. همچنین رویکردهای یادگیری عمیق و روش‌های نیمه‌نظارتی مناسبی نیز در این زمینه به‌کاررفته است. در صنعت مالی، به‌ویژه، در زمینه تشخیص جرائم اقتصادی مانند تقلب مالی و پول‌شویی، روش‌های یادگیری ماشین به‌طور موفقیت‌آمیزی به کار گرفته‌شده‌اند؛ اما ظهور انواع جدیدی از تقلب‌ها با رشد سریع بازارهای الکترونیکی، باعث محبوبیت روش‌های یادگیری عمیق شده است که منجر به ارائه روش‌های نوآورانه‌ای در تشخیص ناهنجاری‌ها گردیده است [۲۰]. به‌ویژه، روش شبکه عصبی مقادیر کرانگین تعمیم‌یافته[4] (GEV-NN) که از توزیع گامبل به‌عنوان تابع فعال‌ساز استفاده می‌کند، در زمینه داده‌های نامتعادل نتایج پیشرفته‌ای را به دست آورده است [۱۸].

ازآنجاکه ماهیت مجموعه داده‌های مربوط به ریزش مشتریان در حوزه‌های مختلف مانند بانک، خرید پوشاک، بیمه، خدمات آنلاین و ... نامتعادل است، بنابراین برای متعادل‌سازی داده‌ها از روش‌های بازنمونه‌گیری، کم‌نمونه‌گیری و روش ترکیبی بازنمونه‌گیری استفاده‌شده است و عملکرد الگوریتم‌های یادگیری ماشین در دو حالت بدون نمونه‌گیری و بازنمونه‌گیری موردبررسی و ارزیابی قرارگرفته است. مطالب این مقاله در هفت بخش تنظیم‌شده است.

در بخش اول به‌مرور مفاهیم پایه در خصوص ریزش مشتری و الگوپردازی پرداخته می‌شود. در بخش دوم، داده‌های نامتعادل و روش‌های متعادل‌سازی داده‌ها معرفی می‌شود. در بخش سوم، روش‌شناسی و الگوریتم‌های رده‌بندی مربوط به ریزش مشتری معرفی می‌شود. بخش چهارم، روش‌شناسی الگوپردازی مشتریان معرفی‌شده است. بخش‌های پنجم، ششم و هفتم به ترتیب به معرفی داده‌ها، نتایج مدل‌بندی ریزش مشتری و الگوپردازی و در نهایت نتیجه‌گیری و پیشنهادها پرداخته‌شده است.

۱.۱ کلیات

تحلیل ریزش مشتری [۲] به معنای بررسی احتمال ترک یک محصول یا خدمت، توسط مشتریان است. به بیان ساده‌تر، این مفهوم بدان معناست که مشتریان به دلیل انتخاب رقبای شرکت، شرکت را ترک می‌کنند [۱۹]. هدف از این تحلیل، شناسایی این وضعیت پیش از ریزش مشتری (ترک محصول یا خدمات، توسط مشتری) و سپس انجام اقدامات پیشگیرانه است.

مشتریان ریزشی را می‌توان به دو گروه اصلی ریزش اختیاری و ریزش غیراختیاری تقسیم کرد [۹]. ریزش غیراختیاری به‌راحتی قابل‌شناسایی است؛ زیرا شامل مشتریانی می‌شود که دریافت خدماتشان توسط شرکت قطع‌شده است. دلایل متعددی از جمله سوءاستفاده مشتری از خدمات یا عدم پرداخت هزینه‌ها توسط مشتری

[4] Generalized Extreme Value Neural Network

وجود دارد که ممکن است شرکت، ارائه خدمات به مشتری را لغو کند. اما ریزش اختیاری پیچیده‌تر است، زیرا در این حالت مشتری با تصمیم آگاهانه، دریافت خدمات ارائه‌شده را قطع می‌کند. این نوع ریزش به دو دسته اصلی تقسیم می‌شود: ریزش اتفاقی و ریزش عمدی.

۱. ریزش اتفاقی زمانی رخ می‌دهد که تغییراتی در شرایط مشتری رخ می‌دهد که نیاز او به خدمات ارائه‌شده را از بین ببرد. برای مثال، ممکن است مشتری به دلیل تغییرات مالی، دیگر توانایی پرداخت هزینه خدمات را نداشته باشد یا به منطقه‌ای نقل‌مکان کند که در آنجا خدمات شرکت در دسترس نیست. بااین‌حال، ریزش اتفاقی معمولاً درصد کمی از ریزش اختیاری مشتریان را شامل می‌شود.

۲. ریزش عمدی مشکلی است که اکثر راه‌حل‌های مدیریت ریزش، برای مقابله با آن طراحی‌شده‌اند. این نوع ریزش زمانی رخ می‌دهد که مشتری تصمیم می‌گیرد به شرکت رقیب مراجعه کند.

الگوپردازی مشتری [۱۳] شامل تجزیه‌وتحلیل ویژگی‌های مشتری به‌منظور طراحی راهکارهای بازاریابی مناسب است که به حفظ مشتری و مدیریت ارتباط با مشتری کمک می‌کند. الگوپردازی بخش‌ها، بر شناسایی ویژگی گروه‌های خاص مشتریان تمرکز دارد تا راهکارهای بازاریابی به شکل مؤثرتری هدایت شوند. الگوپردازی رفتار خریداران، عوامل اجتماعی‌ای مانند زمان‌بندی، مزایای مورد انتظار، میزان استفاده، وفاداری و نگرش مشتریان را برای اجرای بازاریابی هدفمند موردتوجه قرار می‌دهند.

۲ داده‌های نامتعادل

در داده‌های نامتعادل، یک رده، تعداد نمونه بیشتری نسبت به رده‌های دیگر دارد. این امر منجر به توزیع نابرابر رده‌ها و بد رده‌بندی نمونه‌های اقلیت می‌شود؛ زیرا یک رده‌بند تمایل دارد که به‌صورت اریب به نفع نمونه‌های رده اکثریت عمل کند. همچنین سیستم رده‌بند تمایل دارد نمونه‌های رده اقلیت را نادیده بگیرد و آن‌ها را به‌عنوان نوفه شناسایی کند. به‌منظور مقابله با عدم تعادل داده‌ها، روش‌هایی مانند روش بازنمونه‌گیری، روش یادگیری حساس به هزینه و مدل‌های ترکیبی طراحی‌شده‌اند تا به همراه روش‌های مختلف داده‌رهنمون، داده‌های نامتعادل را مدیریت کنند. برای ارزیابی عملکرد داده‌های نامتعادل، معمولاً از معیارهای بازیابی، دقت، معیار F، میانگین هندسی و منحنی ROC [۵] استفاده می‌شود [۱۶].

۱.۲ روش‌های بازنمونه‌گیری

روش‌های بازنمونه‌گیری[۶] به سه دستهٔ کلی بیش‌نمونه‌گیری[۷]، کم‌نمونه‌گیری[۸] و ترکیبی[۹] تقسیم می‌شوند. در روش بیش‌نمونه‌گیری، نمونه‌های ردهٔ اقلیت به روش‌های مختلفی افزایش می‌یابد که این افزایش از طریق تولید نمونه‌های جدید یا تکرار نمونه‌های موجود از ردهٔ اقلیت صورت می‌گیرد. در روش کم‌نمونه‌گیری، نمونه‌های ردهٔ اکثریت به روش‌های مختلفی کاهش می‌یابد. روش ترکیبی، شامل ترکیب دو روش بیش‌نمونه‌گیری و کم‌نمونه‌گیری است؛ بدین ترتیب که ابتدا با اجرای روش بیش‌نمونه‌گیری، نمونه‌های ردهٔ اقلیت افزایش می‌یابد و سپس با اجرای روش کم‌نمونه‌گیری بر روی مجموعه داده حاصل، نمونه‌های ردهٔ اکثریت کاهش می‌یابد تا داده‌ها متعادل شوند. از جمله روش‌های بیش‌نمونه‌گیری، روش‌های تصادفی، روش بیش‌نمونه‌گیری اقلیت مصنوعی[۱۰] (SMOTE) و نمونه‌گیری مصنوعی سازوار[۱۱] (ADASYN) است. از جمله روش‌های کم‌نمونه‌گیری، روش کم‌نمونه‌گیری تصادفی، روش قاعدهٔ پاک‌سازی همسایگی[۱۲] (NCR)، روش NearMiss و پیوندهای تومک[۱۳] است. لازم به ذکر است که هر یک از روش‌های بازنمونه‌گیری تنها بر روی مجموعه آموزشی اعمال می‌شوند. در ادامه به معرفی مختصری از برخی از روش‌های بازنمونه‌گیری پرداخته می‌شود.

روش بیش‌نمونه‌گیری مصنوعی سازوار (ADASYN) [۱۱]، نمونه‌های اقلیت سازوار را بر اساس توزیع آن‌ها تولید می‌کند. در این روش برای نمونه‌های ردهٔ اقلیت که یادگیری آن‌ها سخت‌تر است، داده‌های مصنوعی بیشتری نسبت به نمونه‌های اقلیتی که یادگیری آن‌ها

[5] recursive operating characteristic curve
[6] resampling
[7] over sampling
[8] under Sampling
[9] Hybrid
[10] Synthetic Minority Oversampling Technique
[11] Adaptive Synthetic Sampling
[12] Neighborhood Cleaning Rule
[13] Tomek Links

آسان‌تر است تولید می‌شود. روش ADASYN نه تنها اریبی یادگیری را که از توزیع نامتعادل داده‌های اصلی نشئت می‌گیرد کاهش می‌دهد، بلکه به‌طور سازوار، مرز تصمیم‌گیری را تغییر می‌دهد تا تمرکز بیشتری روی نمونه‌های دشوارتر برای یادگیری وجود داشته باشد.

