Journal of Business Management

Print ISSN: 2008-5907
Online ISSN: 2423-5091

Predicting Bank Customer Value Using Machine Learning: A Novel Approach Based on Adaptive Synthetic Sampling and Feature Importance

Ahmad Jafarnjad*
*Corresponding Author, Prof., Department of Industrial Management, Faculty of Industrial Management and Technology, College of Management, University of Tehran, Tehran, Iran. E-mail: jafarnjd@ut.ac.ir

Arman Rezasoltani
Ph.D. Candidate, Department of Industrial Management, Faculty of Industrial Management and Technology, College of Management, University of Tehran, Tehran, Iran. E-mail: armanrezasoltani@ut.ac.ir

Amir Mohammad Khani
Ph.D. Candidate, Department of Industrial Management, Faculty of Industrial Management and Technology, College of Management, University of Tehran, Tehran, Iran. E-mail: amir.mo.khani@ut.ac.ir

Seyedeh Hoda Hosseinian
Ph.D. Candidate, Department of Industrial Management, Kish International Campus, University of Tehran, Tehran, Iran. E-mail: hoda.hosseinian@ut.ac.ir

Abstract

Objective

Accurate prediction of customer value in the banking industry is one of the fundamental challenges that can contribute to optimal decision-making in customer management and resource allocation. This study aims to develop a comprehensive approach for predicting the value of banking customers. The primary focus of this research is on addressing the challenge of imbalanced data, improving the performance of machine learning models, and selecting key features that are effective in predicting customer value for real-world applications in banking environments.

Citation: Jafarnjad, Ahmad; Rezasoltani, Arman; Khani, Amir Mohammad & Hosseinian, Seyedeh Hoda (2026). Predicting Bank Customer Value Using Machine Learning: A Novel Approach Based on Adaptive Synthetic Sampling and Feature Importance. *Journal of Business Management,* 18(3), 1039 – 1066. https://doi.org/10.22059/jibm.2025.387468.4891 (*in Persian*)

Journal of Business Management, 2026, Vol. 18, No.3, pp. 1039-1066
Published by University of Tehran, College of Management
https://doi.org/10.22059/jibm.2025.387468.4891
Article Type: Research Paper

Received: December 23, 2024
Revised: March 24, 2025
Accepted: April 13, 2025
Published online: September 06, 2026

Methodology

This study used a dataset from a bank comprising 2,000 customers and 14 features related to customer transactions and activities. Data preprocessing was first performed, followed by feature selection, data imbalance handling, and the application of the Adaptive Synthetic Sampling (ADASYN) technique. Correlation analysis and feature importance based on the Random Forest algorithm were subsequently employed to refine the feature selection process. Through this procedure, highly correlated variables were identified, and the final set of relevant features was selected. Subsequently, 11 machine learning algorithms, including CatBoost, XGBoost, Random Forest, LightGBM, and both linear and nonlinear models, were applied to predict customer value. To improve model performance, Optuna was used for hyperparameter tuning, and five-fold cross-validation was employed to ensure robust model evaluation. Model performance was assessed using four evaluation metrics: accuracy, precision, recall, and the F1-score.

Findings

The results showed that ensemble learning-based algorithms provided the best performance in predicting customer value. The CatBoost model, with an F1 Score of 0.9324 and an accuracy of 0.909, was identified as the best-performing model. This model achieved a proper balance between precision and recall, with a precision of 0.9677 and a recall of 0.8998 in predicting valuable customers. The XGBoost and Random Forest models also demonstrated similar performance to CatBoost, with F1 Scores of 0.9322 and 0.932, respectively. The use of a combined approach for feature selection and the application of the ADASYN method for data balancing played a significant role in improving the performance of these models.

Conclusion

These results show that a different approach to data preprocessing with the help of the ADASYN algorithm in combination with modern machine learning methods can positively affect the effectiveness of models predicting customer value. The selection of correlated variables and the feature importance based on Random Forest was important in improving the general performance of the models. This revolution allowed strengthening the work of models through the elimination of features and information that had less impact on the final decisions, making the latter more precise. Based on the results of its evaluation, it can be concluded that ensemble learning models, including CatBoost, XGBoost, and Random Forest, are the most appropriate for banking settings because of their efficiency and effectiveness in dealing with large-scale, complex, and imbalanced datasets. Thus, this study extends previous research by addressing the challenges of imbalanced data and feature selection to enhance customer value prediction and customer management in the banking sector, thereby contributing to the development of an effective approach to this problem. The findings provide useful insights for identifying high-value bank customers and for developing more effective customer retention and service strategies.

Keywords: Machine learning, Customer value prediction, Data imbalance, ADASYN, CatBoost.

مدیریت بازرگانی
شاپا چاپی: ۵۹۰۷-۲۰۰۸
شاپا الکترونیکی: ۵۰۹۱-۲۴۲۳

پیش‌بینی ارزش مشتریان بانکی با یادگیری ماشین: رویکرد نوین مبتنی بر نمونه‌گیری مصنوعی تطبیقی و اهمیت ویژگی‌ها

احمد جعفرنژاد *
* نویسنده مسئول، استاد، گروه مدیریت صنعتی، دانشکدهٔ مدیریت صنعتی و فناوری، دانشکدگان مدیریت، دانشگاه تهران، تهران، ایران. رایانامه: jafarnjd@ut.ac.ir

آرمان رضاسلطانی
دانشجوی دکتری، گروه مدیریت صنعتی، دانشکدهٔ مدیریت صنعتی و فناوری، دانشکدگان مدیریت، دانشگاه تهران، تهران، ایران. رایانامه: armanrezasoltani@ut.ac.ir

امیرمحمد خانی
دانشجوی دکتری، گروه مدیریت صنعتی، دانشکدهٔ مدیریت صنعتی و فناوری، دانشکدگان مدیریت، دانشگاه تهران، تهران، ایران. رایانامه: amir.mo.khani@ut.ac.ir

سیده هدی حسینیان
دانشجوی دکتری، گروه مدیریت صنعتی، پردیس بین‌المللی کیش، دانشگاه تهران، تهران، ایران. رایانامه: hoda.hosseinian@ut.ac.ir

چکیده

هدف: پیش‌بینی دقیق ارزش مشتریان در صنعت بانکداری، یکی از چالش‌های اساسی است که می‌تواند به تصمیم‌گیری بهینه در حوزهٔ مدیریت مشتریان و تخصیص منابع کمک کند. این تحقیق با هدف توسعهٔ یک رویکرد جامع برای پیش‌بینی ارزش مشتریان بانکی انجام شده است. تمرکز اصلی این مطالعه بر مدیریت چالش عدم تعادل داده‌ها، بهبود عملکرد مدل‌های یادگیری ماشین و انتخاب ویژگی‌های کلیدی مؤثر بر پیش‌بینی ارزش مشتریان برای کاربردهای واقعی در محیط‌های بانکی است.

روش: در این پژوهش، داده‌های مربوط به ۲هزار مشتری یک بانک، شامل ۱۴ ویژگی کلیدی مرتبط با تراکنش‌ها و رفتار مشتریان، تحلیل شد. مراحل پژوهش شامل پیش‌پردازش داده‌ها، انتخاب ویژگی‌های مهم و مدیریت عدم تعادل داده‌ها با استفاده از تکنیک ADASYN بود. انتخاب ویژگی‌های مهم با استفاده از ترکیب تحلیل هم‌بستگی و روش Feature Importance مبتنی بر الگوریتم Random Forest انجام شد. در این فرایند، ابتدا ویژگی‌هایی با هم‌بستگی بالا شناسایی شدند و در ادامه، بر اساس میزان اهمیت آن‌ها، ویژگی‌های کلیدی انتخاب شدند. سپس، ۱۱ الگوریتم یادگیری ماشین، از جمله CatBoost، XGBoost، Random Forest،

استناد: جعفرنژاد، احمد؛ رضاسلطانی، آرمان؛ خانی، امیرمحمد و حسینیان، سیده هدی (۱۴۰۵). پیش‌بینی ارزش مشتریان بانکی با یادگیری ماشین: رویکرد نوین مبتنی بر نمونه‌گیری مصنوعی تطبیقی و اهمیت ویژگی‌ها. *مدیریت بازرگانی*، ۱۸(۳)، ۱۰۳۹-۱۰۶۶.

مدیریت بازرگانی، ۱۴۰۵، دوره ۱۸، شماره ۳، صص. ۱۰۳۹-۱۰۶۶
ناشر: دانشکدگان مدیریت، دانشگاه تهران
نوع مقاله: علمی پژوهشی

تاریخ دریافت: ۱۴۰۳/۱۰/۰۳
تاریخ ویرایش: ۱۴۰۴/۰۱/۰۴
تاریخ پذیرش: ۱۴۰۴/۰۱/۲۴
تاریخ انتشار: ۱۴۰۵/۰۶/۱۵
doi: https://doi.org/10.22059/jibm.2025.387468.4891

LightGBM و سایر مدل‌های خطی و غیرخطی، برای پیش‌بینی ارزش مشتریان به‌کار گرفته شد. به‌منظور بهینه‌سازی عملکرد مدل‌ها، از چارچوب Optuna برای تنظیم خودکار هایپرپارامترها و از اعتبارسنجی متقاطع پنج برابری برای ارزیابی دقیق مدل‌ها استفاده شد. عملکرد مدل‌ها براساس ۴ شاخص ارزیابی اصلی شامل صحت (Accuracy)، دقت (Precision)، فراخوانی (Recall) و امتیاز F۱ سنجیده شد.

یافته‌ها: نتایج نشان داد که الگوریتم‌های مبتنی بر یادگیری جمعی، بهترین عملکرد را در پیش‌بینی ارزش مشتریان ارائه می‌دهند. مدل CatBoost با امتیاز F۱ برابر ۰/۹۳۲۴ و صحت ۰/۹۰۹ به‌عنوان بهترین مدل شناسایی شد. این مدل توانست تعادلی مناسب میان دقت و فراخوانی ایجاد کند؛ به‌گونه‌ای که دقت مدل در پیش‌بینی مشتریان ارزشمند به ۰/۹۶۷۷ و فراخوانی آن به ۰/۸۹۹۸ رسید. مدل‌های XGBoost و Random Forest نیز عملکردی مشابه با CatBoost داشتند و امتیاز F۱ آن‌ها به‌ترتیب ۰/۹۳۲۲ و ۰/۹۳۲ بود. استفاده از رویکرد ترکیبی برای انتخاب ویژگی‌ها و استفاده از روش ADASYN به‌منظور ایجاد تعادل در داده‌ها، نقش مهمی در بهبود عملکرد این مدل‌ها ایفا کرد.

نتیجه‌گیری: مطالعهٔ حاضر نشان داد که استفاده از رویکردهای نوین یادگیری ماشین همراه با تکنیک‌های پیش‌پردازش تطبیقی مانند ADASYN می‌تواند به‌طور چشمگیری عملکرد مدل‌های پیش‌بینی ارزش مشتری را بهبود بخشد. انتخاب دقیق ویژگی‌ها با استفاده از ترکیب تحلیل هم‌بستگی و اهمیت ویژگی‌ها مبتنی‌بر الگوریتم Random Forest در بهبود عملکرد مدل‌ها نقش مهمی داشت. این فرایند با شناسایی و حذف ویژگی‌های تکراری و کم‌اهمیت، مدل‌ها را قادر ساخت تا با تمرکز بر اطلاعات کلیدی و مؤثر، دقت پیش‌بینی را افزایش دهند. مدل‌های یادگیری جمعی مانند CatBoost، XGBoost و Random Forest به‌دلیل دقت زیاد و توانایی در مدیریت داده‌های پیچیده و نامتعادل، بهترین گزینه‌ها برای کاربرد در محیط‌های بانکی هستند. این تحقیق با رفع محدودیت‌های پژوهش‌های پیشین و ارائهٔ رویکردی جامع برای مدیریت داده‌های نامتعادل و انتخاب ویژگی‌های کلیدی، گامی مؤثر در جهت بهینه‌سازی استراتژی‌های مدیریت مشتریان در صنعت بانکداری برداشته است. نتایج به‌دست‌آمده می‌تواند به بانک‌ها کمک کند تا با شناسایی دقیق مشتریان ارزشمند، سیاست‌های بهتری برای حفظ مشتریان و تخصیص منابع تدوین کنند.

کلیدواژه‌ها: یادگیری ماشین، پیش‌بینی ارزش مشتری، نامتعادلی داده‌ها، ADASYN، CatBoost.

مقدمه

در دنیای رقابتی امروز، شناسایی ارزش مشتریان به‌عنوان یکی از اولویت‌های کلیدی در صنعت بانکداری شناخته می‌شود (خاشعی ورنامخواستی و فارسی، ۱۴۰۳؛ یعقوب‌پور، خداداد حسینی، جنتی‌فر و ثانوی فرد، ۱۴۰۳). بانک‌ها و مؤسسه‌های مالی برای بهبود مدیریت منابع و ارائه خدمات بهینه به مشتریان، به ابزارهای تحلیلی قوی نیاز دارند. پیش‌بینی ارزش مشتریان و رفتار آن‌ها، مانند احتمال ترک خدمات یا ادامه همکاری، نه‌تنها به بانک‌ها کمک می‌کند تا در بازار پررقابت باقی بمانند، بلکه به بهینه‌سازی استراتژی‌های بازاریابی، مدیریت ریسک و تخصیص منابع نیز کمک شایانی می‌کند (پروکورنکووا، گوسف، ووروبف، دوروگوش و گولین[1]، ۲۰۱۸). علاوه‌براین، ظهور یادگیری ماشین به‌عنوان ابزاری کارآمد برای تحلیل داده‌های حجیم، امکان پیش‌بینی دقیق‌تر رفتار مشتریان را فراهم آورده و به تصمیم‌گیری مؤثرتر مدیران بانکی کمک کرده است (دوروگوش، ارشوف و گولین[2]، ۲۰۱۸). با توجه به رقابت شدید در صنعت بانکداری و نیاز به حفظ مشتریان ارزشمند، شناسایی مشتریان با ارزش بالا و طراحی راه‌کارهای هدفمند برای حفظ آن‌ها از ضروریات است. به‌ویژه، در داده‌های بانکی معمولاً عدم‌توازن شدیدی میان کلاس‌ها (مانند مشتریان پرارزش و کم‌ارزش) وجود دارد که پیش‌بینی دقیق و اثربخش را چالش‌برانگیز می‌کند (ایلیاس، ضیا، منیر، لتچمونان و اون[3]، ۲۰۲۰).