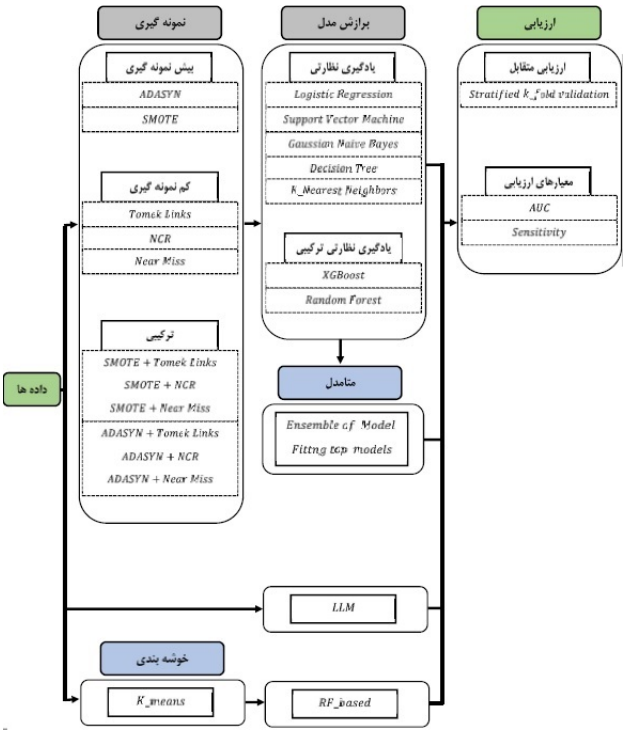
برای روش‌های کم‌نمونه‌گیری، رویکردهای پیشرفته‌ای مانند روش قاعدهٔ پاک‌سازی همسایگی (NCR) و پیوندهای تومک[۱۰] پیشنهادشده است. NCR ترکیبی از دو قاعده است که نمونه‌های اضافی و مبهم را از ردهٔ اکثریت حذف می‌کند. قاعدهٔ اول، نزدیک‌ترین همسایه چگال[۱۴] (CNN) [۲۴] است که در این قاعده زیرمجموعه‌ای از نمونه‌های ردهٔ اکثریت انتخاب می‌شوند که نمی‌توانند به‌درستی رده‌بندی شوند و این نمونه‌ها به‌عنوان نمونه‌های مهم برای یادگیری در نظر گرفته می‌شوند. قاعده دوم، نزدیک‌ترین همسایه ویرایش شده[۱۵] (ENN) است که در این قاعده، نمونه‌های مبهم با استفاده از رویکرد K-نزدیک‌ترین همسایه حذف می‌شود. اگر نمونه‌ای از ردهٔ اکثریت توسط همسایگان خود اشتباه رده‌بندی‌شده باشد، آن نمونه از مجموعه داده حذف می‌شود و اگر نمونه‌ای از ردهٔ اقلیت توسط همسایگان ردهٔ اکثریت خود اشتباه رده‌بندی شود، آن همسایگان از ردهٔ اکثریت حذف خواهند شد.

پیوندهای تومک با استفاده از قاعدهٔ CNN، جفت نمونه‌هایی از ردهٔ اکثریت و اقلیت را شناسایی می‌کند. این جفت‌ها متشکل از یک نمونه از ردهٔ اکثریت و نزدیک‌ترین همسایه‌اش از ردهٔ اقلیت است. نمونه‌هایی از ردهٔ اکثریت که در پیوندهای تومک قرار دارند، داده‌های نویز محسوب شده و باید حذف شوند.

۳ فرایند یادگیری ماشین برای پیش‌بینی و تحلیل ریزش مشتری

هدف اصلی این مقاله، ارزیابی چندین روش یادگیری ماشین برای پیش‌بینی ریزش مشتری است. شکل ۱ فرایند یادگیری ماشین برای پیش‌بینی و تحلیل ریزش مشتری را شرح داده است. با توجه به اینکه، داده‌های ریزش عمدتاً نامتعادل هستند، برای مقابله با مسئله عدم تعادل داده‌ها، هر روش یادگیری با رویکردهای مختلف بازنمونه‌گیری برای متعادل کردن توزیع رده‌ها مقایسه شده است و نشان داده‌شده است که با متعادل کردن داده‌ها، عملکرد روش‌های رده‌بندی بالا می‌رود. با توجه به پیچیدگی علل ریزش، برای بهبود پیش‌بینی ریزش، عملکرد رویکردهای یادگیری ترکیبی[۱۶] موردبررسی قرارگرفته است. در نهایت، قبل از پیاده‌سازی یک رویکرد یادگیری ماشین ترکیبی (گروهی)، به‌منظور دسته‌بندی مشتریان، از یک روش خوشه‌بندی و غیرنظارتی استفاده‌شده است.

در این مقاله، ابتدا چندین روش مختلف در فرایند پیش‌بینی ریزش مشتری با استفاده از یادگیری ماشین ارزیابی می‌شود که شامل سه مرحله بازنمونه‌گیری، برازش مدل و یک فرایند ارزیابی است (شکل ۱ ؛ بخش‌های سبز و خاکستری). برای حل مشکل عدم تعادل داده‌ها [۱]، هر روش یادگیری با روش‌های بازنمونه‌گیری ترکیب‌شده است تا توزیع رده‌ها متعادل شود، چراکه این موضوع تأثیر قابل‌توجهی در عملکرد رده‌بندهای استاندارد دارد. بر اساس نتایج این آزمایش‌ها، بهترین روش‌های یادگیری ماشین را شناسایی کرده و یک روش ترکیبی ارائه‌شده است که می‌تواند با موفقیت بر روی مجموعه داده‌های مشابه ریزش مشتری اعمال شود. در نهایت برای الگوپردازی ریزش مشتریان، از رویکرد خوشه‌بندی بر روی مجموعه داده مرجع بانک استفاده‌شده است (شکل ۱ ؛ بخش‌های سبز و آبی).



شکل ۱. فرایند یادگیری ماشین برای پیش‌بینی و تحلیل ریزش مشتری

روش‌های یادگیری ماشین متعددی در حوزهٔ ریزش مشتری استفاده

[14] Condensed Nearest Neighbor
[15] Edited Nearest Neighbors
[16] ensemble learning

می‌شوند. این روش‌ها شامل رویکردهای نظارتی و غیرنظارتی است که به پیش‌بینی ریزش مشتری یا الگوپردازی رفتار مشتری می‌پردازد. در این مقاله، عملکرد هفت روش یادگیری ماشین نظارتی شامل دو رویکرد ترکیبی باهم مقایسه می‌شوند. مدل‌های به کار گرفته‌شده عبارت‌اند از رده‌بند بیز ساده[۱۷] (Gnb)، رگرسیون لوژستیک[۱۸] (LR)، K-نزدیک‌ترین همسایه[۱۹] (K-NN)، ماشین بردار پشتیبان[۲۰] (SVM) با دو هسته شعاعی (SVMR) و چندجمله‌ای (SVMP) و بدون هسته (SVML) و درخت تصمیم[۲۱] (DT). علاوه بر این، استفاده از رویکردهای ترکیبی مانند جنگل تصادفی، روش تقویت سازوار [۲۵]، تقویت گرادیان [۱۷] یا تقویت گرادیان شدید [۲۲] (XGBoost،[۷]) برای پیش‌بینی ریزش مشتری مطرح‌شده است. همچنین رویکردهای یادگیری عمیق [۱۷، ۷] و روش‌های نیمه‌نظارتی نیز در این زمینه به‌کاررفته است. در این مقاله از بسیاری از روش‌های نام‌برده استفاده‌شده است. درخت تصمیم و رگرسیون لوژستیک، دو الگوریتم بسیار محبوب برای پیش‌بینی ریزش مشتری هستند که هم عملکرد پیش‌بینی و هم قابلیت تفسیر آن‌ها خوب است. باوجوداین نقاط قوت، درخت‌های تصمیم معمولاً در تشخیص روابط خطی بین متغیرها دچار مشکل می‌شوند و رگرسیون لوژستیک نیز در شناسایی اثر متقابل بین متغیرها ضعف‌هایی دارد؛ بنابراین، یک الگوریتم ترکیبی جدید به نام مدل برگ لوژستیک[۲۳] (LLM) [۴] پیشنهادشده است تا داده‌ها را بهتر رده‌بندی کند. در ادامه این روش معرفی می‌شود.

۱.۳ مدل برگ لوژستیک

اخیراً، چندین مطالعه بر روی مدل‌های پیش‌بینی ریزش مشتری انجام‌شده است تا بتواند تعادل خوبی بین عملکرد پیش‌بینی و تفسیرپذیری نتایج از دیدگاه الگوپردازی مشتریان برقرار کنند. به‌عنوان‌مثال، دکگینی و همکاران [۴]، مدل LLM را طراحی کردند که از دو مرحلهٔ بخش‌بندی داده‌ها و به دنبال آن پیشگویی است. در روش LLM، بخش‌بندی بر اساس اِفرازی است که در برگ‌های درخت تصمیم به دست می‌آید که از برچسب ریزش از داده‌های ورودی استفاده می‌کند. در این مرحله از درخت تصمیم استفاده می‌شود تا بخش‌های همگن مشتریان را شناسایی کند. سپس برای هر زیرمجموعه داده، یک مدل رگرسیون لوژستیک برازش داده می‌شود که قابلیت پیشگویی و تفسیر را فراهم می‌آورد. روش LLM همچنین شامل مرحلهٔ کاهش تصادفی نمونه‌های ردهٔ اکثریت و انتخاب ویژگی است.

درخت تصمیم یکی از مدل‌های پیش‌بینی است که عمدتاً به دلیل سادگی، هم از نظر کارایی و هم از نظر قابلیت تفسیر، عملکرد خوبی دارد. در فرایند این مدل، داده‌ها به‌صورت بازگشتی به زیرمجموعه‌های کوچک‌تر و خالص‌تر تقسیم می‌شوند، به‌طوری‌که از یک جستجوی حریصانه برای یافتن شاخه‌های ممکن در فضای درخت تصمیم استفاده کرده و تقسیم‌بندی بهینه را بر اساس یک معیار تقسیم انتخاب می‌کند. این فرایند از گره ریشه شروع می‌شود که گره‌ای بدون گره والد است و مکرر معیارهای تقسیم‌بندی بهینه را تعیین می‌کند تا داده‌ها را به دو گره فرزند تقسیم کند. این فرایند زمانی پایان می‌یابد که تقسیم‌بندی جدید دیگر مطلوب یا ممکن نباشد. در این صورت مجموعه‌ای از گره‌ها بدون گره فرزند به نام گره پایانی یا برگ به دست می‌آید.

به‌این‌ترتیب، توسط ساختار درخت که شامل مجموعه‌ای از برگ‌ها یا گره‌های پایانی (T) است، مجموعه مشتریان (S) که توسط تمامی ویژگی‌ها پوشش داده‌شده‌اند، به زیرمجموعه‌های غیر هم‌پوشان (S_t) تقسیم می‌شوند، به‌طوری‌که هر زیرمجموعه توسط یک برگ (t) در درخت نمایش داده می‌شود:

$$S = \bigcup_{t \in T} S_t; \quad \forall t \neq t' : S_t \cap S_{t'} = \emptyset$$

تقسیم‌بندی‌های مکرر، به‌تدریج مدل‌های پیچیده‌تری تولید می‌کند که می‌تواند منجر به بیش‌برازش شود.

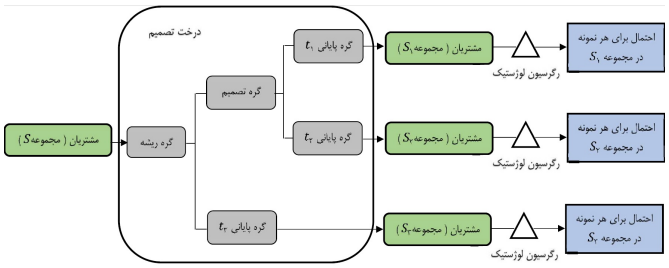
این پیچیدگی نامطلوب، از طریق تعیین چندین ابرپارامتر که فرایند تقسیم‌بندی را کنترل می‌کنند و یا با استفاده از روش هرس کردن پس از اجرای الگوریتم قابل‌کنترل است. هرس، یک روش کاهش پیچیدگی است که بخش‌هایی از درخت را که توانایی کافی برای رده‌بندی مشتریان ندارند، حذف می‌کند. پیچیدگی درخت همچنین تحت تأثیر پارامترهای مختلفی از الگوریتم درخت تصمیم مانند پارامتر حداقل اندازه برگ که حداقل تعداد مشاهدات در گره‌ها را کنترل می‌کند، قرار دارد که این پارامتر یک معیار توقف برای فرایند تقسیم‌بندی است.

در درخت تصمیم، برای یک مشاهده، بر اساس تمایل رده در گره

[17] Naive Bayes
[18] Logistic Regression
[19] K-Nearest Neighbors
[20] Support Vector Machine
[21] Decision Tree
[22] Extreme Gradient Boosting
[23] Logit Leaf Model

پایانی‌ای که آن مشاهده در آن قرار دارد، پیش‌بینی صورت می‌گیرد؛ اما در الگوریتم LLM، در مرحله دوم، یک رگرسیون لوژستیک با انتخاب ویژگی به روش پیشرو در هر گره پایانی برازش می‌شود.

رگرسیون لوژستیک یک روش پیش‌بینی مستقل است که به‌طور گسترده در بازاریابی استفاده می‌شود. در رگرسیون لوژستیک، احتمال‌های پسین به‌طور مستقیم برآورد می‌شوند که این ویژگی آن را بسیار قابل تفسیرتر از روش‌های پیچیده‌تر و «جعبه سیاه» می‌کند. در الگوریتم LLM، مدل‌های رگرسیون لوژستیک از انتخاب ویژگی پیشرو استفاده می‌کنند. شکل ۲، یک نمایش مفهومی از LLM است که جریان داده‌ها را نشان می‌دهد. در این شکل، مجموعهٔ مشتریان (S) توسط درخت تصمیم به سه زیرمجموعهٔ S_1، S_2 و S_3 تقسیم می‌شوند. سپس برای هر زیرمجموعه به‌طور جداگانه یک رگرسیون لوژستیک برازش داده می‌شود و احتمال برای هر نمونه در این زیرمجموعه‌ها مشخص می‌شود.

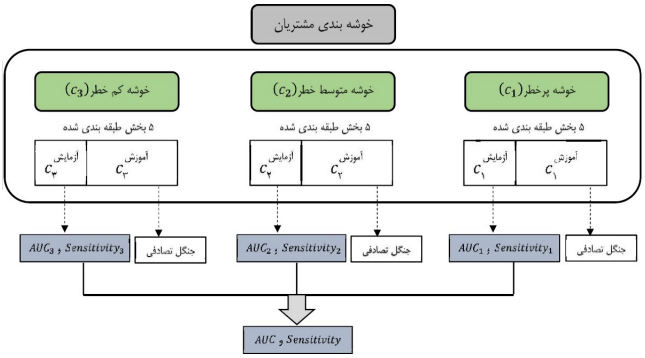


شکل ۲. نمایش مفهومی مدل برگ لوژستیک

۲.۳ روش مبتنی بر جنگل تصادفی

پس از طراحی روش LLM، یولاه و همکاران [۲۳] مدلی برای پیشگویی ریزش مشتری با استفاده از روش جنگل‌های تصادفی [۲۴] طراحی کردند که هدف آن ارائه هم‌زمان تفسیرپذیری و بهبود پیشگویی است. آن‌ها برای ریزش مشتریان از روش K – میانگین استفاده کرده و داده‌ها را از نظر ریزش به سه گروه کم‌خطر، متوسط خطر و پُرخطر تقسیم کرده‌اند. در روش مبتنی بر جنگل تصادفی، ابتدا مشاهدات با استفاده الگوریتم K – میانگین به سه خوشه تقسیم می‌شوند. سپس، در هر خوشه، یک مدل جنگل تصادفی اجرا می‌گردد و عملکرد هرکدام از این مدل‌ها در هر خوشه با استفاده از معیارهای AUC و Sensitivity مورد ارزیابی قرار می‌گیرد.

در نهایت میانگین مقادیر AUC و Sensitivity در سه خوشه به‌عنوان نتایج نهایی مدل مبتنی بر جنگل تصادفی در نظر گرفته می‌شود (شکل ۳). مدل مبتنی بر جنگل تصادفی ارائه‌شده توسط یولاه و همکاران [۲۳]، همانند مدل LLM از انتخاب ویژگی‌ها استفاده می‌کند.



شکل ۳. مدل ترکیبی مبتنی بر جنگل تصادفی لوژستیک

۳.۳ الگوریتم تقویت گرادیان شدید

الگوریتم تقویت گرادیان شدید (XGBoost) برای اولین بار توسط تیانکی چن و کارلوس گاسترین در سال ۲۰۱۱ معرفی شد و در مطالعات بعدی توسط بسیاری از دانشمندان به‌طور مستمر بهبود یافت. این الگوریتم یک روش یادگیری بر پایهٔ مدل‌های درخت تقویت‌شده است. در مدل‌های سنتی درخت تقویت‌شده، تنها از اطلاعات مشتق مرتبهٔ اول استفاده می‌شود. در این مدل‌ها، هنگام آموزش درخت nام، به دلیل استفاده از باقیمانده $n-1$ درخت قبلی، اجرای آموزش توزیع‌شده دشوار است. در مقابل، XGBoost با استفاده از بسط تیلور مرتبه دوم بر روی تابع زیان، به‌طور خودکار از پردازندهٔ چندهسته‌ای (CPU) برای محاسبات موازی استفاده می‌کند. علاوه بر این، XGBoost از روش‌های مختلفی برای جلوگیری از بیش‌برازش نیز استفاده می‌کند.

۴.۳ رویکرد فراترکیبی یادگیری ماشین

در این بخش، به‌عنوان نمونه به ترکیب دو یا سه مدل از سه رویکرد رگرسیون لوژستیک (LR)، تقویت گرادیان شدید (XGB) و جنگل تصادفی (RF) پرداخته‌شده است. الگوریتم‌هایی که ترکیب می‌شوند، می‌تواند رویکردهای برتر در مثال‌های واقعی و کاربردی باشند. در این رویکرد، با استفاده از روش رأی‌دهی نرم، پیش‌بینی‌های دو یا سه مدل مذکور باهم ترکیب می‌شوند. در روش رأی‌دهی نرم، پیش‌بینی نهایی برای هر مشاهده، میانگین احتمالات پیش‌بینی‌شده توسط دو یا سه مدل از این سه رویکرد یا به عبارتی مجموع وزنی از پیش‌بینی مدل‌های پایه (LR,RF,XGB) است. این وزن‌ها بر اساس میزان شباهت (همبستگی) پیش‌بینی‌های مدل‌ها به یکدیگر محاسبه می‌شوند.

[24] RF-based

به بیان دقیق‌تر وزن هر مدل بر اساس میزان شباهت پیش‌بینی‌های آن مدل با پیش‌بینی‌های سایر مدل‌ها تعیین می‌شود. این شباهت‌ها با استفاده از همبستگی پیرسون محاسبه می‌شود. به این معنی که اگر دو مدل، پیش‌بینی‌های مشابه‌تری داشته باشند، همبستگی بین آن‌ها بالا خواهد بود و در نتیجه وزن بیشتری به آن‌ها داده می‌شود و اگر پیش‌بینی‌های مدل‌ها متفاوت باشند، وزن کمتری به آن‌ها تعلق می‌گیرد.

برای مثال، سه مدل $M^۱, M^۲$ و $M^۳$ را در نظر بگیرید. در روش رأی‌دهی نرم [۸]، امتیاز پیش‌بینی‌شده $\hat{y}_{ens}$ به‌صورت مجموع وزنی از امتیازات پیش‌بینی‌شده توسط هر مدل محاسبه می‌شود. وزن هر مدل در این ترکیب با توجه به همبستگی آن مدل با سایر مدل‌ها تعیین می‌گردد؛ بدین ترتیب که وزن هر مدل، معکوس همبستگی آن مدل با سایر مدل‌ها است. برای نرمال‌سازی وزن‌ها، از تابع softmax استفاده می‌شود تا مجموع وزن‌ها برابر با یک شود. این کار موجب می‌شود وزن‌ها به‌طور متعادل بین مدل‌ها توزیع شوند و از این طریق اطمینان حاصل می‌شود که مدل‌ها به‌طور منصفانه در پیش‌بینی نهایی مشارکت دارند. فرمول پیش‌بینی نهایی به‌صورت زیر است:

$$\hat{y}_{ens} = \omega_۱ \hat{y}_{M^۱} + \omega_۲ \hat{y}_{M^۲} + \omega_۳ \hat{y}_{M^۳}$$

که در آن

$$\omega_۱, \omega_۲, \omega_۳ = \text{softmax}(\widetilde{\omega_۱}, \widetilde{\omega_۲}, \widetilde{\omega_۳}),$$

$$\widetilde{\omega_k} = \frac{۱}{\rho(\hat{Y}_{M^k}, \hat{Y}_{M^\alpha}) + \rho(\hat{Y}_{M^k}, \hat{Y}_{M^\beta})}.$$

در این فرمول، $\omega_۱, \omega_۲, \omega_۳$ وزن‌های محاسبه‌شده برای هر مدل و $\hat{y}_{M^۱}, \hat{y}_{M^۲}, \hat{y}_{M^۳}$ پیش‌بینی مدل‌های $M^۱, M^۲, M^۳$ هستند. ρ نشان‌دهنده همبستگی پیرسون است که برای محاسبه شباهت (همبستگی) بین پیش‌بینی‌های هر مدل با سایر مدل‌ها به کار می‌رود و α و β مدل‌هایی غیر از مدل k هستند $(\alpha \neq k, \beta \neq k)$.

۴ روش‌های یادگیری ماشین برای الگوپردازی ریزش

داده‌های ریزش مشتری معمولاً ساختار پیچیده‌ای دارد که نشان‌دهندهٔ عدم تعادل بین رده‌ها و همچنین تقسیم‌بندی ذاتی داده‌ها به دلیل تعدد الگوهای رفتار مشتریان است.