با وجود پیشرفت‌های چشمگیر در این حوزه، چالش‌های متعددی همچنان در مسیر تحلیل و پیش‌بینی ارزش مشتریان وجود دارد. یکی از اصلی‌ترین چالش‌ها، نامتعادلی داده‌ها است که اغلب در داده‌های بانکی مشاهده می‌شود. در مسئلهٔ پیش‌بینی مشتریان باارزش، معمولاً تعداد مشتریانی که به‌عنوان «باارزش» طبقه‌بندی می‌شوند، بسیار کمتر از سایر گروه‌هاست. این عدم‌تعادل باعث می‌شود که مدل‌های یادگیری ماشین تمایل به شناسایی مشتریان در گروه اکثریت داشته باشند و دقت پیش‌بینی در گروه اقلیت کاهش یابد (لیوی، خوش‌گفتار، باودر و سلیا[4]، ۲۰۱۸). علاوه‌براین، اکثر مدل‌های موجود بیشتر بر افزایش دقت تمرکز دارند و کمتر به تفسیرپذیری مدل‌ها برای استفاده در محیط‌های عملیاتی و مدیریتی توجه شده است. شفافیت در پیش‌بینی و تبیین خروجی مدل برای مدیریت ریسک و تصمیم‌گیری در محیط‌های بانکی از اهمیت ویژه‌ای برخوردار است (قیصر و صفات[5]، ۲۰۲۳).

مطالعات اخیر در این حوزه نشان داده‌اند که روش‌های یادگیری ماشین، به‌ویژه مدل‌های مبتنی‌بر یادگیری جمعی، مانند XGBoost، Random Forest و CatBoost، می‌توانند به‌طور چشمگیری عملکرد پیش‌بینی رفتار مشتریان را بهبود بخشند (دیوی، سانتیکو، صفا، گوناوان و ستیاوان[6]، ۲۰۲۴؛ وو[7]، ۲۰۲۴). در مطالعهٔ دیگری، استفاده از متامدل‌های ترکیبی مانند Stacking Ensemble، کارایی مدل‌های پایه را به‌طور شایان توجهی افزایش داده است (گلال، رادی و

1. Prokhorenkova, Gusev, Vorobev, Dorogush & Gulin
2. Dorogush, Ershov & Gulin
3. Ilyas, Zia, Muneer, Letchmunan & Un
4. Leevy, Khoshgoftaar, Bauder & Seliya
5. Kaisar & Sifat
6. Diwy, Santiko, Safa, Gunawan & Setiawan
7. Vu

عارف[1]، ۲۰۲۴). پژوهش‌هایی نظیر پنگ، پنگ و لی[2] (۲۰۲۳) نیز نشان داده‌اند که استفاده از ابزارهای توضیح‌پذیری مدل‌ها، مانند SHAP، می‌تواند به شفافیت بیشتر پیش‌بینی‌ها کمک کند. با این‌حال، بسیاری از این پژوهش‌ها عمدتاً در داده‌های آزمایشگاهی انجام شده‌اند و کارایی مدل‌ها در محیط‌های واقعی و پویا کمتر بررسی شده است. با وجود این پیشرفت‌ها، شکاف‌های مهمی در تحقیقات پیشین شناسایی می‌شود. اول، بیشتر مطالعات از روش‌های ساده‌ای مانند SMOTE برای متوازن‌سازی داده‌ها استفاده کرده‌اند که گاهی به تولید نمونه‌های مصنوعی غیرواقعی منجر می‌شود. دوم، کمبود استفاده از رویکردهای جامع و ترکیبی برای انتخاب ویژگی‌های کلیدی که می‌توانند تأثیر شایان توجهی بر بهبود عملکرد مدل‌های یادگیری ماشین داشته باشند، به وضوح مشاهده می‌شود. سوم، ارزیابی مدل‌ها در داده‌های ایستا و آزمایشگاهی، به محدودیت در سنجش عملکرد آن‌ها در شرایط واقعی و پویا منجر شده است. در نهایت، بسیاری از پژوهش‌ها تنها به یک بُعد خاص از ارزش مشتری، مانند احتمال ریزش یا ارزش طول عمر، پرداخته‌اند و تحلیل جامعی از ابعاد مختلف ارزش مشتری ارائه نداده‌اند (موسوی و امیری عقدایی، ۱۳۹۹).

این پژوهش با هدف رفع شکاف‌های موجود و ارائه یک رویکرد جامع برای پیش‌بینی ارزش مشتریان بانکی انجام شده است. سؤال‌های اصلی پژوهش عبارت‌اند از:

- چگونه می‌توان از روش‌های پیشرفته نمونه‌گیری تطبیقی، مانند ADASYN، برای مدیریت نامتعادلی داده‌ها استفاده کرد؟
- کدام یک از مدل‌های یادگیری ماشین (مانند CatBoost، XGBoost و Random Forest) عملکرد بهتری در پیش‌بینی ارزش مشتریان بانکی دارد؟
- چگونه می‌توان با استفاده از یک فرایند ترکیبی شامل تحلیل هم‌بستگی و اهمیت ویژگی‌ها مبتنی بر الگوریتم جنگل تصادفی، ویژگی‌های کلیدی و مؤثر بر پیش‌بینی ارزش مشتریان را شناسایی کرد؟

این مقاله بدین شرح زیر سازمان‌دهی شده است: در بخش مقدمه، اهمیت موضوع، چالش‌های موجود و اهداف تحقیق بررسی می‌شود. در ادامه، بخش پیشینهٔ پژوهش، به بررسی ادبیات مرتبط با استفاده از مدل‌های یادگیری ماشین در بانکداری و مطالعات پیشین مرتبط می‌پردازد. در بخش روش‌شناسی، تکنیک‌های مورد استفاده شامل نمونه‌برداری تطبیقی، فرایند ترکیبی انتخاب ویژگی‌ها و مدل‌های یادگیری ماشین ارائه می‌شود. بخش نتایج، عملکرد مدل‌های پیشنهادی را بر اساس داده‌های واقعی بانکی تحلیل و ارزیابی می‌کند. در نهایت، در بخش بحث و نتیجه‌گیری، نتایج تحقیق تفسیر شده و پیشنهادهایی برای تحقیقات آینده ارائه می‌شود.

پیشینهٔ پژوهش

در سال‌های اخیر، استفاده از روش‌های یادگیری ماشین در حوزهٔ بانکداری و پیش‌بینی رفتار مشتریان رشد چشمگیری داشته است. با افزایش رقابت در صنعت بانکداری، شناسایی ارزش مشتریان و پیش‌بینی رفتار آن‌ها از جمله ترک یا

1. Galal, Rady & Aref
2. Peng, Peng & Li

ماندگاری، به یکی از موضوعات کلیدی تبدیل شده است. با این حال، نامتعادلی داده‌ها، چالشی جدی در تحلیل‌های مبتنی بر یادگیری ماشین ایجاد می‌کند؛ زیرا موارد نادر ولی حیاتی مانند ترک مشتریان، به‌دلیل کم بودن در داده‌ها به‌درستی شناسایی نمی‌شوند. پژوهش‌های پیشین نشان داده‌اند که استفاده از تکنیک‌های نمونه‌گیری تطبیقی و روش‌های پیشرفته یادگیری ماشین، می‌تواند به بهبود دقت و کارایی مدل‌ها در پیش‌بینی ارزش و رفتار مشتریان کمک کند. جدول ۱ به مرور اجمالی از پژوهش‌های کلیدی انجام‌شده در این زمینه می‌پردازد.

جدول۱. پیشینه پژوهش

نویسندگان	عنوان مقاله	نتایج	مدل‌های مورد استفاده
دیوی و همکاران، ۲۰۲۴	پیش‌بینی ترک مشتریان با استفاده از یادگیری ماشین در داده‌های نامتعادل	روش XGBoost بهترین عملکرد را در تشخیص موارد مثبت ارائه داد.	XGBoost، رگرسیون لجستیک، جنگل تصادفی، SMOTE
وو، ۲۰۲۴	تکنیک کارآمد پیش‌بینی ریزش مشتری با استفاده از یادگیری ماشین ترکیبی در بانک‌های تجاری	مدل انباشته دقت ۹۸/۷۴درصد و صحت ۹۵/۱۳درصد را به‌دست آورد که از رویکردهای تک مدل سنتی بهتر عمل کرد.	Stacked model combining KNN, XGBoost, Random Forest, and SVM
صدیقی، حقی، خان، عادل و شعیب[1]، ۲۰۲۴	استراتژی‌های مختلف مبتنی بر ML برای پیش‌بینی ریزش مشتری در بخش بانکداری	ویژگی‌های کلیدی مرتبط با چرنرها را شناسایی و بهترین مدل‌های یادگیری ماشین را برای پیش‌بینی ریزش مشخص کردند.	درخت تصمیم، جنگل تصادفی، SVM، KNN، XGBoost، رگرسیون لجستیک
اشرف[2]، ۲۰۲۴	پیش‌بینی ریزش مشتریان بانک با استفاده از چارچوب یادگیری ماشین	پیش‌بینی ریزش مشتری را با ساخت و مقایسه ده مدل طبقه‌بندی یادگیری ماشین با پنج تکنیک نمونه برداری تجزیه و تحلیل کرد.	مدل‌های مختلف یادگیری ماشین با تکنیک‌های نمونه‌گیری
تکوابو، گرگینا، تولنی، ماتا و مارتینز[3]، ۲۰۲۲	به‌سمت یادگیری ماشین قابل توضیح برای پیش‌بینی ریزش مشتریان بانکی با استفاده از روش‌های توازن داده و مبتنی بر مجموعه‌ای	بهبود عملکرد طبقه‌بندی از طریق رسیدگی به متغیرهای ناهمگن در سیستم‌های مدیریت ارتباط با مشتری.	یادگیری ماشین قابل توضیح با مدل‌سازی فازی
ترن، لی و نگوین[4]، ۲۰۲۳	پیش‌بینی ترک مشتریان در بخش بانکی با استفاده از مدل‌های طبقه‌بندی مبتنی بر یادگیری ماشین	مدل Random Forest با دقت ۹۷درصد بهترین عملکرد را ارائه داد.	جنگل تصادفی، رگرسیون لجستیک، KNN، ماشین بردار پشتیبان (SVM)
پنگ و همکاران، ۲۰۲۳	تحلیل پیش‌بینی و تفسیر مدل ترک مشتریان در بانکداری	GA-XGBoost بهترین عملکرد را با دقت ۹۹درصدAUC ارائه داد.	GA-XGBoost، SMOTE، Shapley Additive Explanations (SHAP)
گلال و همکاران، ۲۰۲۲	بهبود پیش‌بینی ترک مشتریان در بانکداری دیجیتال با استفاده از مدل‌های هیبریدی	ترکیب یادگیرنده‌های پایه و مدل‌های Voting به دقت ۸۷/۹درصد دست یافت.	رگرسیون لجستیک، جنگل تصادفی، KNN، Adaboost، تقویت گرادیان