هدف از الگوپردازی مشتریان، تفکیک آن‌ها به گروه‌های مشخص با رفتارهای مشابه و شناسایی ویژگی‌های مهم و تأثیرگذار در ریزش مشتریان بانک است. خوشه‌بندی، مجموعه‌ای گسترده از تکنیک‌ها برای یافتن زیرگروه‌ها یا خوشه‌ها در یک مجموعه داده است. در فرایند خوشه‌بندی مشاهدات، هدف این است که داده‌ها را بر اساس ویژگی‌ها به گروه‌های جداگانه‌ای تقسیم کنیم به‌طوری‌که مشاهدات در هر گروه شباهت زیادی به یکدیگر داشته باشند، درحالی‌که مشاهدات در گروه‌های مختلف با یکدیگر دارای تفاوت قابل‌توجهی باشند. ازآنجاکه خوشه‌بندی در بسیاری از حوزه‌ها موردتوجه است، روش‌های متعددی برای انجام خوشه‌بندی وجود دارند. یکی از آن‌ها که در این مقاله مورداستفاده قرارگرفته است، خوشه‌بندی K-میانگین است. در خوشه‌بندی K-میانگین، هدف تقسیم‌بندی مشاهدات به تعداد مشخص و از پیش تعیین‌شده خوشه‌ها است.

باید به یاد داشته باشیم که الگوریتم استاندارد K- میانگین، فرض می‌کند که خوشه‌ها در یک مدل آمیختهٔ گاوسی کروی یکنواخت با نسبت‌های برابر قرار دارند؛ بنابراین، زمانی که خوشه‌ها به‌راحتی از هم جدا نمی‌شوند، باید از فرضیات استاندارد K- میانگین، فاصله گرفت و از مدل‌های جدیدی استفاده کرد که به غیرخطی بودن ساختار داده‌ها توجه داشته باشد. روش‌های خوشه‌بندی جدیدی برای پردازش داده‌هایی که فرضیات ضعیفی درباره شکل خوشه‌ها دارند و نیازمند فیلتر کردن ویژگی‌های نامربوط هستند، پیشنهادشده است که در آن از خودرمزگذارهای عمیق (DAE) [۱۲] استفاده می‌کنند. خودرمزگذارهای عمیق می‌توانند نمایش مناسب‌تری از خوشه‌بندی داده‌ها (یا کدگذاری) را به شیوه‌ای غیرنظارتی ایجاد کند درحالی‌که به‌طور خودکار ویژگی‌های مهم را یاد می‌گیرد. این نوع شبکه عصبی خودنظارتی به‌گونه‌ای آموزش می‌بیند که ورودی خود را در خروجی بازتولید کند و درعین‌حال یک تابع هزینه را بهینه‌سازی می‌کند. در ادامه روش خوشه‌بندی K-میانگین معرفی می‌شود.

۱.۴ خوشه‌بندی K-میانگین

خوشه‌بندی K-میانگین[۲۵] روشی ساده برای تقسیم یک مجموعه داده به K خوشهٔ مجزا و بدون هم‌پوشانی است. برای اجرای خوشه‌بندی K-میانگین، ابتدا باید تعداد خوشه‌های موردنظر (K) مشخص شود؛ سپس الگوریتم K-میانگین هر مشاهده را به‌طور دقیق به یکی از K خوشه اختصاص می‌دهد. روش خوشه‌بندی K-میانگین از یک مسئله ریاضی ساده و شهودی به دست می‌آید. فرض کنید $C_۱, \ldots, C_K$ مجموعه‌هایی باشند که نمایانگر شاخص‌های مربوط به مشاهدات هر خوشه هستند.

[25] k-means clustering

این مجموعه‌ها دو ویژگی مهم دارند:

۱. $C_1 \cup C_2 \cup \cdots \cup C_K = \{1, \ldots, n\}$. به‌عبارت‌دیگر، هر مشاهده حداقل به یکی از K خوشه تعلق دارد.

۲. $\forall k \neq k'; C_k \cap C_{k'} = \emptyset$ به‌عبارت‌دیگر، خوشه‌ها هم‌پوشانی ندارند و هیچ مشاهده‌ای به بیش از یک خوشه تعلق ندارد.

برای مثال، اگر مشاهده iام در خوشه kام قرار داشته باشد، آنگاه $i \in C_k$ است. ایده اصلی خوشه‌بندی K-میانگین این است که یک خوشه‌بندی در صورتی مطلوب است که در آن تغییرات درون خوشه‌ای به حداقل برسد. تغییرات درون خوشه‌ای برای خوشه C_k، اندازه $W(C_k)$ است که میزان تفاوت مشاهدات درون یک خوشه را از یکدیگر نشان می‌دهد؛ بنابراین، هدف حل مسئله زیر است:

$$\min_{C_1, \ldots, C_K} \left\{ \sum_{k=1}^{K} W(C_k) \right\} \qquad (1)$$

به‌عبارت‌دیگر، این فرمول بیان می‌کند که هدف، تقسیم مشاهدات به K خوشه به‌گونه‌ای است که مجموع تغییرات درون خوشه‌ای برای K خوشه به حداقل برسد. حل معادله (۱) یک ایده منطقی به نظر می‌رسد، اما برای اجرایی کردن آن، نیاز به تعریف تغییرات درون خوشه‌ای است. روش‌های مختلفی برای تعریف این مفهوم وجود دارد اما رایج‌ترین آن‌ها، استفاده از توان دوم فاصله اقلیدسی است؛ به‌عبارت‌دیگر $W(C_k)$ به‌صورت زیر تعریف می‌شود:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \qquad (2)$$

که $|C_k|$ تعداد مشاهدات در خوشه kام را نشان می‌دهد؛ به‌عبارت‌دیگر، تغییرات درون خوشه‌ای برای خوشه kام برابر با مجموع توان دوم فاصله اقلیدسی هر جفت از مشاهدات موجود در خوشه kام است که بر تعداد کل مشاهدات در خوشه kام تقسیم می‌شود. از ترکیب معادلات (۱) و (۲)، مسئله بهینه‌سازی زیر به دست می‌آید که خوشه‌بندی K-میانگین را تعریف می‌کند:

$$\min_{C_1, \ldots, C_K} \left\{ \sum_{k=1} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\} \qquad (3)$$

اکنون هدف یافتن الگوریتمی برای حل معادله (۳) است؛ یعنی یافتن روشی برای تقسیم مشاهدات به K خوشه به‌گونه‌ای که معادله (۳) به حداقل برسد. این مسئله در واقع یک مسئله بسیار مشکل است زیرا تقریباً K^n راه برای تقسیم n مشاهده به K خوشه وجود دارد. این تعداد بسیار بزرگ است مگر اینکه K و n بسیار کوچک باشند! خوشبختانه یک الگوریتم بسیار ساده وجود دارد که می‌تواند یک بهینه‌سازی موضعی یعنی یک راه‌حل نسبتاً خوب برای مشکل بهینه‌سازی K-میانگین در معادله (۳) ارائه دهد. این رویکرد، خوشه‌بندی K-میانگین است که در زیر شرح داده‌شده است. به‌تصادف یک شماره از ۱ تا K به هر یک از مشاهدات اختصاص داده می‌شود. این شماره‌ها به‌عنوان تخصیص اولیه خوشه‌ها برای مشاهدات در نظر گرفته می‌شوند.
مراحل زیر تا زمانی تکرار می‌شوند که تخصیص خوشه‌ها تغییر نکند:

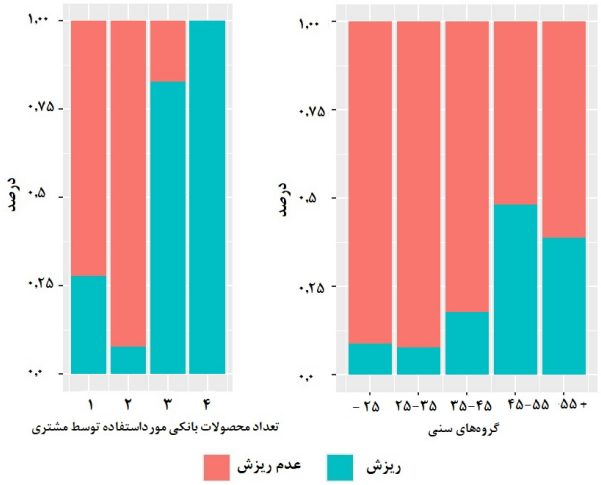
الف) برای هر یک از K خوشه، مرکز خوشه محاسبه می‌شود. مرکز خوشه kام، همان بردار میانگین p ویژگی برای مشاهدات موجود در خوشه kام است.

ب) هر مشاهده به خوشه‌ای که به مرکز آن نزدیک‌تر است اختصاص داده می‌شود (فاصلهٔ هر مشاهده تا مرکز خوشه به کمک فاصلهٔ اقلیدسی محاسبه می‌گردد)

۵ معرفی مجموعه داده

در این مقاله، مجموعه داده با نام Bank که از طریق لینک https://www.kaggle.com در دسترس است، مورداستفاده قرارگرفته است. این مجموعه داده شامل ده هزار نمونه است که حدود ۲۰۰۰ نمونه در رده ریزش و حدود ۸۰۰۰ نمونه در رده غیر ریزش قرار دارند. نسبت ریزش به غیرریزش در این مجموعه داده برابر ۰٫۲۵ است. متغیر هدف، متغیر دودویی است که نشان می‌دهد آیا مشتری، حساب خود را بسته است یا همچنان مشتری بانک است. این مجموعه داده شامل ۱۴ متغیر شامل شماره ردیف، شناسه و نام خانوادگی مشتری، امتیاز اعتباری مشتری، کشور محل سکونت مشتری، جنسیت و سن مشتری، تعداد سال‌هایی که مشتری در بانک بوده است، موجودی بانکی مشتری، تعداد محصولات بانکی که مشتری استفاده می‌کند، داشتن یا نداشتن کارت اعتباری، عضویت فعال یا غیرفعال بودن مشتری، حقوق برآوردی مشتری به دلار و بستن یا نبستن حساب بانکی توسط مشتری است که نشان‌دهندهٔ ترک یا وفاداری مشتری به بانک است.

شکل ۴. مصورسازی نرخ ریزش به تفکیک جنس و کشور



شکل ۵. مصورسازی نرخ ریزش به تفکیک گروه‌های سنی و تعداد محصولات بانکی که مشتری استفاده می‌کند

در این داده‌ها، متوسط سن مشتریان حدود ۴۰ سال است. متوسط سال‌هایی که مشتریان در بانک بوده‌اند، حدود ۵ سال است و نیمی از مشتریان، تنها از یک محصول بانک و حدود ۴۶ درصد مشتریان از دو محصول بانک استفاده کرده‌اند. همچنین، ۲۰ درصد مشتریان، حساب بانکی خود را بسته‌اند. محل سکونت ۵۰ درصد مشتریان، فرانسه و محل سکونت ۲۵ درصد مشتریان در آلمان و ۲۵ درصد دیگر در اسپانیا است. ۵۵ درصد مشتریان مرد هستند و حدود ۷۰ درصد مشتریان دارای کارت اعتباری هستند. عضویت ۵۲ درصد مشتریان، فعال است. بر اساس نتایج شکل‌های ۴ و ۵، ریزش مشتریان آلمانی دو برابر بیشتر از مشتریان فرانسوی و اسپانیایی است. همچنین ریزش مشتریان زن حدود ۲۵ درصد و بیشتر از مردان است. میزان ریزش مستری با افزایش سن تا سن ۵۵ سالگی افزایش و بعدازآن کاهشی می‌شود. در نهایت، مشتریانی که از ۳ یا ۴ محصول استفاده می‌کنند باید به‌دقت غربال شوند زیرا نرخ ریزش برای این دو گروه به ترتیب ۱۰۰ و ۸۲ درصد است.

۶ نتایج مدل‌بندی ریزش مشتریان

در این بخش نتایج مدل‌بندی ریزش مشتریان با استفاده از الگوریتم‌های یادگیری ماشین تکی و ترکیبی و همچنین روش‌های مختلف تکی بدون نمونه‌گیری و بازنمونه‌گیری تکی و ترکیبی ارائه‌شده است. با توجه به اینکه مجموعه داده‌های ریزش مشتری، نامتعادل است، در تمامی نتایج ارائه‌شده در این مقاله، روش اعتبارسنجی متقابل طبقه‌بندی‌شده با در نظر گرفتن $K = ۵$ زیرنمونه به‌کاررفته است؛ چراکه در این روش نسبت توزیع دو رده در بخش‌ها تقریباً محفوظ می‌ماند و بنابراین نتایج قابل‌اعتمادتری به دست خواهد آمد.

جدول ۱. معیارهای AUC و Sensitivity برای ارزیابی الگوریتم‌های یادگیری ماشین در حالت بدون متعادل کردن داده‌ها

الگوریتم‌های یادگیری ماشین نظارتی	بدون نمونه‌گیری	
الگوریتم‌های ساده	AUC	Sensitivity
LR	۰٫۸۳۰	۰٫۳۷۸
SVML	۰٫۶۴۹	۰٫۰۴۹
SVMR	۰٫۷۴۵	۰٫۰۵۳
SVMP	۰٫۵۵۹	۰٫۳۰۸
DT	۰٫۷۵۷	۰٫۴۴۶
Gnb	۰٫۸۱۴	۰٫۲۲۱
$K-nn; K=۲$	۰٫۴۹۲	۰٫۲۲۵
$K-nn; K=۵$	۰٫۴۸۱	۰٫۰۸۷
الگوریتم‌های ترکیبی	AUC	Sensitivity
XGBoost	۰٫۸۳۹	۰٫۴۷۳
RF	۰٫۸۴۷	۰٫۴۴۱

نتایج جدول ۱ به مقایسهٔ الگوریتم‌های یادگیری ماشین بر اساس معیار AUC و Sensitivity بدون متعادل کردن داده‌ها (بدون نمونه‌گیری) پرداخته است. با مقایسهٔ مقادیر AUC، در روش‌های یادگیری ماشین، به ترتیب جنگل‌های تصادفی (RF)، تقویت گرادیان شدید (XGBoost) و رگرسیون لوژستیک (LR) نسبت به سایر روش‌ها، بالاترین AUC و در نتیجه بهترین عملکرد را دارند. بر اساس معیار Sensitivity، به ترتیب روش‌های تقویت گرادیان شدید (XGBoost)، درخت تصمیم (DT)، جنگل‌های تصادفی (RF)، و رگرسیون لوژستیک (LR) نسبت به سایر روش‌ها، بالاترین Sensitivity و در نتیجه بهترین عملکرد را دارند.

جدول‌های ۲ و ۳ به مقایسهٔ الگوریتم‌های یادگیری ماشین به ترتیب بر اساس متعادل‌سازی داده‌ها به روش بیش‌نمونه‌گیری و کم‌نمونه‌گیری

می‌پردازد. به‌طورکلی، مقادیر AUC و Sensitivity برای جدول ۲ مربوط به روش بیش‌نمونه‌گیری بزرگ‌تر از روش کم‌نمونه‌گیری است (جدول ۳). نتایج نشان می‌دهد، در تمام روش‌های بیش‌نمونه‌گیری و کم‌نمونه‌گیری به ترتیب سه روش یادگیری ماشین جنگل‌های تصادفی (RF)، تقویت گرادیان شدید (XGBoost) و رگرسیون لوژستیک (LR)، نسبت به سایر روش‌ها، بالاترین AUC و در نتیجه بهترین عملکرد را دارند و روش یادگیری ماشین K–نزدیک‌ترین همسایه نسبت به سایر روش‌ها کمترین AUC و در نتیجه ضعیف‌ترین عملکرد را دارد. همچنین، از مقایسه مقادیر AUC سه مدل برتر می‌توان نتیجه گرفت در بین روش‌های بیش‌نمونه‌گیری، روش SMOTE و در بین روش‌های کم‌نمونه‌گیری، روش NCR بالاترین عملکرد را دارند. از طرف دیگر، در هر دو روش متعادل‌سازی کم‌نمونه‌گیری و بیش‌نمونه‌گیری، الگوریتم‌های ترکیبی تقویت گرادیان شدید و جنگل‌های تصادفی (RF) نسبت به الگوریتم ساده رگرسیون لوژستیک (LR)، بر اساس هر دو معیار ارزیابی، عملکرد بهتری دارد. جدول ۴ به مقایسه الگوریتم‌های مختلف یادگیری ماشین با متعادل کردن داده‌ها به روش ترکیب روش بیش‌نمونه‌گیری ADASYN با سه روش کم‌نمونه‌گیری Tomek، nearmiss و NCR می‌پردازد. این در حالی است که در جدول ۵ به مقایسهٔ الگوریتم‌های مختلف یادگیری ماشین با متعادل کردن داده‌ها به روش ترکیب روش بیش‌نمونه‌گیری SMOTE با سه روش کم‌نمونه‌گیری Tomek، nearmiss و NCR پرداخته‌شده است.

نتایج جدول ۴ و ۵ نشان می‌دهند:

- در تمام روش‌های بازنمونه‌گیری ترکیبی (ترکیب روش بیش‌نمونه‌گیری ADASYN و SMOTE با سه روش کم‌نمونه‌گیری Tomek، nearmiss و NCR) به ترتیب سه روش یادگیری ماشین جنگل‌های تصادفی (RF)، تقویت گرادیان شدید و رگرسیون لوژستیک (LR) نسبت به سایر روش‌ها، بالاترین AUC و در نتیجه بهترین عملکرد را دارند و روش یادگیری ماشین K–نزدیک‌ترین همسایه نسبت به سایر روش‌ها کمترین AUC و در نتیجه ضعیف‌ترین عملکرد را داشته است.
- بر اساس معیار Sensitivity، به ترتیب سه روش یادگیری ماشین تقویت گرادیان شدید، جنگل‌های تصادفی (RF)، و رگرسیون لوژستیک (LR) نسبت به سایر روش‌ها، بهترین عملکرد را داشته است. در مجموع بر اساس هر دو معیار، روش‌های ترکیبی، عملکرد بهتری نسبت به الگوریتم‌های ساده داشته است.

از مقایسه مقادیر AUC سه مدل برتر در جدول‌های ۴ و ۵ می‌توان نتیجه گرفت:

- در بین روش‌های بازنمونه‌گیری ترکیبی (ترکیب روش بیش‌نمونه‌گیری ADASYN با سه روش کم‌نمونه‌گیری Tomek، nearmiss و NCR)، روش nearmiss-ADASYN و در بین روش‌های بازنمونه‌گیری ترکیبی (ترکیب روش بیش‌نمونه‌گیری SMOTE با سه روش کم‌نمونه‌گیری Tomek، nearmiss و NCR) روش nearmiss-SMOTE دارای بالاترین عملکرد است.
- در مقایسه چهار روش بازنمونه‌گیری برتر SMOTE، NCR، nearmiss-ADASYN و nearmiss-SMOTE، روش بیش‌نمونه‌گیری SMOTE بهترین عملکرد را دارد.

ازآنجاکه یکی از راه‌های افزایش دقت برآوردها، استفاده از روش‌های ترکیبی الگوریتم‌های یادگیری ماشین و همچنین استفاده از روش‌های ترکیبی الگوریتم‌های بازنمونه‌گیری برای متعادل‌سازی داده‌ها است، در جدول ۶ به مقایسهٔ معیارهای AUC و Sensitivity برای الگوریتم‌های ترکیبی سه مدل یادگیری ماشین LR، RF و XGBoost به‌عنوان بهترین الگوریتم‌های یادگیری ماشین در چهار حالت متعادل‌سازی داده‌ها به‌صورت بدون نمونه‌گیری، بیش‌نمونه‌گیری، کم‌نمونه‌گیری و روش‌های ترکیبی بازنمونه‌گیری پرداخته‌شده است. نتایج به‌دست‌آمده از جدول ۶ نشان می‌دهد در تمام روش‌های بدون نمونه‌گیری و بازنمونه‌گیری، ترکیب سه مدل LR، RF و XGBoost در مقایسه با ترکیب هر جفت از این سه مدل، بالاترین AUC و در نتیجه بهترین عملکرد را دارد. همچنین ترکیب سه مدل LR,RF,XGBoost با دو روش بیش‌نمونه‌گیری SMOTE و روش بازنمونه‌گیری ترکیبی nearmiss-SMOTE در مقایسه با سایر روش‌های بدون‌نمونه‌گیری و بازنمونه‌گیری بهترین عملکرد را دارند. در تمام روش‌های بدون‌نمونه‌گیری و بازنمونه‌گیری، ترکیب دو مدل LR و XGBoost در مقایسه با ترکیب هر جفت از این سه مدل، بالاترین Sensitivity و در نتیجه بهترین عملکرد را دارد. همچنین ترکیب دو مدل LR,XGBoost با دو روش بازنمونه‌گیری ترکیبی nearmiss-ADASYN و nearmiss-SMOTE بهترین عملکرد را دارند. جدول ۷ به مقایسهٔ معیار AUC و Sensitivity دو مدل ترکیبی LLM و RF-based می‌پردازد. نتایج نشان می‌دهد، بر اساس معیار AUC، مدل LLM و بر اساس معیار Sensitivity، مدل RF-based عملکرد بهتری دارد.