1. Siddiqui, Haque, Khan, Adil & Shoaib
2. Ashraf
3. Tékouabou, Gherghina, Toulni, Mata & Martins
4. Tran, Le & Nguyen

نویسندگان	عنوان مقاله	نتایج	مدل‌های مورد استفاده
مونیر و همکاران[1]، ۲۰۲۲	استفاده از یادگیری ماشین برای پیش‌بینی ترک مشتریان در بانکداری	مدل Random Forest بهترین عملکرد را با دقت F1 معادل ۹۱/۹درصد ارائه داد.	Random Forest، AdaBoost، SMOTE
دیاس، گودینیو و تورس[2]، ۲۰۲۰)	استفاده از یادگیری ماشین برای پیش‌بینی ترک مشتریان در بانکداری خرد	روش Stochastic Boosting بهترین نتایج را با داده‌های چالش‌برانگیز ارائه داد.	تقویت تصادفی، جنگل تصادفی، تقویت گرادیان، درخت تصمیم
الیاس و همکاران، ۲۰۲۰	پیش‌بینی تراکنش‌های آینده در داده‌های بانکی نامتعادل بزرگ	بهینه‌سازی مدل LightGBM به دقت ۸۹درصد منجر شد.	LightGBM، جنگل تصادفی، رگرسیون لجستیک، Adaboost
شینگی[3]، ۲۰۲۰	یادگیری ماشین برای پیش‌بینی پیش‌فرض وام‌ها در بانک‌ها	روش Federated Learning به همراه SMOTE عملکرد پیش‌بینی را بهبود بخشید.	یادگیری فدرال، SMOTE، تجمع وزنی
امیرحسنخانی، طلوعی اشلقی، رادفر و پورابراهیمی، ۱۴۰۳	ارائه یک مدل ترکیبی مبتنی بر یادگیری ماشینی برای طبقه‌بندی مشتریان مشترک صنعت بانکداری و بیمه	نتایج به‌دست‌آمده از طبقه‌بندی مشتریان با استفاده از مدل پیشنهادی در این تحقیق، دقت بالای ۹۹ درصد را نشان می‌دهد.	در این تحقیق، از یک مدل ترکیبی مبتنی بر ماشین بردار پشتیبان (SVM) و الگوریتم ژنتیک استفاده شده است
نصیرالاسلامی، صنیعی، عباسیان، فتح‌پور کاشانی و قیصری، ۱۴۰۳	پیش‌بینی سپرده‌های بانکی به روش یادگیری ماشین	این مقاله نشان می‌دهد که مدل‌های یادگیری ماشین می‌توانند با دقت بالایی میزان سپرده‌های بانکی را پیش‌بینی کنند.	الگوریتم‌های مختلف یادگیری ماشین
چهره و سرآبادانی، ۱۴۰۲	ارائه مدلی مبتنی بر الگوریتم جنگل تصادفی و بهینه‌سازی جایا برای پیش‌بینی ریزش مشتریان بانکی	مدل پیشنهادی با ترکیب الگوریتم جنگل تصادفی و بهینه‌سازی جایا توانسته است با دقت بالایی ریزش مشتریان بانکی را پیش‌بینی کند.	الگوریتم جنگل تصادفی و بهینه‌سازی جایا
جعفری اسکندری و روحی، ۱۳۹۶	مدیریت ریسک اعتباری مشتریان بانکی با استفاده از روش ماشین بردار تصمیم بهبودیافته با الگوریتم ژنتیک با رویکرد داده‌کاوی	این پژوهش الگویی برای پیش‌بینی شاخص نرخ وصول مشتریان ارائه می‌دهد که دقت بیشتری نسبت به روش‌های قبلی دارد.	ماشین بردار تصمیم بهبودیافته با الگوریتم ژنتیک
خانلری، احراری و میرپور، ۱۳۹۵	پیش‌بینی ارزش طول عمر مشتریان بانکی با استفاده از تکنیک دسته‌بندی گروهی داده‌ها (GMDH) در شبکهٔ عصبی	این مطالعه نشان می‌دهد که استفاده از شبکه عصبی GMDH می‌تواند با دقت بالایی ارزش طول عمر مشتریان بانکی را پیش‌بینی کند.	شبکه عصبی GMDH
غلامیان و مظفری، ۱۳۹۵	پیش‌بینی ارزش مشتریان جدید بانک بر مبنای مدل آر.اف.ام با استفاده از درخت تصمیم بهبودیافته در راستای کاهش حداکثر حافظه مورد نیاز	این پژوهش روشی برای بهبود درخت تصمیم با استفاده از شبکه عصبی ارائه می‌دهد که به کاهش حداکثر حافظه مورد نیاز منجر می‌شود.	درخت تصمیم با استفاده از شبکه عصبی

1. Muneer et al.
2. Dias, Godinho & Torres
3. Shingi

اگرچه درحال‌حاضر در استفاده از تکنیک‌های یادگیری ماشین در زمینه پیش‌بینی ارزش و رفتار مشتریان بانکی پیشرفت‌های چشمگیری ایجاد شده است، محدودیت‌هایی وجود دارد که به شکاف‌هایی در تحقیقات در این زمینه منجر شده است. به عبارت دیگر، تمرکز در درجهٔ اول به‌سمت استفاده از مدل‌های پیچیده یادگیری ماشینی است، درحالی‌که از ارزیابی جامع تکنیک‌های پیش‌پردازش داده غفلت می‌شود. در واقع، بسیاری از مطالعات از روش‌هایی مانند SMOTE برای مدیریت عدم تعادل داده‌ها استفاده کرده‌اند که به‌دلیل تولید نمونه‌های مصنوعی نامعتبر، می‌تواند بر کیفیت پیش‌بینی مدل‌ها تأثیر منفی بگذارد. یکی دیگر از نقاط ضعف پژوهش‌های گذشته، کم‌توجهی به فرایند جامع انتخاب ویژگی‌های مؤثر است. پژوهش‌هایی مانند گلال و همکاران (۲۰۲۲) و پنگ و همکاران (۲۰۲۳) عمدتاً براستفاده از مجموعه داده‌های کامل تمرکز داشته‌اند و به ندرت از تکنیک‌های انتخاب ویژگی مؤثر استفاده کرده‌اند. این در حالی است که انتخاب دقیق ویژگی‌ها، نقش کلیدی در بهبود عملکرد مدل‌ها و کاهش پیچیدگی آن‌ها دارد. یکی دیگر از شکاف‌های موجود، عدم‌تحلیل جامع ابعاد مختلف ارزش مشتری است. پژوهش‌هایی نظیر خانلری و همکاران (۱۳۹۵) تنها بر یک جنبه خاص از ارزش مشتری، مانند ارزش طول عمر مشتری یا احتمال ریزش، تمرکز کرده‌اند. این در حالی است که ارتباط بین شاخص‌های مختلف اقتصادی و ارزش مشتریان نیازمند رویکردهای جامع‌تری است.

همچنین، استفاده محدود از رویکردهای تطبیقی برای مدیریت داده‌های نامتعادل، یکی دیگر از چالش‌های پژوهش‌های گذشته است. تحقیقات مونیر و همکاران (۲۰۲۲) و دیاس و همکاران (۲۰۲۰) نشان می‌دهد که داده‌های نامتعادل، همچنان در این حوزه چالش اساسی محسوب می‌شود. بااین‌حال، بهره‌گیری از رویکردهای تطبیقی پیشرفته نظیر یادگیری انتقالی و یادگیری تعاملی کمتر بررسی شده است. در نهایت، بسیاری از پژوهش‌ها، مانند اشرف (۲۰۲۴) و ترن و همکاران (۲۰۲۳)، عملکرد مدل‌ها را در چارچوب داده‌های آزمایشگاهی ارزیابی کرده‌اند؛ اما ارزیابی این مدل‌ها در محیط‌های واقعی و تحت شرایط پویا و نامتعادل به‌طور کامل انجام نشده است.

بر اساس تحلیل فوق، می‌توان شکاف‌های اصلی پژوهشی را به‌صورت زیر دسته‌بندی کرد:

۱. فقدان رویکردهای جامع که تکنیک‌های پیش‌پردازش پیشرفته را با مدل‌های یادگیری ماشین ترکیب کند.
۲. کمبود استفاده از روش‌های جامع برای انتخاب ویژگی‌های کلیدی که می‌توانند تأثیر مستقیم بر بهبود عملکرد مدل‌های یادگیری ماشین داشته باشند.
۳. تمرکز محدود بر یک یا چند شاخص مشتری به‌جای رویکرد جامع‌نگر به ارزش مشتری.
۴. عدم استفاده گسترده از تکنیک‌های تطبیقی پیشرفته برای مدیریت داده‌های نامتعادل و ناهمگن.
۵. نبود ارزیابی جامع مدل‌ها در محیط‌های واقعی با داده‌های پویا.

برخی ویژگی‌ها این مطالعه را به تلاش بهتری برای رسیدگی به شکاف‌های موجود تبدیل می‌کند. اول، جنبه‌های مربوط به استفاده از روش‌های نمونه‌گیری تطبیقی مدرن، کیفیت داده‌های ورودی را افزایش می‌دهد، در حالی که بر موانعی که در طول رویکردهای سنتی، از جمله SMOTE با آن مواجه می‌شوند، غلبه می‌کنند. دوم، استفاده از یک فرایند ترکیبی برای انتخاب ویژگی‌های کلیدی با استفاده از تحلیل هم‌بستگی و اهمیت ویژگی‌ها مبتنی بر الگوریتم جنگل تصادفی، به شناسایی و حذف ویژگی‌های غیرضروری کمک می‌کند و باعث بهبود عملکرد مدل‌ها می‌شود. سوم،

پیش‌بینی رویکرد چندبعدی ارزش طول عمر مشتری و احتمال ریزش در یک مدل جامع ادغام می‌شود. چهارم، استفاده از مدل‌ها در زمینه‌های واقعی و داده‌های پویا، شکاف بین تحقیقات نظری و کاربردی را کاهش می‌دهد. در نهایت، بهره‌گیری از ترکیب مدل‌های یادگیری ماشینی تطبیقی و رویکردهای ترکیبی، دقت و کارایی پیش‌بینی‌ها را بهبود می‌بخشد. این مقاله با پرداختن به محدودیت‌ها و ارائه یک رویکرد منسجم، گامی مؤثر در جهت افزایش پیش‌بینی ارزش مشتری بانک و مدیریت داده‌های نامتعادل برمی‌دارد.

روش‌شناسی پژوهش

این تحقیق کاربردی است و هدف اصلی آن پیش‌بینی ارزش مشتریان بانک با استفاده از الگوریتم‌های یادگیری ماشین و مقابله با عدم تعادل داده‌ها است. روش تحقیق مورد استفاده در این تحقیق کتابخانه‌ای است. جامعهٔ آماری شامل داده‌های مربوط به ۲هزار مشتری بانک تجارت ایران و اطلاعات حساب بانکی آن‌ها در یک سال است. در این مجموعه داده، ۱۳۹۷ مشتری ارزشمند و ۶۰۳ مشتری غیرارزشمند شناسایی شدند. برای هر مشتری ۱۴ ویژگی مرتبط با رفتار و تراکنش‌های مالی در نظر گرفته شده است که جزئیات آن‌ها در جدول ۲ ارائه شده است.

جدول ۲. اطلاعات مربوط به ویژگی‌های مجموعه داده

نام ویژگی	توضیحات	نوع
آخرین تأخر	آخرین تعداد روزهایی که مشتری تراکنشی نداشته است	عددی
حداکثر تأخر	بیشترین تعداد روزهایی که مشتری تراکنشی نداشته است	عددی
کمترین تأخر	کمترین تعداد روزهایی که مشتری تراکنشی نداشته است	عددی
تعداد دفعات مراجعه	تعداد دفعات مراجعه به بانک توسط مشتری	عددی
جمع کل واریزی‌ها	جمع کل مبلغ واریزی‌های مشتری	عددی
تعداد کل واریزی‌ها	تعداد کل واریزی‌های مشتری	عددی
جمع کل برداشت‌ها	جمع کل مبلغ برداشت‌های مشتری	عددی
تعداد کل برداشت‌ها	تعداد کل برداشت‌های مشتری	عددی
کمترین مانده	کمترین مانده حساب مشتری	عددی
بیشترین مانده	بیشترین مانده حساب مشتری	عددی
تعداد کل خدمات	تعداد کل خدماتی که مشتری استفاده کرده است	عددی
کل تراکنش‌های غیرحضوری	کل تراکنش‌های غیرحضوری مشتری	عددی
آخرین مانده	آخرین مانده حساب مشتری	عددی
ارزش مشتری	ارزش مشتری (۱ با ارزش و ۰ بی‌ارزش)	باینری

در این مطالعه، از زبان برنامه‌نویسی پایتون[1] به‌عنوان ابزار اصلی برای پیاده‌سازی مدل‌ها و تحلیل داده‌ها استفاده شد. فرایند انتخاب ویژگی‌های کلیدی با استفاده از یک رویکرد ترکیبی شامل تحلیل هم‌بستگی و روش Feature

1. Python

Importance مبتنی‌بر الگوریتم جنگل تصادفی انجام شد. این فرایند به حذف ویژگی‌های تکراری یا کم‌اهمیت و انتخاب ویژگی‌هایی منجر شد که تأثیر بیشتری بر پیش‌بینی ارزش مشتریان داشتند. در این پژوهش، داده‌ها با استفاده از روش استانداردسازی Z-score نرمال‌سازی شدند تا کیفیت مدل‌ها و دقت پیش‌بینی‌ها بهبود یابد. برای مقابله با مشکل عدم تعادل کلاس‌ها، از روش نمونه‌گیری تطبیقی مصنوعی (ADASYN) استفاده شد. هدف ADASYN ایجاد نمونه‌های مصنوعی برای نقاط بحرانی است که معمولاً توسط مدل‌ها به‌سختی طبقه‌بندی می‌شوند. این روش باعث بهبود عملکرد مدل‌ها در شناسایی مشتریان ارزشمند (کلاس اقلیت) شد. در ادامه، ۱۱ الگوریتم یادگیری ماشین شامل CatBoost، XGBoost، Random Forest، LightGBM و سایر مدل‌های خطی و غیرخطی برای پیش‌بینی ارزش مشتریان به‌کار گرفته شدند. به‌منظور بهینه‌سازی عملکرد این الگوریتم‌ها، از چارچوب Optuna برای تنظیم خودکار هایپرپارامترها استفاده شد. برای اطمینان از ارزیابی دقیق مدل‌ها و به حداقل رساندن اثرهای تصادفی بودن در داده‌ها، از یک رویکرد اعتبارسنجی متقاطع پنج‌برابری استفاده شد. در نهایت، عملکرد مدل‌ها با استفاده از ۴ معیار اصلی ارزیابی شامل صحت[1]، دقت[2]، فراخوانی[3] و امتیاز F1 سنجیده شد. بهترین مدل‌ها براساس عملکرد کلی شناسایی و گزارش شدند که در میان آن‌ها مدل CatBoost بهترین عملکرد را ارائه داد. این پژوهش با ارائه رویکردی جامع، گامی مؤثر در بهبود پیش‌بینی ارزش مشتریان بانکی و مدیریت داده‌های نامتعادل برداشته است.