با توجه به جداول ۶ و ۷ از مقایسه دو مدل ترکیبی LLM و RF-based با مدل ترکیبی سه مدل برتر (مدل پیشنهادی) می‌توان نتیجه

گرفت مدل ترکیبی سه مدل برتر در تمام حالات (ترکیب دوتایی و سه‌تایی) نسبت به دو مدل LLM و RF-based، دارای AUC بالاتر و در نتیجه دارای عملکرد بهتر است.

در یک جمع‌بندی نهایی از نتایج کل جداول می‌توان گفت در مقایسه تمام مدل‌های برتر ساده و ترکیبی ذکرشده در روش‌های بدون‌نمونه‌گیری و بازنمونه‌گیری، ترکیب سه مدل برتر LR, RF, XGBoost با روش بیش‌نمونه‌گیری SMOTE و روش بازنمونه‌گیری ترکیبی nearmiss-SMOTE با مقدار ۰٫۸۸۶ برای AUC بهترین مدل است.

۱.۶ نتایج الگوپردازی مشتریان بر اساس خوشه‌بندی

در این بخش به ارائه نتایج مدل‌بندی الگوپردازی مشتریان پرداخته‌شده است. در این مقاله به‌منظور الگوپردازی ریزش مشتریان بانک، از الگوریتم خوشه‌بندی K-میانگین استفاده‌شده است. هدف از خوشه‌بندی مشتریان، تفکیک آن‌ها به گروه‌های مشخص با رفتارهای مشابه و شناسایی ویژگی‌های مهم و تأثیرگذار در ریزش مشتریان بانک است. قبل از اجرای الگوریتم K-میانگین، ابتدا داده‌ها استانداردسازی شدند تا تمامی ویژگی‌ها در مقیاس یکسانی قرار گیرند و تأثیر ویژگی‌هایی با مقیاس‌های مختلف بر خوشه‌بندی کاهش یابد. پس از استانداردسازی، داده‌ها به سه خوشه تقسیم شدند و نرخ ریزش در هر خوشه محاسبه گردید. بر اساس این نرخ، خوشه‌ها به سه سطح پُرخطر، متوسط خطر و کم‌خطر از نظر ریزش دسته‌بندی شدند:

- نرخ ریزش بالای ٪۵۰: پُرخطر (خوشه ۱)
- نرخ ریزش بین ٪۲۰ تا ٪۵۰: متوسط خطر (خوشه ۲)
- نرخ ریزش کمتر از ٪۲۰: کم‌خطر (خوشه ۳)

با شناسایی ویژگی‌های مشتریان در هر گروه، می‌توان برنامه‌های بازاریابی مؤثرتر و مناسب‌تری طراحی کرد که به حفظ مشتریان و افزایش تعامل آن‌ها کمک می‌کند. همچنین این تحلیل به بانک‌ها کمک می‌کند مشتریانی با سطح ریسک بالا (خوشه پُرخطر) را شناسایی کنند و با تدوین برنامه‌های هدفمند، احتمال ریزش آن‌ها را کاهش دهند. جدول ۸ به تحلیل توصیفی خوشه‌بندی مشتریان بانک در سه خوشهٔ معرفی‌شده می‌پردازد. نتایج جدول ۸ نشان می‌دهد که حدود ۳ درصد از مشتریان در خوشهٔ پُرخطر، حدود ۵۷ درصد در خوشه متوسط خطر و حدود ۴۰ درصد در خوشه کم‌خطر هستند. نتایج نشان می‌دهد که ۸۶ درصد مشتریان پُرخطر، ۲۷ درصد مشتریان متوسط خطر و ۶ درصد مشتریان کم‌خطر، ریزش دارند.

جدول ۸. تحلیل توصیفی خوشه‌بندی مشتریان

	خوشه	تعداد	درصد	نرخ ریزش
بدون نمونه‌گیری	۱	۳۲۶	۰٫۰۳	۰٫۸۶
	۲	۵۷۲۶	۰٫۵۷	۰٫۲۷
	۳	۳۹۴۸	۰٫۴۰	۰٫۰۶

پس از خوشه‌بندی مشتریان به سه خوشه پُرخطر، متوسط خطر و کم‌خطر، به‌منظور بررسی دقیق تأثیر هر ویژگی بر ریزش مشتریان و استخراج الگوی مشخص برای ریزش، در هر خوشه به‌طور جداگانه یک مدل رگرسیون لوژستیک اجرا و نتایج آن در جدول ۹ ارائه‌شده است. نتایج ویژگی‌های رسته‌ای مشتریان در این سه خوشه به شرح زیر است:

- در خوشه پُرخطر:

چهار ویژگی سن مشتری، موجودی بانکی مشتری، نوع عضویت و کشور محل سکونت مشتری تأثیر معناداری در ریزش مشتری دارند. افزایش سن و موجودی بانکی، احتمال ریزش مشتریان را افزایش می‌دهد. احتمال ریزش مشتریان با عضویت فعال، کمتر از مشتریانی است که عضویت غیرفعال دارند و احتمال ریزش مشتریان آلمانی کمتر از مشتریان فرانسوی است. نتایج نشان می‌دهد با فرض ثابت بودن سایر ویژگی‌ها، به ازای یک سال افزایش سن، بخت ریزش مشتریان ۱٫۱۰۳ ≈ $e^{۰٫۰۹۸}$ برابر می‌شود؛ یعنی افزایش سن، احتمال ریزش مشتریان را افزایش می‌دهد. همچنین، به ازای یک واحد افزایش در موجودی بانکی، بخت ریزش مشتریان ۱٫۰۳۴ ≈ $e^{۰٫۰۳۴}$ برابر می‌شود؛ یعنی افزایش موجودی بانکی، احتمال ریزش مشتریان را افزایش می‌دهد. بخت ریزش مشتریانی که عضویت فعال دارند، ۰٫۳۹۷ ≈ $e^{-۰٫۹۲۰}$ برابر بخت ریزش مشتریانی است که عضویت غیرفعال دارند. به‌عبارت‌دیگر، احتمال ریزش مشتریانی که عضویت فعال دارند، کمتر از مشتریانی است که عضویت غیرفعال دارند. همچنین، بخت ریزش مشتریان آلمانی، ۰٫۱۱۵ ≈ $e^{-۲٫۱۶۰}$ برابر بخت ریزش مشتریان فرانسوی (سطح‌مبنا) است؛ یعنی احتمال ریزش مشتریان آلمانی کمتر از مشتریان فرانسوی است.

- در خوشه متوسط خطر:

هفت ویژگی امتیاز اعتباری مشتری، سن و موجودی بانکی مشتری، نوع عضویت و کشور محل سکونت مشتری، جنسیت مشتری و تعداد محصولات بانکی که مشتری استفاده می‌کند، تأثیر معناداری در ریزش مشتری دارند. افزایش سن، احتمال ریزش مشتریان را افزایش می‌دهد اما افزایش امتیاز اعتباری

و موجودی بانکی احتمال ریزش مشتریان را کاهش می‌دهد. از طرفی احتمال ریزش مشتریان با عضویت فعال، کمتر از مشتریانی است که عضویت غیرفعال دارند و احتمال ریزش مشتریان مرد کمتر از مشتریان زن است. همچنین احتمال ریزش مشتریانی که از دو محصول بانکی استفاده می‌کنند، کمتر از مشتریانی است که از یک محصول بانکی استفاده می‌کنند و احتمال ریزش مشتریان آلمانی بیشتر از مشتریان فرانسوی است.

- در خوشه کم‌خطر:

در خوشه کم‌خطر، پنج ویژگی سن مشتری، نوع عضویت و کشور محل سکونت مشتری، جنسیت مشتری و تعداد محصولات بانکی که مشتری استفاده می‌کند، تأثیر معناداری در ریزش مشتری دارند. نتایج نشان می‌دهد با فرض ثابت بودن سایر ویژگی‌ها، به ازای یک سال افزایش سن، بخت ریزش مشتریان ۱٫۰۷ $\approx e^{۰٫۰۶۸}$ برابر می‌شود. همچنین بخت ریزش مشتریانی که عضویت فعال دارند، ۰٫۲۶۹ $\approx e^{-۱٫۳۱۲}$ برابر بخت ریزش مشتریانی است که عضویت غیرفعال دارند؛ به‌عبارت‌دیگر احتمال ریزش مشتریانی که عضویت فعال دارند، کمتر از مشتریانی است که عضویت غیرفعال دارند. بخت ریزش مشتریان اسپانیایی ۳٫۹۰۶ $\approx e^{۱٫۳۶۲}$ برابر بخت ریزش مشتریان فرانسوی (سطح‌مبنا) است. بخت ریزش مشتریان مرد ۰٫۵۰۹ $\approx e^{-۰٫۶۷۶}$ برابر بخت ریزش مشتریان زن است. همچنین بخت ریزش مشتریانی که از دو محصول بانکی استفاده می‌کنند، ۰٫۲۸۰ $\approx e^{-۱٫۲۷۸}$ برابر بخت ریزش مشتریانی است که از یک محصول بانکی استفاده می‌کنند (سطح‌مبنا). لازم به ذکر است که در خوشه کم‌خطر فقط ۲ نفر مشتری آلمانی حضور داشتند که حذف گردیدند.

۷ نتیجه‌گیری و پیشنهادها

در این مطالعه، چندین روش رایج یادگیری ماشین برای پیش‌بینی و الگوپردازی ریزش مشتری بررسی، ارزیابی و مقایسه شده‌اند. رویکرد دیگری که در زمینه پیش‌بینی ریزش باید موردتوجه قرارگیرد، روش‌های یادگیری عمیق است. در صنعت مالی، به‌ویژه، در زمینه تشخیص جرائم اقتصادی مانند تقلب مالی و پول‌شویی، روش‌های یادگیری ماشین به‌طور موفقیت‌آمیزی به‌کارگرفته‌شده‌اند. روش‌های سنتی مانند رگرسیون لوژستیک، بیز ساده و ماشین بردار پشتیبان از پرکاربردترین روش‌ها در این زمینه هستند. بااین‌حال، ظهور انواع جدیدی از تقلب‌ها با رشد سریع بازارهای الکترونیکی، باعث محبوبیت روش‌های یادگیری عمیق شده است که منجر به ارائه روش‌های نوآورانه‌ای در تشخیص ناهنجاری‌ها گردیده است. همچنین، باید توجه داشت که اکثر چارچوب‌های پیش‌بینی ریزش مشتری، معمولاً فقط داده‌های ساختاریافته را در نظر می‌گیرند درحالی‌که بخش بزرگی از داده‌های موجود در دنیای امروز شامل داده‌های متنی و غیرساختاریافته است، بنابراین باید روش‌هایی برای استفاده از اطلاعات این نوع داده‌ها نیز ارائه شود. به‌ویژه، با گسترش روزافزون ارتباطات آنلاین بین مشتریان و شرکت‌ها یا بانک‌ها، این موضوع بسیار حائز اهمیت است. بهره‌گیری از ویژگی‌های شبکه‌های اجتماعی – نظیر نظرات، اشتراک‌گذاری دوستان – نیز می‌تواند با کشف اطلاعات علی، پیش‌بینی ریزش را بهبود بخشد. چراکه تأثیرات اجتماعی یکی از دلایل اصلی رفتار ریزش مشتریان به شمار می‌رود. این مباحث می‌تواند موضوع یک مطالعه جذاب و کوتاه‌مدت در آینده باشد.

جدول ۲. معیارهای AUC و Sensitivity برای ارزیابی الگوریتم‌های یادگیری ماشین به روش بیش‌نمونه‌گیری

الگوریتم‌های یادگیری ماشین نظارتی	ADASYN		SMOTE	
الگوریتم‌های ساده	AUC	Sensitivity	AUC	Sensitivity
LR	۰٫۸۶۴	۰٫۶۱۲	۰٫۸۶۶	۰٫۶۱۸
SVML	۰٫۶۷۹	۰٫۳۰۶	۰٫۶۲۵	۰٫۳۸۴
SVMR	۰٫۸۰۴	۰٫۳۱۵	۰٫۸۰۷	۰٫۳۱۹
SVMP	۰٫۵۸۶	۰٫۴۰۰	۰٫۵۸۲	۰٫۴۱۰
DT	۰٫۸۰۸	۰٫۵۳۶	۰٫۷۹۰	۰٫۵۲۸
Gnb	۰٫۸۵۰	۰٫۴۱۷	۰٫۸۵۲	۰٫۴۲۹
$K-nn; K=۲$	۰٫۴۶۶	۰٫۳۷۱	۰٫۴۶۴	۰٫۳۸۵
$K-nn; K=۵$	۰٫۴۵۵	۰٫۲۳۲	۰٫۴۷۱	۰٫۲۴۴
الگوریتم‌های ترکیبی	AUC	Sensitivity	AUC	Sensitivity
XGBoost	۰٫۸۷۴	۰٫۶۴۱	۰٫۸۷۷	۰٫۶۳۴
RF	۰٫۸۸۳	۰٫۶۱۵	۰٫۸۸۳	۰٫۶۱۲

جدول ۳. معیارهای AUC و Sensitivity برای ارزیابی الگوریتم‌های یادگیری ماشین به روش کم‌نمونه‌گیری

الگوریتم‌های یادگیری ماشین نظارتی	Tomek		nearmiss		NCR	
الگوریتم‌های ساده	AUC	Sensitivity	AUC	Sensitivity	AUC	Sensitivity
LR	۰٫۸۳۳	۰٫۵۱۷	۰٫۸۳۰	۰٫۵۱۵	۰٫۸۳۵	۰٫۵۱۹
SVML	۰٫۶۷۴	۰٫۲۴۶	۰٫۶۳۰	۰٫۲۵۳	۰٫۶۴۹	۰٫۱۰۳
SVMR	۰٫۷۴۴	۰٫۱۰۸	۰٫۷۴۳	۰٫۱۱۲	۰٫۷۵۱	۰٫۱۰۵
SVMP	۰٫۵۷۰	۰٫۴۳۲	۰٫۵۵۷	۰٫۴۱۲	۰٫۵۷۲	۰٫۴۲۳
DT	۰٫۷۶۳	۰٫۵۳۵	۰٫۷۴۴	۰٫۵۰۵	۰٫۷۵۲	۰٫۵۰۹
Gnb	۰٫۸۱۷	۰٫۳۰۰	۰٫۸۱۵	۰٫۲۶۰	۰٫۸۱۸	۰٫۳۰۴
$K-nn; K=۲$	۰٫۴۸۲	۰٫۳۲۹	۰٫۴۹۸	۰٫۳۲۵	۰٫۴۸۱	۰٫۳۴۲
$K-nn; K=۵$	۰٫۴۷۱	۰٫۲۰۱	۰٫۴۸۶	۰٫۱۹۵	۰٫۴۶۶	۰٫۲۲۲
الگوریتم‌های ترکیبی	AUC	Sensitivity	AUC	Sensitivity	AUC	Sensitivity
XGBoost	۰٫۸۳۹	۰٫۵۶۹	۰٫۸۳۷	۰٫۵۷۲	۰٫۸۴۱	۰٫۵۷۴
RF	۰٫۸۵۰	۰٫۵۴۲	۰٫۸۴۹	۰٫۵۵۰	۰٫۸۵۴	۰٫۵۵۲

جدول ۴. معیارهای AUC و Sensitivity برای ارزیابی الگوریتم‌های یادگیری ماشین به روش بازنمونه‌گیری ترکیبی

الگوریتم‌های یادگیری ماشین نظارتی	Tomek-ADASYN		nearmiss-ADASYN		NCR-ADASYN	
الگوریتم‌های ساده	AUC	Sensitivity	AUC	Sensitivity	AUC	Sensitivity
LR	۰٫۸۵۷	۰٫۵۹۰	۰٫۸۶۴	۰٫۶۲۰	۰٫۸۵۴	۰٫۵۹۰
SVML	۰٫۶۶۸	۰٫۳۰۶	۰٫۶۷۹	۰٫۱۹۴	۰٫۶۶۲	۰٫۲۷۰
SVMR	۰٫۷۸۸	۰٫۲۵۸	۰٫۸۰۳	۰٫۳۱۵	۰٫۷۸۶	۰٫۲۴۵
SVMP	۰٫۵۷۹	۰٫۳۹۹	۰٫۵۵۸	۰٫۴۳۸	۰٫۸۵۳	۰٫۵۳۷
DT	۰٫۷۸۲	۰٫۵۱۹	۰٫۷۸۳	۰٫۵۲۳	۰٫۷۶۹	۰٫۵۱۹
Gnb	۰٫۸۴۳	۰٫۳۶۸	۰٫۸۵۱	۰٫۴۳۰	۰٫۸۴۰	۰٫۳۶۸
$K-nn; K=۲$	۰٫۴۶۲	۰٫۳۷۷	۰٫۴۶۶	۰٫۳۷۸	۰٫۴۶۰	۰٫۳۷۹
$K-nn; K=۵$	۰٫۴۴۴	۰٫۲۳۰	۰٫۴۷۵	۰٫۲۳۴	۰٫۴۵۴	۰٫۲۴۵
الگوریتم‌های ترکیبی	AUC	Sensitivity	AUC	Sensitivity	AUC	Sensitivity
XGBoost	۰٫۸۶۵	۰٫۶۱۴	۰٫۸۷۵	۰٫۶۴۲	۰٫۸۶۳	۰٫۶۲۳
RF	۰٫۸۷۵	۰٫۵۹۶	۰٫۸۸۳	۰٫۶۱۹	۰٫۸۷۳	۰٫۵۹۲

جدول ۵. معیارهای AUC و Sensitivity برای ارزیابی الگوریتم‌های یادگیری ماشین به روش بازنمونه‌گیری ترکیبی (ترکیب روش بیش‌نمونه‌گیری SMOTE با سه روش کم‌نمونه‌گیری

الگوریتم‌های یادگیری ماشین نظارتی	Tomek-SMOTE		nearmiss-SMOTE		NCR-SMOTE	
الگوریتم‌های ساده	AUC	Sensitivity	AUC	Sensitivity	AUC	Sensitivity
LR	۰٫۸۵۸	۰٫۵۹۸	۰٫۸۶۶	۰٫۶۲۱	۰٫۸۵۶	۰٫۵۸۹
SVML	۰٫۶۲۸	۰٫۱۰۲	۰٫۶۱۷	۰٫۱۹۹	۰٫۶۱۴	۰٫۱۶۱
SVMR	۰٫۷۹۱	۰٫۲۵۷	۰٫۸۰۸	۰٫۳۲۵	۰٫۷۸۳	۰٫۲۴۵
SVMP	۰٫۶۰۲	۰٫۴۹۱	۰٫۵۳	۰٫۴۴۶	۰٫۵۷۳	۰٫۴۳۵
DT	۰٫۸۱۰	۰٫۵۱۳	۰٫۷۸۲	۰٫۵۰۹	۰٫۷۸۶	۰٫۵۲۴
Gnb	۰٫۸۴۴	۰٫۳۸۷	۰٫۸۵۳	۰٫۴۳۵	۰٫۸۴۱	۰٫۳۷۰
$K-nn; K=۲$	۰٫۴۶۱	۰٫۳۹۳	۰٫۴۶۵	۰٫۳۸۷	۰٫۴۵۸	۰٫۳۸۲
$K-nn; K=۵$	۰٫۴۷۱	۰٫۲۴۶	۰٫۴۷۱	۰٫۲۴۹	۰٫۴۴۲	۰٫۲۴۹
الگوریتم‌های ترکیبی	AUC	Sensitivity	AUC	Sensitivity	AUC	Sensitivity
XGBoost	۰٫۸۶۸	۰٫۶۲۱	۰٫۸۷۶	۰٫۶۴۷	۰٫۸۶۵	۰٫۶۲۳
RF	۰٫۸۷۷	۰٫۵۹۸	۰٫۸۸۴	۰٫۶۲۲	۰٫۸۷۴	۰٫۵۹۵

جدول ۶. مقایسهٔ معیارهای AUC و Sensitivity مدل ترکیبی سه مدل برتر (مدل فراترکیبی پیشنهادی)، بدون و با متعادل کردن داده‌ها و از طریق اعتبارسنجی متقابل به روش طبقه‌بندی با $K = ۵$ زیرنمونه