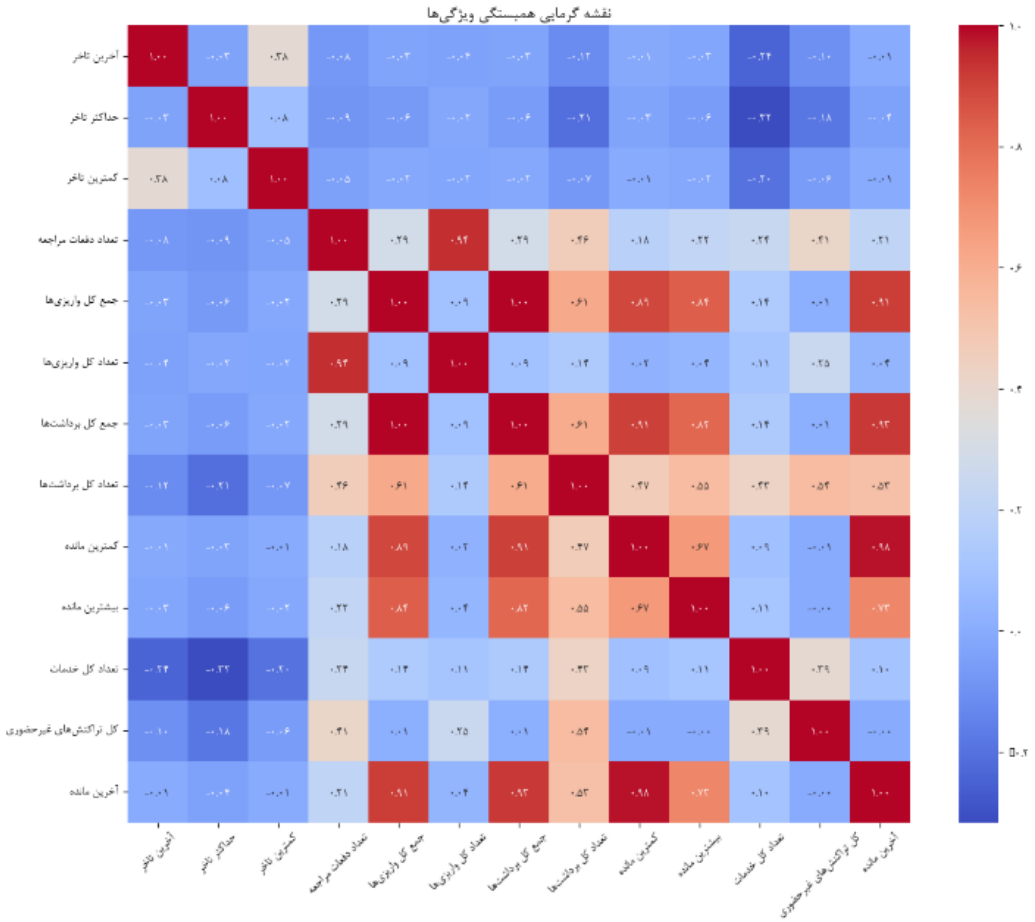
پیش‌پردازش داده‌ها

پیش پردازش داده‌ها تکنیکی است که قابلیت استفاده از مجموعه داده را بیشتر می‌کند. در این فرایند، داده‌هایی که ناقص هستند، قالب‌بندی اشتباه دارند، نامربوط هستند یا اطلاعات تکراری دارند، گاهی تغییر یا پاک می‌شوند. این فرایند همچنین می‌تواند مستلزم تبدیل داده‌های متنی به مقادیر عددی و طراحی مجدد ویژگی‌ها باشد (کوتسیانتیس[4]، ۲۰۱۱). در فرایند پیش‌پردازش داده‌ها، هیچ داده‌ای از دست نرفته و همچنین هیچ داده تکراری وجود نداشته است. تعداد داده‌های پرت نیز نسبت به کل داده‌ها بسیار کم بوده و قابل چشم‌پوشی است (هاج و آستین[5]، ۲۰۰۴). روش دیگری که اغلب در مرحله پیش‌پردازش داده‌ها اتفاق می‌افتد، مقیاس‌بندی داده‌ها است. مقیاس‌بندی روشی برای استانداردسازی دامنه متغیرها یا ویژگی‌های مستقل است. از این‌رو در این مطالعه، StandardScaler روشی برای استانداردسازی ویژگی‌های عددی بود. StandardScaler از کتابخانه scikit-learn در پایتون برای ایجاد مقادیر ویژگی استاندارد شده در مقیاسی با میانگین صفر و انحراف استاندارد یک استفاده می‌کند. به عبارت دیگر، داده‌ها را به‌گونه‌ای تبدیل می‌کند که همه مقادیر ویژگی یک توزیع استاندارد را به اشتراک بگذارند (آلدی، نظومی، سنتوسا و جنیدی[6]، ۲۰۲۳).

1. Accuracy
2. Precision
3. Recall
4. Kotsiantis
5. Hodge and Austin
6. Aldi, Nozomi, Sentosa & Junaidi

انتخاب مهم‌ترین ویژگی‌ها

برای انتخاب ویژگی‌های کلیدی که بیشترین تأثیر را در پیش‌بینی دارند، فرایندی گام‌به‌گام انجام شد. ابتدا بـرای بررسـی هم‌بستگی بین ویژگی‌ها، از ماتریس هم‌بستگی استفاده شد. هدف از این مرحله، شناسایی ویژگی‌هایی بود که به‌شدت به یکدیگر وابسته‌اند و ممکن است اطلاعات تکراری ارائه دهند (مهرگان و خانی، ۱۴۰۳). با اسـتفاده از ضـریب هم‌بسـتگی (پیرسون)، روابط بین ویژگی‌ها محاسبه شد و ویژگی‌هایی که هم‌بستگی بالاتر از ۸۰درصد داشتند، بـه‌عنوان ویژگی‌هـای بالقوه تکراری شناسایی شدند (لونگ و همکاران[1]، ۲۰۲۲؛ خانی، کـزازی و تقـوی فـرد، ۱۴۰۱). شـکل ۱ نقشـه گرمـایی هم‌بستگی ویژگی‌ها را به تصویر کشیده است و نشان می‌دهد کدام ویژگی‌ها با یکدیگر هم‌بستگی بالایی دارند.

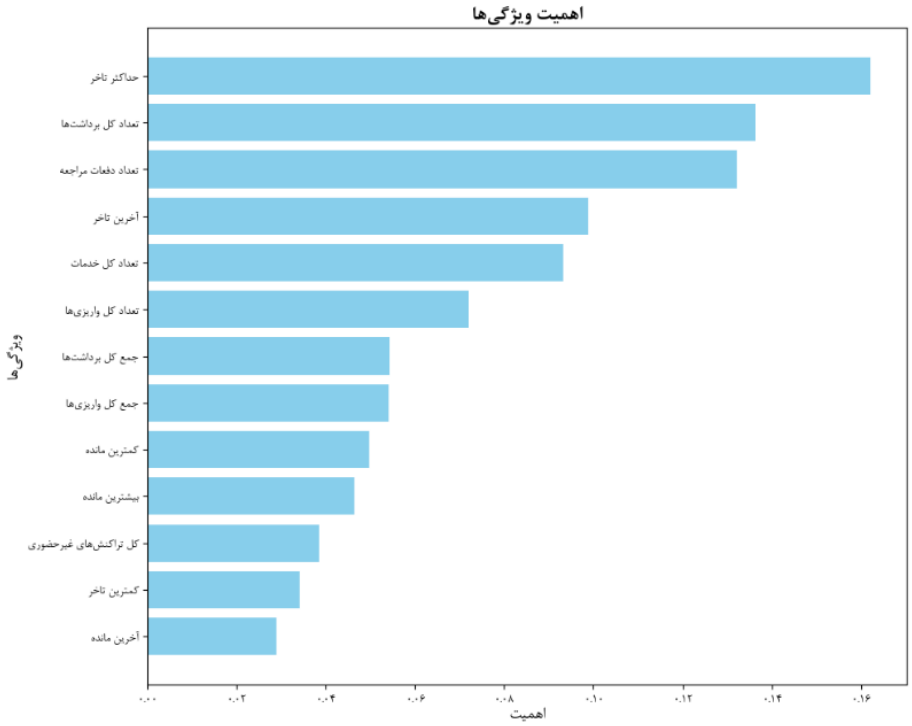


شکل ۱. نقشهٔ حرارتی هم‌بستگی ویژگی‌های مختلف مشتریان بانکی

در مرحلهٔ بعد، از روش انتخاب ویژگی مبتنی‌بر اهمیـت ویژگی‌هـا بـا اسـتفاده از الگـوریتم جنگـل تصـادفی بـرای تصمیم‌گیری در مورد نگه‌داشتن یا حذف این ویژگی‌های هم‌بسته استفاده شد. الگوریتم جنگل تصادفی با محاسبه اهمیت هر ویژگی، مشخص می‌کند که هر ویژگی چه میزان در کاهش خطا و بهبود پیش‌بینی نقش دارد. ویژگی‌هایی که اهمیت بیشتری داشتند، در مجموعه ویژگی‌ها نگه داشته شدند و سایر ویژگی‌های هم‌بسـته حـذف شـدند (کـان، نگـون، لـی و

1. Luong et al.

رامیرز[1]، ۲۰۱۹). شکل ۲ اهمیت ویژگی‌ها را نشان می‌دهد که بر اساس الگوریتم جنگل تصادفی محاسبه شده است. این نمودار اهمیت نسبی هر ویژگی در پیش‌بینی ارزش مشتریان را نمایش می‌دهد.



شکل ۲. اهمیت ویژگی‌ها براساس الگوریتم جنگل تصادفی

این فرایند به انتخاب ۸ ویژگی اصلی منجر شد که شامل آخرین تأخر، حداکثر تأخر، کمترین تأخر، تعداد دفعات مراجعه، جمع کل برداشت‌ها، تعداد کل برداشت‌ها، تعداد کل خدمات، کل تراکنش‌های غیرحضوری هستند. با این رویکرد ترکیبی (تحلیل هم‌بستگی و اهمیت ویژگی‌ها)، مدل از اطلاعات کلیدی و غیرتکراری بهره‌مند شد که تأثیر مستقیمی بر کاهش پیچیدگی مدل، بهبود دقت پیش‌بینی و جلوگیری از اثرهای منفی هم‌خطی ویژگی‌ها داشت.

مدیریت داده‌های نامتعادل

یکی از چالش‌های اصلی در این پژوهش، نامتعادل بودن داده‌ها بود. در مجموعه داده‌های مورد استفاده، تعداد نمونه‌های مربوط به مشتریان «ارزشمند» به‌مراتب بیشتر از نمونه‌های «غیرازشمند» بود. این عدم‌تعادل باعث می‌شود که مدل‌های یادگیری ماشین به سمت پیش‌بینی کلاس اکثریت (مشتریان ازشمند) تمایل پیدا کنند و درنتیجه، دقت پیش‌بینی برای کلاس اقلیت (مشتریان غیرارزشمند) کاهش یابد. برای مقابله با این چالش، از رویکرد نمونه‌گیری مصنوعی تطبیقی یا ADASYN به‌عنوان رویکرد اصلی مدیریت داده‌های نامتعادل استفاده شد. ADASYN یک روش بیش‌نمونه‌گیری برای متوازن‌سازی داده‌های نامتعادل است که با تولید نمونه‌های مصنوعی برای کلاس اقلیت، مشکل نامتعادل بودن

1. Conn, Ngun, Li & Ramirez

داده‌ها را برطرف می‌کند. برخلاف روش‌های یکنواخت مانند SMOTE، ADASYN روی نمونه‌هایی تمرکز دارد که طبقه‌بندی آن‌ها دشوارتر است (نقاط نزدیک به مرز تصمیم‌گیری یا نقاطی با تراکم پایین) (مجاهد و همکاران[1]، ۲۰۲۴). این ویژگی باعث می‌شود مدل یادگیری ماشین بتواند نمونه‌های چالش‌برانگیز را بهتر بشناسد. مراحل روش ADASYN به شرح زیر است:

۱. محاسبه نسبت عدم تعادل کلاس‌ها

در ابتدا، تعداد نمونه‌های کلاس اکثریت (N_{maj}) و کلاس اقلیت (N_{min}) محاسبه می‌شود. سپس نسبت عدم تعادل (d) بین این دو کلاس به‌صورت زیر تعریف می‌شود:

رابطهٔ ۱)
$$d = \frac{N_{maj} - N_{min}}{N_{min}}$$

۲. محاسبه سختی نمونه‌های کلاس اقلیت

برای هر نمونه x_i از کلاس اقلیت، تعداد همسایگان کلاس اکثریت (k) در یک محدوده مشخص (معمولاً با استفاده از kنزدیک‌ترین همسایه‌ها) محاسبه می‌شود. سختی هر نمونه r_i به‌صورت زیر تعریف می‌شود:

رابطهٔ ۲)
$$r_i = \frac{\text{تعداد همسایگان کلاس اکثریت برای } x_i}{k}$$

۳. نرمال‌سازی وزن‌ها

وزن r_i برای هر نمونه از کلاس اقلیت نرمال‌سازی می‌شود تا مجموع وزن‌ها برابر ۱ شود. G_i وزن نرمال‌شده نمونه x_i است. این کار به‌صورت زیر انجام می‌شود:

رابطهٔ ۳)
$$G_i = \frac{r_i}{\sum_{j=1}^{N_{min}} r_j}$$

۴. تولید نمونه‌های مصنوعی

نمونه‌های مصنوعی با استفاده از ترکیب خطی نقاط اصلی و نزدیک‌ترین همسایگان آن تولید می‌شوند. برای هر نمونه x_i، نمونه مصنوعی x_{new} با فرمول زیر ایجاد می‌شود که در آن x_{nn} یکی از نزدیک‌ترین همسایگان x_i و λ یک مقدار تصادفی بین صفر و یک که تنوع نمونه‌های مصنوعی را تضمین می‌کند.

رابطهٔ ۴)
$$x_{new} = x_i + \lambda(x_{nn} - x_i)$$

۵. تعداد نمونه‌های مصنوعی تولید شده

تعداد کل نمونه‌های مصنوعی G که باید تولید شود، به‌صورت زیر محاسبه می‌شود که در آن β پارامتری است که میزان تعادل نهایی را تعیین می‌کند (معمولاً مقدار آن بین صفر و یک انتخاب می‌شود).

1. Mujahid et al.

$$G = (N_{maj} - N_{min}) * \beta$$ رابطهٔ ۵)

الگوریتم‌های یادگیری ماشین

در این پژوهش، برای پیش‌بینی ارزش مشتریان بانکی، ۱۱ الگوریتم مختلف یادگیری ماشین استفاده شدند. انتخاب این الگوریتم‌ها براساس توانایی آن‌ها در طبقه‌بندی، پایداری، و تطبیق‌پذیری در برابر داده‌های مختلف صورت گرفت. علاوه‌براین، برای بهینه‌سازی عملکرد هر مدل، از ابزار قدرتمند Optuna استفاده شد که فرایند بهینه‌سازی هایپرپارامترها را به‌صورت خودکار و کارآمد انجام می‌دهد. Optuna یک ابزار پیشرفته و قدرتمند برای بهینه‌سازی هایپرپارامترها است که از الگوریتم‌های جست‌وجوی کارآمد مانند TPE[1] برای یافتن بهترین مقادیر پارامترها استفاده می‌کند. مراحل استفاده از Optuna در این پژوهش به شرح زیر بود:

۱. تعریف تابع هدف[2]: در این پژوهش، تابع هدف برای هر مدل شامل تنظیم هایپرپارامترها، آموزش مدل، و ارزیابی آن بر اساس معیار F1 Score بود.

۲. جست‌وجوی هایپرپارامترها: Optuna به‌صورت خودکار بهترین ترکیب پارامترها را با استفاده از چندین مرحله آزمایش (Trials) پیدا کرد.

۳. انتخاب بهترین پارامترها: بهترین هایپرپارامترهای هر مدل براساس حداکثر مقدار F1 Score انتخاب شدند.

در جدول ۳، یازده الگوریتم یادگیری ماشین مورد استفاده به همراه توضیح مختصر و هایپرپارامترهای بهینه به دست آمده با استفاده از Optuna آورده شده است:

جدول ۳. الگوریتم‌های یادگیری ماشین مورد استفاده

مدل	هایپرپارامترهای بهینه	توضیح مختصر	منابع
CatBoost	{'iterations': 138, 'depth': 3, 'learning_rate': 0.2677}	الگوریتم Gradient Boosting بهینه‌سازی شده برای داده‌های دسته‌ای و عددی با عملکرد بالا.	پروکورنکوا و همکاران (۲۰۱۸)، دوروگوش و همکاران (۲۰۱۸)
XGBoost	{'n_estimators': 181, 'max_depth': 10, 'learning_rate': 0.0846}	نسخه‌ای پیشرفته از Gradient Boosting با تمرکز بر بهینه‌سازی حافظه و سرعت.	بنتژاک، چورگ و مارتینز مونیز[3] (۲۰۲۰)، جعفرنژاد، رضاسلطانی و خانی (۱۴۰۳)
Random Forest	{'n_estimators': 155, 'max_depth': 17}	مدل Ensemble مبتنی بر درخت‌های تصمیم که برای داده‌های پیچیده و متنوع مناسب است.	ایرانزاد و لیو[4] (۲۰۲۴)، شونلاو و زو[5] (۲۰۲۰)

1. Tree-structured Parzen Estimator
2. Objective Function
3. Bentéjac, Csörgő & Martínez-Muñoz
4. Iranzad & Liu
5. Schonlau & Zou

مدل	هایپرپارامترهای بهینه	توضیح مختصر	منابع
LightGBM	{'n_estimators': 126, 'max_depth': 20, 'learning_rate': 0.0778}	نسخه‌ای سریع‌تر و سبک‌تر از Gradient Boostingکه برای داده‌های بزرگ بسیار کاربردی است.	یانگ، چن، یانگ و تیان[1] (۲۰۲۳)، نگوین، نگوین، لی و بوی[2] (۲۰۲۲)
AdaBoost	{'n_estimators': 172, 'learning_rate': 0.9875}	الگوریتمی مبتنی بر تقویت طبقه‌بندهای ساده مانند درخت تصمیم برای بهبود دقت.	واینر، اولسون، بلایچ و میز[3] (۲۰۱۷)، هو، هو و میبانک[4] (۲۰۰۸)
Extra Trees	{'n_estimators': 157, 'max_depth': 20}	الگوریتمی مشابه Random Forest با استفاده از تصمیم‌گیری‌های تصادفی‌تر برای هر گره درخت.	سان مورینو، مارنیسا و سوناردی[5] (۲۰۲۳)، ماستلینی و همکاران[6] (۲۰۲۲)
MLP	{'hidden_layer_sizes': (50,), 'learning_rate_init': 0.028, 'max_iter': 202}	یک شبکه عصبی چندلایه که برای یادگیری روابط غیرخطی و پیچیده میان داده‌ها استفاده می‌شود.	کروس، مستغیم، بورگلت، برونه و اشتاینبرچر[7] (۲۰۲۲)، پرزیبیلا کاسپرک و مارفو[8] (۲۰۲۴)
Decision Tree	{'max_depth': 4, 'min_samples_split': 16}	الگوریتمی ساده و تفسیری که از ساختار درختی برای تصمیم‌گیری استفاده می‌کند.	مینیه و جره[9] (۲۰۲۴)، کوتسیانتیس (۲۰۱۱)
SVM	{'C': 37.6175, 'kernel': 'rbf'}	الگوریتمی قدرتمند برای طبقه‌بندی با استفاده از یک ابرصفحه جداکننده میان کلاس‌ها.	گویدو، فریسی، لوفارو و کونفورتی[10] (۲۰۲۴)، آواد و خانا[11] (۲۰۱۵)
Logistic Regression	{'C': 1.6467, 'max_iter': 104}	یک مدل ساده خطی که برای مسائل طبقه‌بندی استفاده می‌شود و احتمال کلاس‌ها را محاسبه می‌کند.	تولز و مئورر[12](۲۰۱۶)، ویتینهوف و مک‌کالاک[13] (۲۰۰۶)
KNN	{'n_neighbors': 15}	الگوریتمی مبتنی بر مقایسه نزدیک‌ترین همسایگان یک نمونه برای تعیین کلاس آن.	هالدر، اُدین، اُدین، آریال و خریسات[14] (۲۰۲۴)، سیریوپولوس، کالامپالیکیس، کوتسیانتیس و ورهاتیس[15] (۲۰۲۳)

1. Yang, Chen, Yang & Tian
2. Nguyen, Nguyen, Le & Bui
3. Wyner, Olson, Bleich & Mease
4. Hu, Hu & Maybank
5. Sanmorino, Marnisah & Sunardi
6. Mastelini et al.
7. Kruse, Mostaghim, Borgelt, Braune & Steinbrecher
8. Przybyła-Kasperek & Marfo
9. Mienye & Jere
10. Guido, Ferrisi, Lofaro & Conforti
11. Awad & Khanna
12. Tolles & Meurer
13. Vittinghoff & McCulloch
14. Halder, Uddin, Uddin, Aryal & Khraisat
15. Syriopoulos, Kalampalikis, Kotsiantis & Vrahatis

در این پژوهش، داده‌های اصلی به دو مجموعه جداگانه برای آموزش و آزمون مدل تقسیم شدند. یعنی در Train-Test Split، داده‌ها به‌صورت ۸۰درصد در آموزش و ۲۰درصد در تست تقسیم شدند. روش اعتبارسنجی متقاطع با پنج برابر برای ارزیابی مدل برای کاهش تأثیرات تصادفی داده‌ها انجام شد. این روش یکی از رایج‌ترین و تحسین‌شده‌ترین روش‌های ارزیابی در یادگیری ماشینی است که امکان ارزیابی عملکرد مدل را در شرایط مختلف فراهم می‌کند. در این تحقیق داده‌ها به پنج قسمت مساوی تقسیم شدند. در هر مرحله از چهار قسمت به‌عنوان داده‌های آموزشی و یک قسمت به‌عنوان داده‌های آزمون استفاده شد. این فرایند به‌صورت دایره‌ای انجام شد به‌طوری که تمام قسمت‌ها یک بار به‌عنوان داده‌های آزمایشی مورد استفاده قرار گرفتند.

برای ارزیابی عملکرد مدل‌های یادگیری ماشین در این پژوهش، از چهار شاخص اصلی استفاده شده است. برای محاسبه این شاخص‌ها، از مفاهیم True Positive (TP) و True Negative (TN) برای پیش‌بینی‌های صحیح مدل استفاده شده است. TP نشان‌دهندهٔ تعداد مشتریان ارزشمندی است که مدل به‌درستی آن‌ها را شناسایی کرده و TN تعداد مشتریان غیرارزشمندی است که مدل به‌درستی آن‌ها را پیش‌بینی کرده است. همچنین، False Positive (FP) نشان‌دهندهٔ تعداد مشتریان غیرارزشمندی است که به اشتباه به‌عنوان مشتری ارزشمند پیش‌بینی شده‌اند و False Negative (FN) تعداد مشتریان ارزشمندی است که مدل به اشتباه آن‌ها را غیرارزشمند پیش‌بینی کرده است. این مقادیر به‌عنوان مبنای محاسبه شاخص‌های صحت (Accuracy)، دقت (Precision)، فراخوانی (Recall)، و امتیاز F1 به کار گرفته شده‌اند تا عملکرد مدل‌ها در پیش‌بینی ارزش مشتریان به‌طور جامع ارزیابی شود. جدول ۴ شاخص‌های ارزیابی مدل‌های یادگیری ماشین را نشان می‌دهد.

جدول ٤. شاخص‌های ارزیابی مدل‌های یادگیری ماشین

شاخص	تعریف	فرمول
صحت	نسبت پیش‌بینی‌های صحیح (چه مثبت و چه منفی) به کل نمونه‌ها.	$\frac{TP+TN}{TP+FP+FN+TN}$
دقت	نسبت پیش‌بینی‌های درست برای یک کلاس به کل نمونه‌هایی که به‌عنوان آن کلاس پیش‌بینی شده‌اند.	$\frac{TP}{TP+FP}$
فراخوانی	نسبت پیش‌بینی‌های درست برای یک کلاس به کل نمونه‌های واقعی آن کلاس.	$\frac{TP}{TP+FN}$
امتیاز F1	میانگین هارمونیک بین Precision و Recall برای ایجاد تعادل بین آن‌ها.	$\frac{2*Precision*Recall}{Precision+Recall}$

یافته‌های پژوهش

در این بخش، به بررسی نتایج مدل‌های مختلف یادگیری ماشین پرداخته شده است. تمامی مدل‌ها بر روی سیستمی با پردازنده Intel Core i7-13700H و ۱۶ گیگابایت رم و پایتون ۳/۱۲ اجرا شدند. در ادامه، نتایج عملکرد مدل‌ها آورده شده است. نتایج در جدول ۵ نشان داده شده است.

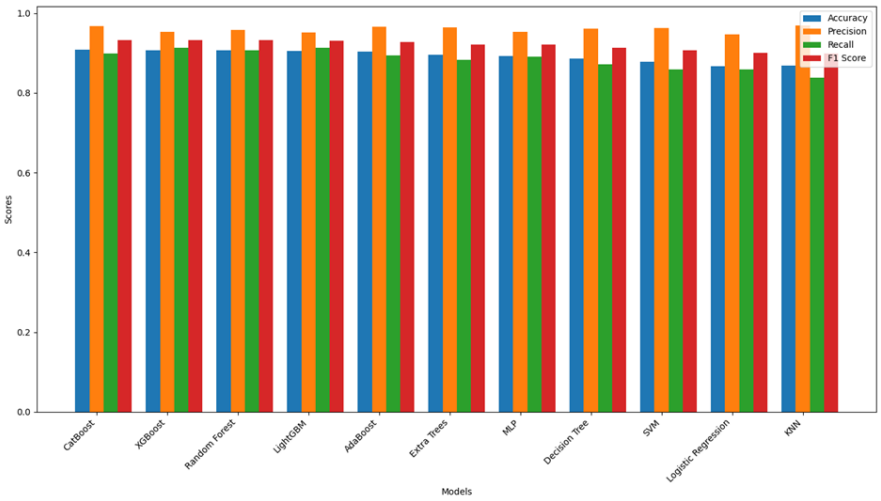
جدول ۵. مقایسه عملکرد مدل‌های مختلف یادگیری ماشین بر اساس معیارهای ارزیابی

مدل	صحت	دقت	فراخوانی	امتیاز F1
CatBoost	۰/۹۰۹	۰/۹۶۷۷	۰/۸۹۹۸	۰/۹۳۲۴
XGBoost	۰/۹۰۷۵	۰/۹۵۳۲	۰/۹۱۲۷	۰/۹۳۲۲
Random Forest	۰/۹۰۷۵	۰/۹۵۸۶	۰/۹۰۶۹	۰/۹۳۲
LightGBM	۰/۹۰۶	۰/۹۵۱	۰/۹۱۲۶	۰/۹۳۱۲
AdaBoost	۰/۹۰۳۵	۰/۹۶۵۴	۰/۸۹۴۱	۰/۹۳۸۳
Extra Trees	۰/۸۹۵۵	۰/۹۶۴۲	۰/۸۸۳۳	۰/۹۲۱۹
MLP	۰/۸۹۳	۰/۹۵۲۹	۰/۸۹۱۲	۰/۹۲۰۸
Decision Tree	۰/۸۸۵۵	۰.۹۶۰۷	۰/۸۷۱۸	۰/۹۱۳۹
SVM	۰/۸۷۷۵	۰/۹۶۲۵	۰/۸۵۸۳	۰/۹۰۷۳
Logistic Regression	۰/۸۶۷۵	۰/۹۴۷۲	۰/۸۵۸۳	۰/۹۰۰۵
KNN	۰/۸۶۸۵	۰/۹۶۸۸	۰/۸۳۸۹	۰/۸۹۹

در این مطالعه، ۱۱ مدل یادگیری ماشین برای عملکرد در برابر ۴ معیار ارزیابی اصلی مورد ارزیابی قرار گرفتند: صحت (Accuracy)، دقت (Precision)، فراخوانی (Recall) و امتیاز F1. این معیارها توانایی مدل‌ها را در شناسایی مشتریان ارزشمند و دقتی را نشان می‌دهد که در آن مدل، از پیش‌بینی اشتباه مشتریان غیر با ارزش به‌عنوان مشتریان ارزشمند جلوگیری می‌کند. نتایج نشان داد که مدل‌های مبتنی‌بر روش‌های یادگیری گروهی مانند CatBoost، XGBoost و Random Forest بهترین عملکرد را داشتند؛ زیرا توانستند به تعادل قابل ستایش دقت و فراخوانی دست یابند. در این میان بهترین مدل CatBoost با امتیاز F1 برابر ۰/۹۳۲۴ و صحت ۰/۹۰۹ بود. این مدل توانست دقت استثنایی ۰/۹۶۷۷ را در کنار فراخوانی ۰/۸۹۹۸ نشان دهد که گویای عملکرد متعادل در شناسایی با ارزش و جلوگیری از پیش‌بینی‌های اشتباه است. مدل XGBoost نیز با امتیاز F1 برابر ۰/۹۳۲۲ عملکردی نزدیک به CatBoost داشت و توانست دقت بالای ۰/۹۵۳۲ و فراخوانی ۰/۹۱۲۷ را به نمایش بگذارد. این عملکرد نشان‌دهندهٔ توانایی XGBoost در شناسایی مشتریان ارزشمند با کمترین خطای ممکن است. از سوی دیگر، مدل Random Forest نیز با امتیاز F1 برابر ۰/۹۳۲ و دقت ۰/۹۵۸۶ عملکرد بسیار خوبی ارائه داد. با وجود اینکه فراخوانی این مدل (۰/۹۰۶۹) کمی پایین‌تر از XGBoost بود، اما همچنان یکی از بهترین گزینه‌ها برای این مسئله محسوب می‌شود.

مدل‌های LightGBM و AdaBoost نیز عملکرد مناسبی داشتند. LightGBM با امتیاز F1 برابر ۰/۹۳۱۲ و دقت ۰/۹۵۱ توانست تعادلی نزدیک به XGBoost ارائه دهد، در حالی که فراخوانی آن برابر ۰/۹۱۲۶ بود. مدل AdaBoost با امتیاز F1 برابر ۰/۹۲۸۳ و دقت ۰/۹۶۵۴ نشان داد که در شناسایی صحیح مشتریان ارزشمند عملکرد بالایی دارد؛ اما فراخوانی آن برابر ۰/۸۹۴۱ بود که به‌معنای از دست دادن تعداد بیشتری از مشتریان ارزشمند نسبت به مدل‌های برتر است. در میان مدل‌های ساده‌تر، Decision Tree و Logistic Regression عملکرد متوسطی داشتند. Decision Tree با امتیاز F1 برابر ۰/۹۱۳۹ و دقت ۰/۹۶۰۷ توانست پیش‌بینی دقیقی ارائه دهد، اما فراخوانی پایین‌تر آن

(۰/۸۷۱۸) نشان داد که تعداد بیشتری از مشتریان ارزشمند شناسایی نشده‌اند. Logistic Regression نیز با امتیاز F1 برابر ۰/۹۰۰۵ و صحت ۰/۸۶۷۵ عملکرد ضعیف‌تری داشت که به‌دلیل محدودیت این مدل در شناسایی روابط غیرخطی و پیچیده بین ویژگی‌ها است. مدل SVM با امتیاز F1 برابر ۰/۹۰۷۳ و دقت ۰/۹۶۲۵ توانست پیش‌بینی‌های دقیقی انجام دهد، اما فراخوانی پایین‌تر آن (۰/۸۵۸۳)، نشان‌دهندهٔ ضعف در شناسایی تمامی مشتریان ارزشمند بود. مدل KNN نیز با امتیاز F1 برابر ۰/۸۹۹ و دقت ۰/۹۶۸۸، عملکرد قابل‌قبولی از نظر دقت داشت؛ اما به‌دلیل فراخوانی پایین‌تر (۰/۸۳۸۹) نتوانست به اندازه مدل‌های پیچیده‌تر مؤثر باشد.



شکل ۳ . مقایسهٔ مدل‌های یادگیری ماشین براساس معیارهای مختلف ارزیابی

همان‌طور که در شکل ۳ نشان داده شده است، مدل‌های مبتنی‌بر یادگیری جمعی مانند CatBoost، XGBoost و Random Forest بهترین عملکرد را در تمامی شاخص‌ها داشتند و توانستند تعادل مناسبی میان شناسایی مشتریان ارزشمند و جلوگیری از پیش‌بینی اشتباه ایجاد کنند. مدل‌های ساده‌تر، مانند Logistic Regression و KNN، به‌دلیل سرعت و سادگی می‌توانند در مسائل کوچک‌تر کاربردی باشند؛ اما برای مسائل پیچیده‌تر و داده‌های نامتعادل، عملکرد ضعیف‌تری نشان دادند. این نتایج تأکید می‌کند که انتخاب مدل‌های مناسب با روش‌های پیشرفته یادگیری جمعی می‌تواند برای پیش‌بینی دقیق مشتریان بانکی و بهینه‌سازی تصمیمات مدیریتی، تأثیر درخور توجهی داشته باشد.

بحث و نتیجه‌گیری

مطالعات انجام‌شده در سال‌های اخیر تأکید زیادی بر استفاده از روش‌های یادگیری ماشین برای پیش‌بینی رفتار و ارزش مشتریان بانکی داشته‌اند. با توجه به پیچیدگی داده‌های بانکی و اهمیت تصمیم‌گیری دقیق در این حوزه، بهره‌گیری از مدل‌های پیشرفته و روش‌های نوین پیش‌پردازش داده‌ها ضروری است. بااین‌حال، محدودیت‌هایی مانند عدم‌تعادل داده‌ها، عدم‌استفاده کافی از رویکردهای جامع برای انتخاب ویژگی‌های مؤثر و تحلیل ناقص ابعاد مختلف ارزش مشتری همچنان به‌عنوان چالش‌های کلیدی مطرح هستند. در این پژوهش، با استفاده از ۱۱ الگوریتم مختلف یادگیری ماشین،

پیش‌بینی ارزش مشتریان بانکی با تمرکز بر مسئله داده‌های نامتعادل بررسی شد. مدل‌های مورداستفاده شامل روش‌های مبتنی بر Ensemble Learning مانند CatBoost، XGBoost و Random Forest بودند که عملکرد بهتری نسبت به روش‌های خطی و ساده‌تر مانند Logistic Regression و KNN ارائه دادند. استفاده از روش ADASYN برای متوازن‌سازی داده‌ها، یکی از عوامل کلیدی در بهبود عملکرد مدل‌ها بود. این روش با تولید نمونه‌های مصنوعی برای کلاس اقلیت، توانست تعادل بین کلاس‌ها را برقرار کرده و مدل‌ها را قادر سازد تا ارزش مشتریان را با دقت و فراخوانی بهتری شناسایی کنند.

پژوهش حاضر با مقایسهٔ نتایج حاصل از ۱۱ الگوریتم یادگیری ماشین، به‌طور جامع نشان داد که روش‌های مبتنی بر یادگیری جمعی مانند CatBoost، XGBoost، و Random Forest، به‌دلیل توانایی بالا در مدیریت داده‌های پیچیده و نامتعادل، بهترین عملکرد را داشتند. نتایج این پژوهش حاکی از آن است که مدل CatBoost با امتیاز F1 برابر ۰/۹۳۲۴ و صحت ۰/۹۰۹ بهترین عملکرد کلی را ارائه داد. این مدل توانست تعادلی میان دقت بالا و فراخوانی ایجاد کند که به‌معنای توانایی شناسایی صحیح مشتریان ارزشمند و کاهش پیش‌بینی‌های اشتباه است. عملکرد قوی CatBoost ناشی از توانایی آن در استفاده بهینه از داده‌های نامتعادل و حساسیت به ویژگی‌های کلیدی مشتریان است. مدل XGBoost نیز عملکردی نزدیک به CatBoost داشت و با دقت ۰/۹۵۳۲ و فراخوانی ۰/۹۱۲۷ نشان داد که این الگوریتم برای شناسایی مشتریان ارزشمند با حداقل خطا مناسب است. از سوی دیگر، Random Forest با امتیاز F1 برابر ۰/۹۳۲ و دقت ۰/۹۵۸۶ گزینه‌ای مناسب برای مسائلی است که به تفسیرپذیری بالاتر مدل‌ها نیاز دارند، اگرچه فراخوانی کمی پایین‌تر آن، ممکن است به‌معنای از دست دادن تعدادی از مشتریان ارزشمند باشد. مدل‌های LightGBM و AdaBoost نیز عملکرد مناسبی از خود نشان دادند. LightGBM با امتیاز F1 برابر ۰/۹۳۱۲ و دقت ۰/۹۵۱، گزینه‌ای مناسب برای پردازش سریع و داده‌های حجیم است. AdaBoost با دقت ۰/۹۶۵۴ توانست پیش‌بینی‌های دقیقی ارائه دهد، اما فراخوانی پایین‌تر آن (۰/۸۹۴۱) نشان داد که ممکن است در شناسایی مشتریان ارزشمند کمی ضعیف‌تر عمل کند. مدل‌های ساده‌تر مانند Logistic Regression و KNN، علی‌رغم سادگی و سرعت بالا، در پیش‌بینی مشتریان ارزشمند در مقایسه با مدل‌های پیچیده‌تر، عملکرد ضعیف‌تری داشتند. به‌طور مثال، Logistic Regression با امتیاز F1 برابر ۰/۹۰۰۵ و صحت ۰/۸۶۷۵ نشان داد که در مواجهه با داده‌های غیرخطی و پیچیده محدودیت دارد. KNN نیز با دقت ۰/۹۶۸۸ عملکردی متوسط ارائه داد، اما فراخوانی پایین‌تر آن (۰/۸۳۸۹) مانعی برای شناسایی مشتریان ارزشمند بود.

پژوهش‌های پیشین، مانند مطالعات دیوی و همکاران (۲۰۲۴) و وو (۲۰۲۴)، نیز نشان داده‌اند که روش‌های مبتنی بر یادگیری جمعی مانند XGBoost و Random Forest به‌دلیل دقت بالا در تشخیص مشتریان ارزشمند، بر سایر مدل‌ها برتری دارند. بااین‌حال، این مطالعات عمدتاً از روش‌های ساده‌ای مانند SMOTE برای مدیریت داده‌های نامتعادل استفاده کرده‌اند، درحالی‌که پژوهش حاضر با بهره‌گیری از روش پیشرفته ADASYN توانست نمونه‌های مصنوعی معتبری برای کلاس اقلیت ایجاد کند که تأثیر مستقیمی بر بهبود عملکرد مدل‌ها داشت. همچنین، برخلاف پژوهش‌هایی نظیر گلال و همکاران (۲۰۲۲) که بیشتر برعملکرد عددی مدل‌ها تمرکز کرده‌اند، این پژوهش با استفاده از

یک رویکرد ترکیبی برای انتخاب ویژگی‌های کلیدی، شامل تحلیل هم‌بستگی و اهمیت ویژگی‌ها مبتنی بر الگوریتم جنگل تصادفی، گامی فراتر در بهبود عملکرد مدل‌ها برداشت. این فرایند باعث حذف ویژگی‌های کم‌اهمیت و تمرکز مدل‌ها بر اطلاعات کلیدی شد که در مطالعات پیشین کمتر به آن پرداخته شده است. در مقایسه با مطالعه اشرف (۲۰۲۴) که تنها به ارزیابی مدل‌ها در محیط‌های آزمایشگاهی پرداخته است، این پژوهش با ارزیابی مدل‌ها در شرایط واقعی و استفاده از داده‌های پویا، توانست عملکرد عملی مدل‌ها را نیز بررسی کند که یک شکاف مهم در پژوهش‌های پیشین را پر می‌کند. برخلاف مدل‌های ترکیبی گلال و همکاران (۲۰۲۲) یا چهره و سرآبادانی (۱۴۰۲) که بر بهینه‌سازی الگوریتم‌ها تمرکز داشته‌اند، مقالهٔ حاضر با تأکید بر سادگی و کارایی مدل CatBoost، روند آموزش مدل را تسهیل کرده و به خروجی‌هایی قابل استفاده برای سیاست‌گذاری در بانک‌ها منتهی شده است. همچنین، برخلاف مطالعاتی چون شینگی (۲۰۲۰) و امیرحسنخانی و همکاران (۱۴۰۳) که به‌طور خاص بر پیش‌فرض وام یا طبقه‌بندی مشتریان بین صنایع تمرکز دارند، پژوهش حاضر به‌شکل مستقیم، به پیش‌بینی ارزش مشتری با رویکرد عملیاتی برای بانک‌ها پرداخته است. با این حال، مقالهٔ حاضر در مقایسه با مطالعات فوق، از ویژگی‌های مهم و منحصربه‌فرد زیر برخوردار است:

اول، ذکر این نکته مهم است که متفاوت از بسیاری از مطالعات قبلی که از SMOTE در متعادل کردن مجموعه داده استفاده کرده‌اند، این مطالعه از روش ADASYN استفاده می‌کند که نمونه‌های سخت را هدف قرار می‌دهد و داده‌های مصنوعی دقیق‌تری تولید می‌کند.

دوم، انتخاب ویژگی با استفاده از تجزیه‌وتحلیل هم‌بستگی و انتخاب ویژگی اهمیت انجام می‌شود؛ زیرا این فرایند به کاهش پیچیدگی مدل کمک می‌کند و در عین حال دقت بیشتری به دست می‌آورد.

سوم، در حالی که بسیاری از مطالعات ذکر شده در بالا دارای داده‌هایی هستند که ممکن است خوداکسیداسیون یا مبتنی بر معیار باشند، مقالهٔ حاضر از داده‌های واقعی اولیه مشتریان یک بانک ایرانی برای افزایش اعتبار مطالعه و یافته‌های آن استفاده می‌کند.

چهارم، در حالی که سایر مقالات ارائه‌دهنده مقایسه مدل‌ها بر دستیابی به دقت کلی مدل تمرکز دارند، این مقاله حاضر با ارائه راه‌حل‌هایی برای مشکلات بانک‌ها با استفاده از مدل‌هایی برای بازاریابی هدفمند، در تصمیم‌گیری زمان مناسب برای صرف منابع برای حفظ مشتریان ارزشمند و در فرایند تصمیم‌گیری اعتبار، بیش از پیش ارائه می‌دهد. از این رو می‌توان مقاله حاضر را هم از لحاظ نظری و هم از نظر عملی با آثاری که به این جنبه عملی کم‌توجهی کرده‌اند، مانند گلال و همکاران (۲۰۲۲) و تکوابو و همکاران (۲۰۲۲)، متمایز کرد.

بر اساس نتایج این تحقیق، چندین مفهوم را می‌توان از این تحقیق برای رفع نیازهای بانک ها در حوزه مدیریت، بازاریابی و تصمیم‌گیری به‌دست آورد:

اولاً، مدل CatBoost که نتایج خوبی در جست‌وجوی مشتریان ارزشمند نشان داده است، به بانک‌ها اجازه می‌دهد تا بازاریابی هوشمندانه‌ای را برای مشتریان انجام دهند. این برای برنامه‌ریزی دقیق بازاریابی و توانایی برنامه‌ریزی در مورد گروه خاصی از مشتریان و نحوه حفظ آن‌ها مفید است.

دوم، مدیریت صحیح منابع بانک ها از طریق استفاده از نتایج مدل های پیشنهادی حاصل می‌شود. از این منظر

می‌توان ابتدا به مشتریان ارزشمند خدمات رسانی، اعتبار و حمایت کرد، در حالی که شرکت انرژی و پول خود را برای مشتریان کم ارزش هدر نخواهد داد.

سوم، استفاده از یادگیری ماشینی، توسعهٔ مدل‌هایی را امکان‌پذیر می‌سازد که در تعیین مستقیم مشتریان ارزشمندی که در آستانه انحراف هستند، مؤثر هستند. این یک پایه برای سازمان‌دهی برنامه های وفاداری مؤثر و درگیر کردن مشتریان ایجاد می‌کند. می‌تواند به‌طور کامل جریان ثابتی از درآمد را حفظ کند و سطوح فرسایش مشتری را کاهش دهد.

چهارم، متغیرهای به‌دست‌آمده از طریق این مدل‌ها نیز در فرایندهای تصمیم‌گیری اعتباری اتخاذ می‌شوند. به‌کارگیری پیش بینی دقیق در ارزیابی ریسک اعتباری مشتریان به تصمیم‌گیری صحیح در مورد تسهیلات ارائه شده و مدیریت مطالبات کمک می‌کند.

در انتها، از طریق استفاده از ابزارهای هوش مصنوعی توضیح‌پذیر از جمله SHAP در میان ابزارهای دیگر، همراه با یادگیری ماشینی، می‌توان وارد سطح بعدی تحلیل رفتاری مشتری شد. قابلیت‌های فنی این ابزارها عبارت است از تعریف سهم هر ویژگی در پیش‌بینی مدل و پیش‌بینی نتایج احتمالی از طریق توضیح علل زمینه‌ای که رفتار مشتریان بر اساس آن است. این ویژگی به مدیران بانک برای تصمیم گیری بهتر کمک می‌کند. از آنجایی که این پیشنهادها برای یک محیط بانکی واقعی قابل اجرا هستند، می‌توانند به‌عنوان نقطه شروعی برای ایجاد تغییرات در مفاهیم کیفیت خدمات و گسترش کارایی رویکردهای مدیریت مشتری‌مدار در نظر گرفته شوند.

پیشنهادهای تحقیقاتی

تحقیقات انجام‌شده در این مطالعه نشان داد که استفاده از الگوریتم‌های پیشرفته یادگیری ماشین، همراه با تکنیک‌های نوین پیش‌پردازش و نمونه‌گیری، می‌تواند به بهبود دقت و کارایی مدل‌های پیش‌بینی ارزش مشتریان در صنعت بانکداری کمک کند. بااین‌حال، برخی از محدودیت‌ها و چالش‌ها همچنان باقی مانده‌اند که می‌توانند در تحقیقات آینده موردبررسی قرار گیرند. از جمله این محدودیت‌ها، عدم‌ارزیابی مدل‌ها در محیط‌های زمانی پویا، وابستگی به داده‌های ساختاریافته و تراکنشی و عدم استفادهٔ گسترده از روش‌های یادگیری عمیق برای تحلیل داده‌های پیچیده‌تر است. علاوه‌براین، نیاز به مدل‌های قابل‌تفسیرتر و انطباق‌پذیرتر در محیط‌های واقعی بانکداری همچنان احساس می‌شود. با توجه به پیشرفت‌های سریع در حوزهٔ هوش مصنوعی و یادگیری ماشین، پیشنهادهایی برای تحقیقات آتی ارائه می‌شود تا شکاف‌های موجود برطرف شوند و مدل‌های پیش‌بینی ارزش مشتریان به سطوح بالاتری از دقت، تفسیرپذیری، و کاربردپذیری دست یابند.

- ادغام یادگیری ماشین با تحلیل‌های رفتاری و روان‌شناختی مشتریان: تحقیقات آتی باید تحلیل رفتاری و روانشناختی مشتری را همراه با داده‌های تراکنش در نظر بگیرد. این می‌تواند مدل‌های بهتری را برای پیش‌بینی رفتارهای مشتری و تعیین انگیزه‌های آن‌ها ارائه دهد.
- توسعهٔ مدل‌های قابل توضیح‌تر برای کاربردهای واقعی بانکی: با توجه به اهمیت شفافیت در تصمیم‌گیری‌های بانکی، پیشنهاد می‌شود که در تحقیقات آتی از مدل‌های یادگیری ماشین با قابلیت توضیح‌پذیری بیشتر، مانند

XAI[1]، بهره گرفته شود. استفاده از ابزارهایی نظیر LIME و SHAP می‌تواند اعتماد مدیران و کاربران را به تصمیمات مدل افزایش دهد.

- بررسی عملکرد مدل‌ها در شرایط محیطی وابسته به زمان و دینامیک: محدودیت اصلی این مطالعه ماهیت ایستایی داده‌هاست. توصیه می‌شود که تحقیقات آینده عملکرد مدل‌ها را در جریان داده‌ها و زمانی که تغییرات محیطی رخ می‌دهد برای افزایش کاربرد آن‌ها در شرایط دنیای واقعی بررسی کنند.
- یادگیری عمیق برای تجزیه‌وتحلیل داده‌های پیچیده‌تر: شبکه‌های عصبی عمیق، در درجه اول شبکه‌های عصبی مکرر (RNN) یا شبکه‌های مبتنی‌بر توجه[2]، می‌توانند به‌طور مؤثر برای تجزیه‌وتحلیل روابط بسیار پیچیده‌تر و زمانی که شامل ویژگی‌های مشتری است مورد استفاده قرار گیرند.
- ارزیابی تأثیر رویکردهای مبتنی بر داده‌های شخص ثالث و بانکداری باز: در شرایطی که بانک‌ها به داده‌های مشتریان از منابع شخص ثالث[3] دسترسی دارند، بررسی استفاده از این داده‌ها برای بهبود پیش‌بینی‌ها می‌تواند زمینه‌ای مهم برای تحقیقات آتی باشد.

منابع

امیرحسنخانی، حمیدرضا؛ طلوعی اشلقی، عباس؛ رادفر، رضا و پورابراهیمی، علیرضا (۱۴۰۳). ارائه یک مدل ترکیبی مبتنی بر یادگیری ماشینی برای طبقه‌بندی مشتریان مشترک صنعت بانکداری و بیمه. *فصلنامه مدیریت بهره‌وری*، ۱۸(۱)، ۱–۲۸.

جعفرنژاد چقوشی، احمد؛ رضاسلطانی، آرمان و خانی، امیرمحمد (۱۴۰۳). مقایسه مدل‌های یادگیری جمعی برای پیش‌بینی رتبه کشوری دانش‌آموزان در کنکور سراسری. *مدیریت صنعتی*، ۱۶(۳)، ۴۵۷-۴۸۱.

جعفری اسکندری، میثم و روحی، میلاد (۱۳۹۶). مدیریت ریسک اعتباری مشتریان بانکی با استفاده از روش ماشین بردار تصمیم بهبودیافته با الگوریتم ژنتیک با رویکرد داده‌کاوی. *مدیریت دارایی و تأمین مالی*، ۵(۴)، ۱۷–۳۲.

چهره، سپیده و سرآبادانی، علی (۱۴۰۲). ارائه مدلی مبتنی بر الگوریتم جنگل تصادفی و بهینه‌سازی جایا برای پیش‌بینی ریزش مشتریان بانکی. *مدیریت مهندسی و رایانش نرم*، ۹(۲)، ۱۳۲–۱۴۸.

خاشعی ورنامخواستی، وحید و فارسی، سجاد (۱۴۰۳). مدلی برای خلق و اجرای نوآوری دوسوتوان در صنعت بانکداری ایران. *مدیریت بازرگانی*، ۱۶(۴)، ۱۰۰۲– ۱۰۲۸.

خانلری، امیر؛ احراری، مهدی و میرپور، سمیه (۱۳۹۵). پیش‌بینی ارزش طول عمر مشتریان بانکی با استفاده از تکنیک دسته‌بندی گروهی داده‌ها (GMDH) در شبکه عصبی. *مدیریت بازرگانی*، ۸(۴)، ۸۳۳–۸۶۰.

خانی، امیرمحمد؛ کزازی، ابوالفضل و تقوی فرد، محمدتقی (۱۴۰۱). ارزیابی کیفیت خدمات معاونت فرهنگی و اجتماعی شهرداری تهران در حوزه فرهنگ و هنر و تأثیر آن بر رضایت شهروندان. *برنامه‌ریزی رفاه و توسعه اجتماعی*، ۱۳(۵۰)، ۲۰۵-۲۵۰.

1. Explainable AI
2. Attention-based Networks
3. Third-party Data

غلامیان، محمدرضا و مظفری، عظیمه (۱۳۹۵). پیش‌بینی ارزش مشتریان جدید بانک بر مبنای مدل آر.اف.ام با استفاده از درخت تصمیم بهبودیافته در راستای کاهش حداکثر حافظه مورد نیاز. *مطالعات مدیریت کسب‌وکار هوشمند*، ۵(۱۷)، ۹۳–۱۲۱.

موسوی، سید محسن و امیری عقدایی، سیدفتح‌اله (۱۳۹۹). شناسایی عناصر سازنده «ارزش پیشنهادی به مشتری» و تأثیر آن‌ها بر رضایت مشتری با استفاده از تحلیل احساسات بر مبنای متن‌کاوی. *مدیریت بازرگانی*، ۱۲(۴)، ۱۰۹۲–۱۱۱۶.

مهرگان، محمدرضا و خانی، امیرمحمد (۱۴۰۳). بهبود عملکرد سازمانی: نقش زنجیرهٔ تأمین ۴.۰ و تأمین مالی در کاهش ریسک زنجیرهٔ تأمین. *نشریه علمی پژوهشی مدیریت کسب‌وکارهای بین‌المللی*، ۷(۳)، ۳۹–۵۹.

نصیرالاسلامی، ابراهیم؛ صنیعی، احسان؛ عباسیان، عزت‌اله؛ فتح‌پور کاشانی، رضا و قیصری، نگین (۱۴۰۳). پیش‌بینی سپرده‌های بانکی به روش یادگیری ماشین. *فصلنامه مطالعات اقتصادی کاربردی ایران*، ۱۳(۵۰)، ۱۳۷–۱۶۷.

یعقوب‌پور، مریم؛ خداداد حسینی، سیدحمید؛ جنتی‌فر، حسین و ثانوی‌فرد، رسول (۱۴۰۳). طراحی مدل تعالی بازاریابی دیجیتالی (مطالعه موردی: بانکداری تجاری). *مدیریت بازرگانی*، ۱۶(۳)، ۵۹۶–۶۱۷.

References

Aldi, F., Nozomi, I., Sentosa, R. B., & Junaidi, A. (2023). Machine learning to identify monkey pox disease. *Sinkron: jurnal dan penelitian teknik informatika*, *7*(3), 1335-1347.

Amirhasankhan, H., Toloie Eshlaghy, A., Radfar, R. & Pourebrahimi, A. (2024). Presenting a Hybrid Model based on the Machine Learning for the Classification of Banking and Insurance Industry Common Customers, *Journal of Productivity Management,* 18(68), 53-80. magiran.com/p2710297 (*in Persian*)

Ashraf, R. (2024). Bank Customer Churn Prediction Using Machine Learning Framework. *Journal of Applied Finance & Banking,* 14(4), 65-109, 65–109. https://doi.org/10.47260/jafb/1445

Awad, M. & Khanna, R. (2015). Support vector machines for classification. In *Apress eBooks* (pp. 39–66). https: //doi.org/10.1007/978-1-4302-5990-9_3

Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https: //doi.org/10.1007/s10462-020-09896-5

Chehreh, S. & Sarabadani, A. (2024). A model based on random forest algorithm and Jaya optimization to predict bank customer churn. *Journal of Engineering Management and Soft Computing*, *9*(2), 132-148. doi: 10.22091/jemsc.2024.9541.1174 (*in Persian*)

Conn, D., Ngun, T., Li, G. & Ramirez, C. M. (2019). Fuzzy Forests: Extending random Forest feature selection for Correlated, High-Dimensional Data. *Journal of Statistical Software*, *91*(9). https: //doi.org/10.18637/jss.v091.i09

Dewi, O. I. P., Santiko, V. N., Safa, S. W. P., Gunawan, A. A. S. & Setiawan, K. E. (2024). Machine Learning for Imbalanced Data in Telecom Churn Classification. *2024 International Conference on Information Technology Research and Innovation (ICITRI)*, 30–35. https: //doi.org/10.1109/icitri62858.2024.10699226

Dias, J., Godinho, P. & Torres, P. (2020). Machine learning for customer churn prediction in retail banking. In *Lecture notes in computer science* (pp. 576–589). https: //doi.org/10.1007/978-3-030-58808-3_42

Dorogush, A. V., Ershov, V. & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *arXiv (Cornell University)*. https: //doi.org/10.48550/arxiv.1810.11363

Galal, M., Rady, S. & Aref, M. (2022). Enhancing Customer Churn Prediction in Digital Banking using Ensemble Modeling. *2022 4th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, *1*, 21–25. https: //doi.org/10.1109/niles56402.2022.9942408

Gholamian, M. & Mozafari, A. (2016). Prediction of Bank Customer Value based on RFM Model Using Improved Decision Tree to Reduce the Maximum Required Memory. *Business Intelligence Management Studies*, *5*(17), 93-121. https://doi.org/10.22054/ims.2016.6993 (*in Persian*)

Guido, R., Ferrisi, S., Lofaro, D. & Conforti, D. (2024). An Overview on the advancements of support vector machine models in healthcare Applications: a review. *Information*, *15*(4), 235. https: //doi.org/10.3390/info15040235

Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S. & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, *11*(1). https: //doi.org/10.1186/s40537-024-00973-y

Hodge, V. & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, *22*(2), 85-126.

Hu, N. W., Hu, N. W. & Maybank, S. (2008). ADABoost-Based Algorithm for network Intrusion Detection. *IEEE Transactions on Systems Man and Cybernetics Part B (Cybernetics)*, *38*(2), 577–583. https: //doi.org/10.1109/tsmcb.2007.914695

Ilyas, S., Zia, S., Muneer, U., Letchmunan, S. & Un, Z. (2020). Predicting the Future Transaction from Large and Imbalanced Banking Dataset. *International Journal of Advanced Computer Science and Applications*, *11*(1). https: //doi.org/10.14569/ijacsa.2020.0110134

Iranzad, R. & Liu, X. (2024). A review of random forest-based feature selection methods for data science education and applications. *International Journal of Data Science and Analytics*. https: //doi.org/10.1007/s41060-024-00509-w

Jafari Eskandari, M. & Rohii, M. (2017). Credit Risk Management of Banking Customers Using Support Vector Machine Optimized by Genetic Algorithm with Data Mining Approach. *Journal of Asset Management and Financing*, *5*(4), 17-32. doi: 10.22108/amf.2017.21191 (*in Persian*)

Jafarnejad, A., Rezasoltani, A. & Khani, A.M. (2024). Unleashing the Power of Ensemble Learning: Predicting National Ranks in Iran's University Entrance Examination. *Industrial Management Journal,* 16(3), 457-481. https: //doi.org/10.22059/imj.2024.381521.1008178 (*in Persian*)

Kaisar, S. & Sifat, S. T. (2023). Explainable Machine Learning Models for Credit Risk Analysis: A Survey. In *Data Analytics for Management, Banking and Finance* (pp. 51–72). https: //doi.org/10.1007/978-3-031-36570-6_2

Khani, A. M., Kazazi, A. Taqhavi Fard, M. T. (2022). Evaluating the quality of services of the cultural and social deputy of Tehran municipality in the field of culture and art. *Social Development & Welfare Planning*, *13*(50), 205-250. https://doi.org/10.22054/qjsd.2021.58035.2110 (*in Persian*)

Khanlari, A., Ahrari, M. & Mirpoor, S. (2017). Predicting Customer Lifetime Value Based on Financial and Demographic Characteristics Using GMDH Neural Network Case Study: Individual Customers of a Private Bank of Iran. *Journal of Business Management*, *8*(4), 833-860. https://doi.org/10.22059/jibm.2017.61302 (*in Persian*)

Khashei Varnamkhasti, V. & Farsi, S. (2024). A Model for Creating and Implementing Ambidextrous Innovation in Iranian Banking. *Journal of Business Management*, *16*(4), 1002-1028. https://doi.org/10.22059/jibm.2023.364184.4642 (*in Persian*)

Kotsiantis, S. B. (2011). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*(4), 261–283. https: //doi.org/10.1007/s10462-011-9272-4

Kruse, R., Mostaghim, S., Borgelt, C., Braune, C. & Steinbrecher, M. (2022). Multi-layer perceptrons. In *Texts in computer science* (pp. 53–124). https: //doi.org/10.1007/978-3-030-42227-1_5

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A. & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, *5*(1). https: //doi.org/10.1186/s40537-018-0151-6

Luong, H. H., Tran, T. T., Van Nguyen, N., Le, A. D., Nguyen, H. T. T., Nguyen, K. D., Tran, N. C. & Nguyen, H. T. (2022). Feature Selection Using Correlation Matrix on Metagenomic Data with Pearson Enhancing Inflammatory Bowel Disease Prediction. In *Lecture notes in electrical engineering* (pp. 1073–1084). https: //doi.org/10.1007/978-981-16-2183-3_102

Mastelini, S. M., Nakano, F. K., Vens, C. & De Leon Ferreira De Carvalho, A. C. P. (2022). Online Extra Trees Regressor. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(10), 6755–6767. https: //doi.org/10.1109/tnnls.2022.3212859

Mehregan, M. R. & Khani, A. M. (2024). Improving organizational performance: the role of supply chain 4.0 and financing in reducing supply chain risk. *Journal of International Businesses Administration*, *7*(3), 39-59. doi: 10.22034/jiba.2024.60005.2164 (*in Persian*)

Mienye, I. D. & Jere, N. (2024). A survey of Decision trees: Concepts, algorithms, and applications. *IEEE Access*, *12*, 86716–86727. https: //doi.org/10.1109/access.2024.3416838

Mousavi, S. M. & Amiri Aghdaie, S. F. (2021). Identifying the Constructive Elements of "Value Proposition" and their Impact on Customers' Satisfaction using Sentiment Analysis based on Text Mining. *Journal of Business Management*, *12*(4), 1092-1116. https://doi.org/10.22059/jibm.2020.302987.3847 (*in Persian*)

Mujahid, M., Kına, E., Rustam, F., Villar, M. G., Alvarado, E. S., De La Torre Diez, I. & Ashraf, I. (2024). Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering. *Journal of Big Data*, *11*(1). https: //doi.org/10.1186/s40537-024-00943-4

Muneer, A., Ali, R. F., Alghamdi, A., Taib, S. M., Almaghthawi, A. & Ghaleb, E. a. A. (2022). Predicting customers churning in banking industry: A machine learning approach. *Indonesian Journal of Electrical Engineering and Computer Science*, *26*(1), 539. https: //doi.org/10.11591/ijeecs.v26.i1.pp539-549

Nasiroleslami, E., Saniee, E., Abbasian, E., Fathpour Kashani, R. & Gheysari, N. (2024). Prediction of Bank Deposits by Machine Learning Method. *Journal of Applied Economics Studies in Iran*, *13*(50), 137-167. doi: 10.22084/aes.2023.27444.3564 (*in Persian*)

Nguyen, Q., Nguyen, H., Le, D. T. & Bui, Q. (2022). Fine-Tuning LightGBM using an artificial Ecosystem-Based optimizer for forest fire analysis. *Forest Science*, *69*(1), 73–82. https: //doi.org/10.1093/forsci/fxac039

Peng, K., Peng, Y. & Li, W. (2023). Research on customer churn prediction and model interpretability analysis. *PLoS ONE*, *18*(12), e0289724. https: //doi.org/10.1371/journal.pone.0289724

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Neural Information Processing Systems*, *31*, 6639–6649. https: //papers.nips.cc/paper/7898-catboost-unbiased-boosting-with-categorical-features.pdf

Przybyła-Kasperek, M. & Marfo, K. F. (2024). A multi-layer perceptron neural network for varied conditional attributes in tabular dispersed data. *PLoS ONE*, *19*(12), e0311041. https: //doi.org/10.1371/journal.pone.0311041

Sanmorino, A., Marnisah, L. & Sunardi, H. (2023). Feature selection using Extra Trees Classifier for Research Productivity Framework in Indonesia. In *Lecture notes in electrical engineering* (pp. 13–21). https: //doi.org/10.1007/978-981-99-0248-4_2

Schonlau, M. & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal Promoting Communications on Statistics and Stata*, *20*(1), 3–29. https: //doi.org/10.1177/1536867x20909688

Shingi, G. (2020). A federated learning based approach for loan defaults prediction. *2021 International Conference on Data Mining Workshops (ICDMW)*, 362–368. https: //doi.org/10.1109/icdmw51313.2020.00057

Siddiqui, N., Haque, M. A., Khan, S. M. S., Adil, M. & Shoaib, H. (2024). Different ML-based strategies for customer churn prediction in banking sector. *Journal of Data Information and Management*, *6*(3), 217–234. https: //doi.org/10.1007/s42488-024-00126-z

Syriopoulos, P. K., Kalampalikis, N. G., Kotsiantis, S. B. & Vrahatis, M. N. (2023). kNN Classification: a review. *Annals of Mathematics and Artificial Intelligence*. https: //doi.org/10.1007/s10472-023-09882-x

Tékouabou, S. C. K., Gherghina, Ș. C., Toulni, H., Mata, P. N. & Martins, J. M. (2022). Towards explainable machine learning for bank churn prediction using data balancing and Ensemble-Based methods. *Mathematics*, *10*(14), 2379. https: //doi.org/10.3390/math10142379

Tolles, J. & Meurer, W. J. (2016). Logistic regression. *JAMA*, *316*(5), 533. https: //doi.org/10.1001/jama.2016.7653

Tran, H. D., Le, N. & Nguyen, V. (2023). Customer churn prediction in the banking sector using Machine Learning-Based classification models. *Interdisciplinary Journal of Information Knowledge and Management*, *18*, 087–105. https: //doi.org/10.28945/5086

Vittinghoff, E. & McCulloch, C. E. (2006). Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology*, *165*(6), 710–718. https: //doi.org/10.1093/aje/kwk052

Vu, V. (2024). An efficient customer churn prediction technique using combined machine learning in commercial banks. *Operations Research Forum*, *5*(3). https: //doi.org/10.1007/s43069-024-00345-5

Wyner, A. J., Olson, M., Bleich, J. & Mease, D. (2017). Explaining the success of adaboost and random forests as interpolating classifiers. *Journal of Machine Learning Research*, *18*(1), 1558–1590. https: //doi.org/10.5555/3122009.3153004

Yaghoubpour, M., Khodadad Hosseini, S. H., Janatifar, H. Sanavifard, R. (2024). Designing a Digital Marketing Excellence Model (A Case Study of Commercial Banking). *Journal of Business Management*, *16*(3), 596-617. https://doi.org/10.22059/jibm.2024.376605.4785 (*in Persian*)

Yang, H., Chen, Z., Yang, H. & Tian, M. (2023). Predicting coronary heart disease using an improved LightGBM Model: Performance analysis and comparison. *IEEE Access*, *11*, 23366–23380. https: //doi.org/10.1109/access.2023.3253885