		مدل LR,RF,XGBoost		مدل LR,RF		مدل LR,XGBoost		مدل RF,XGBoost	
		AUC	Sensitivity	AUC	Sensitivity	AUC	Sensitivity	AUC	Sensitivity
بدون نمونه‌گیری		۰٫۸۵۱	۰٫۴۳۵	۰٫۸۴۹	۰٫۴۱۰	۰٫۸۴۶	۰٫۴۳۳	۰٫۸۴۸	۰٫۴۵۸
بیش‌نمونه‌گیری	ADASYN	۰٫۸۸۵	۰٫۶۳۱	۰٫۸۸۲	۰٫۶۱۷	۰٫۸۸۱	۰٫۶۳۴	۰٫۸۸۳	۰٫۶۳۵
	SMOTE	۰٫۸۸۶	۰٫۶۲۳	۰٫۸۸۳	۰٫۶۱۷	۰٫۸۸۲	۰٫۶۳۰	۰٫۸۸۴	۰٫۶۲۶
کم‌نمونه‌گیری	Tomek	۰٫۸۵۳	۰٫۵۴۹	۰٫۸۵۱	۰٫۵۳۱	۰٫۸۴۹	۰٫۵۵۳	۰٫۸۴۹	۰٫۵۵۷
	nearmiss	۰٫۸۵۱	۰٫۵۵۳	۰٫۸۵۰	۰٫۵۳۴	۰٫۸۴۶	۰٫۵۵۵	۰٫۸۴۸	۰٫۵۶۵
	NCR	۰٫۸۵۶	۰٫۵۵۳	۰٫۸۵۴	۰٫۵۳۶	۰٫۸۵۲	۰٫۵۵۲	۰٫۸۵۲	۰٫۵۶۶
بازنمونه‌گیری ترکیبی	Tomek-ADASYN	۰٫۸۷۷	۰٫۶۰۹	۰٫۸۷۵	۰٫۵۹۷	۰٫۸۷۳	۰٫۶۰۹	۰٫۸۷۵	۰٫۶۱۲
	nearmiss-ADASYN	۰٫۸۸۴	۰٫۶۳۵	۰٫۸۸۲	۰٫۶۲۶	۰٫۸۸۱	۰٫۶۴۱	۰٫۸۸۳	۰٫۶۳۵
	NCR-ADASYN	۰٫۸۷۵	۰٫۶۰۶	۰٫۸۷۲	۰٫۵۹۵	۰٫۸۷۱	۰٫۶۱۲	۰٫۸۷۲	۰٫۶۱۴
	Tomek-SMOTE	۰٫۸۷۸	۰٫۶۱۰	۰٫۸۷۵	۰٫۵۹۷	۰٫۸۷۵	۰٫۶۱۱	۰٫۸۷۶	۰٫۶۱۵
	nearmiss-SMOTE	۰٫۸۸۶	۰٫۶۳۷	۰٫۸۸۳	۰٫۶۲۳	۰٫۸۸۲	۰٫۶۴۱	۰٫۸۸۴	۰٫۶۳۹
	NCR-SMOTE	۰٫۸۷۶	۰٫۶۰۶	۰٫۸۷۴	۰٫۵۹۴	۰٫۸۷۲	۰٫۶۱۶	۰٫۸۷۳	۰٫۶۱۵

جدول ۷. معیارهای AUC و Sensitivity برای مقایسه مدل ترکیبی LLM و مدل مبتنی بر جنگل‌های تصادفی در روش بدون نمونه‌گیری

		مدل ترکیبی LLM(De Caigny et al)	مدل ترکیبی RF-based(Ullah et al)
بدون نمونه‌گیری	AUC	۰٫۸۴۲ ± ۰٫۰۰۵	۰٫۸۰۴ ± ۰٫۰۱۵
	Sensitivity	۰٫۴۵۷ ± ۰٫۰۲۷	۰٫۴۸۹ ± ۰٫۴۶۹

جدول ۹. آمار توصیفی پروفایل ریزش مشتریان (بدون نمونه‌گیری)، به تفکیک مشتریان مربوط به هر خوشه برای ویژگی‌های کیفی مشتریان

	خوشه پُرخطر		خوشه متوسط خطر		خوشه کم‌خطر	
متغیرها	برآورد ضرایب	انحراف معیار	برآورد ضرایب	انحراف معیار	برآورد ضرایب	انحراف معیار
عدد ثابت	−۰٫۲۸۸	۱٫۷۹۴	−۱٫۳۹۶	۰٫۲۹۹	−۳٫۶۸۰	۰٫۶۱۴
امتیاز اعتباری مشتری	−۰٫۰۰۴	۰٫۰۰۲	−۰٫۰۰۱	۰٫۰۰۰	−۰٫۰۰۱	۰٫۰۰۱
سن مشتری	۰٫۰۹۸	۰٫۰۲۳	۰٫۰۷۰	۰٫۰۰۳	۰٫۰۶۸	۰٫۰۰۷
تعداد سال‌هایی که مشتری در بانک بوده است	−۰٫۰۱۶	۰٫۰۶۸	−۰٫۰۱۶	۰٫۰۱۱	−۰٫۰۳۶	۰٫۰۲۵
موجودی بانکی مشتری به دلار	۰٫۰۳۴	۰٫۰۰۸	−۰٫۰۱۲	۰٫۰۰۱	−۰٫۰۰۱	۰٫۰۰۲
کارت اعتباری	−۰٫۴۹۵	۰٫۴۴۴	−۰٫۰۲۹	۰٫۰۷۳	−۰٫۲۱۸	۰٫۱۵۷
عضویت فعال مشتری	−۰٫۹۲۰	۰٫۳۹۵	−۱٫۰۳۹	۰٫۰۷۰	−۱٫۳۱۲	۰٫۱۶۵
حقوق برآوردی مشتری به دلار	۰٫۰۰۷	۰٫۰۰۴	۰٫۰۰۰	۰٫۰۰۱	−۰٫۰۰۰	۰٫۰۰۱
کشور محل سکونت مشتری: آلمان	−۲٫۱۶۰	۰٫۹۵۸	۰٫۶۱۷	۰٫۰۷۷	–	–
کشور محل سکونت مشتری: اسپانیا	۰٫۳۰۶	۰٫۴۸۴	−۰٫۱۲۴	۰٫۱۰۵	۱٫۳۶۲	۰٫۱۵۹
جنسیت مشتری: مرد	−۰٫۶۲۴	۰٫۳۸۳	−۰٫۴۸۲	۰٫۰۶۷	−۰٫۶۷۶	۰٫۱۴۸
تعداد محصولات بانکی: ۲	–	–	−۰٫۹۶۷	۰٫۰۹۲	−۱٫۲۷۸	۰٫۱۵۲

مراجع

[1] Błaszczyński, J., and Stefanowski, J. (2018). Local data characteristics in learning classifiers from imbalanced data. *Advances in Data Analysis with Computational Intelligence Methods: Dedicated to Professor Jacek Żurada*, Springer, 51-85.

[2] Çelik, O., and Osmanoglu, U. O. (2019). Comparing to techniques used in customer churn analysis. *Journal of Multi-disciplinary Developments*, **4(1)**, 30-38.

[3] Chan, C. C. H. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. *Expert systems with applications*, **34(4)**, 2754-2762.

[4] De Caigny, A., Coussement, K., and De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, **269(2)**, 760-772.

[5] Garvin, D. A. (1988). *Managing Quality: The Strategic and Competitive Edge.* New York: London, Free Press.

[6] Geiler, L., Affeldt, S., and Nadif, M. (2022). An effective strategy for churn prediction and customer profiling. *Data and Knowledge Engineering*, **142**, 102100.

[7] Gregory, B. (2018). Predicting customer churn: Extreme gradient boosting with temporal data. *arXiv preprint*, arXiv:1802.03396.

[8] Guo, C., and Berkhahn, F. (2016). Entity embeddings of categorical variables. *arXiv preprint*, arXiv:1604.06737.

[9] Hadden, J., Tiwari, A., Roy, R., and Ruta, D. (2008). Churn prediction: Does technology matter?. *International Journal of Industrial and Manufacturing Engineering*, **2(4)**, 524-536.

[10] Hart, P. (1968). The condensed nearest neighbor rule (corresp.). *IEEE transactions on information theory*, **14(3)**, 515-516.

[11] He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, 1322-1328.

[12] Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, **313(5786)**, 504-507.

[13] Kasem, M. S., Hamada, M., and Taj-Eddin, I. (2024). Customer profiling, segmentation, and sales prediction using AI in direct marketing. *Neural Computing and Applications*, **36(9)**, 4995-5005.

[14] Kathanassopoulos, A. D. (2000). Customer satisfaction cues to support market segmentation and explain switching behavior. *Journal of business research*, **47(3)**, 191-207.

[15] Laroche, M., Rosenblatt, J. A., and Manning, T. (1986). Services used and factors considered important in selecting a bank: an investigation across diverse demographic segments. *International Journal of bank marketing*, **4(1)**, 35-55.

[16] Mathew, R. M., and Gunasundari, R. (2021, March). A review on handling multiclass imbalanced data classification in education domain. *In 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 752-755.

[17] Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., and Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on neural networks*, **11(3)**, 690-696.

[18] Munkhdalai, L., Munkhdalai, T., and Ryu, K. H. (2020). GEV-NN: A deep neural network architecture for class imbalance problem in binary classification. *Knowledge-Based Systems*, **194**, 105534.

[19] Nettleton, D. (2014). *Commercial Data Mining: Processing, Analysis and Modeling for Predictive Analytics Projects*, Morgan Kaufmann Publishers-Elsevier, Boston.

[20] Pang, G., Shen, C., Cao, L., and Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, **54(2)**, 1-38.

[21] Reichheld, F. F. and Sasser W. E. (1990). Zero defections: quality comes to services. *Harvard business review*. **68(5)**, 105-11.

[22] Reinartz, W. J. and Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. *Journal of marketing*. **67 (1)**, 77-99.

[23] Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., and Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE access*, **7**, 60134-60149.

[24] Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *in IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-2(3)**, 408-421.

[25] Xie, Y., and Li, X. (2008). Churn prediction with linear discriminant boosting algorithm. *International conference on machine learning and cybernetics*, Kunming, China, 228-233.

An effective hybrid method for churn prediction and customer profiling

Ladan Faridi[1] and Zahra Rezaei Ghahroodi[2]

Abstract:

Customer churn is one of the major economic concerns of many companies, including banks, and banks have focused their attention on customer retention, because the cost of attracting a new customer is much higher than the cost of keeping a customer. Customer churn prediction and profiling are two major economic concerns for many companies. Different learning approaches have been proposed; however, a priori choice of the most suitable model to perform both tasks remains non-trivial as it is highly dependent on the intrinsic characteristics of the churn data. Our study compares several machine learning methods with several resampling approaches for data balancing of a public bank data set. Our evaluations, reported in terms of area under the curve (AUC) and sensitivity, explore the influence of rebalancing strategies and difference machine learning methods. This work identifies the most appropriate methods in an attrition context and an effective pipeline based on an ensemble approach and clustering. Our strategy can enlighten marketing or human resources services on the behavioral patterns of customers and their attrition probability.

Keywords: Customer churn, Customer profiling, Machine learning methods, Ensemble approach, Bank customers, Area under the curve (AUC) criteria, Data balancing.

[1] School of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran.
[2] School of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran.