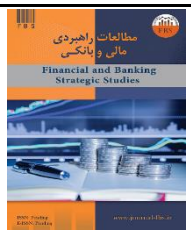
E-ISSN: 2981-1589 | P-ISSN: 2981-1570

Financial and Banking Strategic Studies

www.journal-fbs.com

Fin. Bank. Stra. Stud. Vol. 3, No. 2 (2025) 93–108.



Paper Type: Original Article

Customer Churn Prediction in Banks Using Machine Learning Algorithms

Seyyed Mohsen Hosseini*

Member of Bank Sepah Scholars Club; mohsen.hssn419@gmail.com.

Citation:

Hosseini, S. M. (2025). Customer churn prediction in banks using machine learning algorithms. *Financial and banking strategic studies, 3*(2), 93-108.

Received: 25/08/2024 Reviewed: 28/10/2024 Revised: 09/12/2024 Accepted: 02/01/2025

Abstract

Purpose: In the banking industry, retaining loyal customers is considerably more cost-effective and profitable than acquiring new ones. Customer churn remains a major challenge for banks, directly reducing profitability, increasing marketing expenditures, and lowering market share. This study evaluates the performance of machine learning algorithms for predicting customer churn across branches of a state-owned bank in Iran between 2021 and 2024. By focusing on customer retention and minimizing the costs of attrition, the study aims to develop an efficient, interpretable model to identify customers at risk of churn.

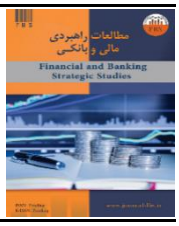
Methodology: This descriptive-analytical, retrospective study analyzed data from 2,025 active customers over 4 years. For each customer, 12 features covering transactional, behavioral, and demographic characteristics were collected. Following data cleaning, z-score normalization was applied. Several machine learning algorithms (including Decision Tree, Random Forest, Support Vector Machine, Multilayer Perceptron, Bayesian Network, and XGBoost (were implemented in R. Their performance was assessed through 10-fold cross-validation based on accuracy, sensitivity, and specificity.

Findings: Among the 2,025 customers examined, 325 (16%) were identified as churners. Statistical tests revealed no significant differences between churners and non-churners in age, relationship duration with the bank, or average deposits over the past six months. Among the models tested, XGBoost demonstrated superior performance with an accuracy of 96.89%, sensitivity of 87.11%, and specificity of 98.71%. The area under the ROC curve (AUC) for this model was 0.9907, indicating excellent discriminatory power.

Originality/Value: The contribution of this study lies in integrating advanced machine learning techniques with rigorous statistical analysis using real-world banking data. To the best of our knowledge, this is among the few studies to systematically compare multiple ML algorithms within the Iranian banking context, emphasizing both interpretability and robust validation. The findings provide practical insights for banking policymakers to design proactive strategies to improve customer retention.

Keywords: Attrition, Customer churn, Machine learning, XGBoost model.

Corresponding Author: mohsen.hssn419@gmail.com 10.22105/fbs.2025.542570.1169



مطالعات راهبردی مالی و بانکی

دوره ۳، شماره (۲)، (۱۴۰۴)، ۱۰۸-۹۳

www.journal-fbs.com

نوع مقاله: پژوهشی

پیش بینی ریزش مشتریان بانک با استفاده از الگوریتم های یادگیری ماشین

سید محسن حسینی*

عضو باشگاه پژوهان بانک سپه.

چکیده

هدف: در صنعت بانکداری، حفظ مشتریان وفادار به‌مراتب کم‌هزینه‌تر و سودآورتر از جذب مشتریان جدید است. روی‌گردانی مشتریان به‌عنوان یکی از چالش‌های اصلی بانک‌ها، تاثیر مستقیمی بر کاهش سودآوری، افزایش هزینه‌های بازاریابی و افت سهم بازار دارد. پژوهش حاضر با هدف ارزیابی عملکرد الگوریتم‌های یادگیری ماشین در پیش‌بینی روی‌گردانی مشتریان شعب یک بانک دولتی در سال‌های ۱۴۰۰ تا ۱۴۰۳ انجام شده است. با توجه به اهمیت حفظ مشتریان وفادار و کاهش هزینه‌های ناشی از ریزش مشتریان، این مطالعه تلاش دارد مدلی کارآمد و تفسیرپذیر برای شناسایی مشتریان در معرض ریزش ارایه دهد.

روش‌شناسی پژوهش: مطالعه حاضر از نوع توصیفی-تحلیلی و گذشته‌نگر است. داده‌های مربوط به ۲۰۲۵ مشتری فعال طی دوره ۱۴۰۰ تا ۱۴۰۳ گردآوری شد. برای هر مشتری ۱۲ ویژگی شامل مشخصات تراکنشی، رفتاری و جمعیت‌شناختی ثبت گردید. داده‌ها پس از پاک‌سازی، با استفاده از روش *z-score* نرمال‌سازی شدند. سپس با پیاده‌سازی الگوریتم‌های مختلف یادگیری ماشین شامل درخت تصمیم، جنگل تصادفی، ماشین بردار پشتیبان، شبکه عصبی پرسپترون چندلایه، شبکه بیزین و *XGBoost* در محیط *R*، عملکرد مدل‌ها با استفاده از روش اعتبارسنجی متقاطع ۱۰تایی و بر مبنای معیارهای صحت، حساسیت و ویژگی مقایسه شد.

یافته‌ها: از بین 2025 مشتری بررسی‌شده، تعداد ۳۲۵ نفر معادل ۱۶%به‌عنوان مشتریان روی‌گردان شناسایی شدند. بررسی آماری نشان داد متغیرهای سن، مدت زمان رابطه با بانک و میانگین سپرده‌ها در ۶ ماه گذشته بین دو گروه تفاوت معناداری ندارند. مدل *XGBoost* با صحت ۹۶/۸۹% حساسیت ۸۷/۱۱% و ویژگی ۹۸/۷۱% بالاترین عملکرد را نسبت به سایر الگوریتم‌ها نشان داد. همچنین سطح زیر منحنی نمودار مشخصه عملکرد برای این مدل برابر با ۰/۹۹۰۷ محاسبه شد که بیانگر دقت بسیار بالا در طبقه‌بندی است.

اصالت/ارزش‌افزوده علمی: در این پژوهش ویژگی‌سازی (*Feature Engineering*) خاص بانکی انجام شده است. متغیرهای جدید از تراکنش‌ها یا رفتار مشتری استخراج‌شده که در مقالات مشابه کمتر استفاده شده است مانند تغییر تعداد تراکنش‌های سه‌ماهه چهارم به سه‌ماهه اول و ... ترکیب رویکردهای پیشرفته یادگیری ماشین و استفاده از داده‌های مربوط به مشتریان یکی از بانک‌های ایران ضرورت و اهمیت این پژوهش را بیشتر نمایان می‌سازد. همچنین این مطالعه یکی از معدود پژوهش‌هایی است که عملکرد چندین الگوریتم *ML* را در محیط بانکی ایران با بهره‌گیری از تحلیل تفسیرپذیر و اعتبارسنجی دقیق مقایسه می‌کند. نتایج آن می‌تواند به سیاست‌گذاران بانکی در طراحی اقدامات پیشگیرانه برای حفظ مشتریان کمک شایانی کند.

کلیدواژه‌ها: روی‌گردانی مشتری، ریزش مشتری، مدل *XGBoost*، یادگیری ماشین.

۱- مقدمه

در فضای رقابتی امروز صنعت بانکداری، حفظ مشتریان موجود به‌عنوان یک مزیت رقابتی کلیدی شناخته می‌شود. تغییرات سریع در رفتار مشتریان، ظهور بانک‌های دیجیتال و گسترش فناوری‌های نوین پرداخت، موجب شده‌اند تا بانک‌ها بیش از گذشته در معرض ریزش مشتری قرار گیرند. روی‌گردانی مشتری به معنای قطع تعامل فعال مشتری با بانک و انتقال خدمات مالی به رقباست. این پدیده می‌تواند اثرات نامطلوبی بر

درآمد، شهرت و رشد بلندمدت بانک‌ها داشته باشد. مطالعات نشان داده‌اند که جذب مشتری جدید به‌طور متوسط پنج تا هفت برابر پرهزینه‌تر از حفظ مشتریان فعلی است و حتی کاهش یک درصدی در نرخ روی‌گردانی می‌تواند تا شش درصد سودآوری بانک را افزایش دهد [1].

مساله وفادار نگه‌داشتن مشتریانی که در بازارهای بسیار رقابتی امروزی، گزینه‌های زیادی را پیش رو دارند و می‌توانند انتخاب‌های زیادی داشته باشند بسیار حایز اهمیت است. یک پایگاه داده که شامل انتخاب‌های مشتری و پروفایل‌های شخصی آن‌ها باشد می‌تواند در حل این مساله مورد استفاده قرار گیرد. الگوهای رفتاری مشتریان می‌تواند برای تشخیص ویژگی‌های هر مشتری و اینکه به کدام محصول رغبت بیشتری نشان می‌دهند مورد استفاده قرار گرفته و مواردی از این دست سبب وفادار ماندن مشتری خواهد شد. موسسات مالی برای چندین دهه راهبردهای متمرکز بر تولید و معاملات را دنبال می‌کردند و چندان بر شیوه ارتباط با مشتریان تمرکز نداشتند. با رشد فناوری و توسعه عوامل رقابتی، نیاز بنگاه‌های اقتصادی به ایجاد و حفظ ارتباط موثر با مشتریان بیش‌ازپیش نمود یافته است و بانک‌ها در بازار رقابتی با سایر بانک‌ها و موسسات مالی باید به شناخت صحیح از مشتریان خود دست یابند. هدف از شناسایی مشتریان، ایجاد تمایز و تشخیص با ارزش‌ترین آن‌ها و اقدام برای نگهداری و جذب آن‌هاست؛ ازاین‌رو مدیریت ارتباط با مشتریان ابزاری مهم و اثرگذاری در رقابت بین بانک‌ها به‌منظور ارایه خدمات بهینه و جذب مشتریان جدید است [2]. مدیریت ارتباط با مشتری زیرساختی است که ارزش مشتری را آشکار می‌کند و افزایش می‌دهد. برای داشتن مدیریت ارتباط موثر با مشتری، جمع‌آوری اطلاعات درباره ارزش مشتری و بخش‌بندی آنان به‌منظور پاسخگویی به نیازهای منحصربه‌فرد هر بخش ضروری است. بازار رقابتی امروزه به‌سرعت در حال تغییر و تحول است و ویژگی‌های خاصی از قبیل تکرار خرید مشتریان در بازه‌های زمانی، حجم بالای مشتریان، اطلاعات باارزش از رفتار خرید مشتریان و ... دارد. در چنین بازارهایی، هدف مدیریت ارتباط با مشتری، درک و پیش‌بینی الگوی خرید و شناسایی نیازهای مشتریان و عرضه متناسب با خواسته و انتظارات مشتری است؛ ازاین‌رو مدیریت ارتباط با مشتری، پیش‌نیازی برای فعالیت‌های بازاریابی از قبیل هدف‌گذاری بخش‌های مشتریان پیاده‌سازی می‌شود [3]. با رشد فناوری اطلاعات، افزایش رقابت بین بانک‌ها و ارایه خدمات در قالب‌های نوین بانکداری الکترونیک، احتمال ریزش مشتریان افزایش یافته است. از سوی دیگر تاثیر عوامل محیطی و روان‌شناختی مانند تبلیغات، ارایه خدمات نوین و... موجب شده است رفتار مشتری در برخی شرایط ثبات نداشته باشد و بانک‌ها در تحلیل و پیش‌بینی رفتار مشتریان با عدم قطعیت مواجه شوند؛ بنابراین باید برای شناخت بهتر نیازها و پیش‌بینی دقیق رفتار مشتری، ماهیت پویای رفتار آن‌ها را بررسی کرد.

با توجه به این واقعیت، بانک‌ها به‌دنبال بهره‌گیری از روش‌های علمی و داده‌محور برای شناسایی مشتریانی هستند که در معرض ریزش قرار دارند. استفاده از داده‌های رفتاری، مالی و خدماتی مشتریان و تحلیل آن‌ها با روش‌های داده‌کاوی و یادگیری ماشین، می‌تواند ابزار قدرتمندی برای پیش‌بینی روی‌گردانی فراهم آورد. این الگوریتم‌ها قادرند الگوهای پیچیده و پنهان در داده‌های مشتریان را شناسایی کرده و خروجی‌هایی با دقت بالا ارایه دهند. پیش‌بینی روی‌گردانی مشتریان به‌عنوان یکی از مهم‌ترین کاربردهای داده‌کاوی در حوزه بازاریابی رابطه‌مند و مدیریت ارتباط با مشتری[1]، در سال‌های اخیر توجه محققان بسیاری را به خود جلب کرده است. سازمان‌ها برای پیروزی در میدان رقابت نیازمند شناخت رفتار مشتریان خود هستند تا بتوانند برای نگهداری آن‌ها زودتر از دیگران خواسته‌ها و رفتارهای آنان را پیش‌بینی کنند. هدف از تحلیل رویگردانی، شناسایی مشتریانی است که احتمال می‌رود در آینده‌ای نزدیک سازمان را ترک کنند و از خدمات رقیب استفاده نمایند. تحلیل رویگردانی می‌تواند به مدیران کمک کند تا دلایل رویگردانی مشتری را درک کنند و از این طریق روابط خود با مشتری را بهبود دهند. حفظ مشتریان قدیمی به‌خصوص در برخی از خدمات که فروش یک‌باره ندارند، به‌جز هزینه‌های جذب مشتری، ارزش فرصت را نیز برای شرکت به همراه دارد. به این معنا که سازمان قادر به ارایه خدمات اضافی و جدید به مشتری و کسب درآمد بیشتر است. به همین دلیل، از دست دادن مشتری نه‌تنها منجر به کاهش درآمد و تحمیل هزینه جذب مشتری جدید به سازمان می‌شود بلکه باعث از دست رفتن درآمدهای بالقوه نیز می‌شود. در منابع مختلف تعریف‌های متفاوتی از رویگردانی مشتری ارایه شده است که برخی از آن‌ها عبارت‌اند از:

۱. رویگردانی زمانی اتفاق می‌افتد که مشتری از یک تامین کننده خدمات به دیگری مراجعه کند [4].
۲. وقتی مشتری ناراضی است یا گزینه‌های بهتری می‌بیند، به یک بانک رقیب مراجعه می‌کند [5].

[1] Customer Relationship Management (CRM)

در مطالعات بانکداری، مشتریان رویگردان (*churn*) به افرادی اطلاق می‌شوند که رابطه بانکی خود را خاتمه داده‌اند (برای مثال با بستن حساب بانکی) یا سطح تعامل آن‌ها (مانند تعداد تراکنش‌ها) به‌طور معناداری کمتر از یک آستانه مشخص است. آستانه‌ی ثابت برای فعالیت مشتری از منظر تعداد تراکنش بوده که از طریق قوانین کسب‌وکار تعیین می‌شود. به‌طوری‌که اگر تعداد تراکنش‌های مشتری کمتر از این حد معین باشد، او به‌عنوان مشتری رویگردان در نظر گرفته می‌شود [6]. در سال‌های اخیر، مطالعات متعددی در حوزه پیش‌بینی روی‌گردانی مشتریان با استفاده از الگوریتم‌های یادگیری ماشین انجام شده است. به‌عنوان نمونه، عسگری و همکاران [7] با استفاده از ترکیب خوشه‌بندی سلسله‌مراتبی، مدل زنجیره مارکوف و الگوریتم درخت تصمیم، مدلی برای طبقه‌بندی مشتریان بانک در وضعیت‌های «فعال»، «نیمه‌فعال»، «در آستانه ریزش» و «ریزش‌یافته» ارایه کردند. نتایج مطالعه آن‌ها بر روی بیش از ۳۸۰ هزار رکورد مشتری نشان داد که بیش از ۷۹% مشتریانی که به وضعیت "ریزش‌یافته" می‌رسند، دیگر به سیستم بانکی بازنمی‌گردند. در مطالعه‌ای دیگر، مطیعی و محمدی [8] با بهره‌گیری از شبکه عصبی مصنوعی[1] و تحلیل توضیحات شاپلی جمع‌پذیر برای تفسیر خروجی مدل، به دقت ۸۸٫۳% در پیش‌بینی ریزش مشتریان یکی از بانک‌های خصوصی ایران دست یافتند. آن‌ها نشان دادند که کاهش تعامل با خدمات الکترونیکی، افت میانگین موجودی و کاهش تنوع محصولات استفاده‌شده از جمله مهم‌ترین متغیرهای تاثیرپذیر از فرآیند رویگردانی مشتریان هستند و می‌توانند به‌عنوان نشانه‌های رفتاری این پدیده موردتوجه قرار گیرند. این مطالعه نشان داد که استفاده از روش‌های تفسیرپذیر مانند *SHAP* می‌تواند تصمیم‌گیری مدیریتی را تسهیل کند.

در سطح بین‌المللی نیز مدل‌هایی مانند *XGBoost* و *CatBoost* به‌دلیل توانایی بالا در یادگیری غیرخطی، دقت پیش‌بینی بالا و امکان ترکیب با روش‌های تحلیل تفسیرپذیر، در حوزه پیش‌بینی روی‌گردانی موفق عمل کرده‌اند. ژانگ و همکاران [9] همچنین لاندبرگ لی [10] با معرفی چارچوب *SHAP*، زمینه تفسیر خروجی مدل‌های پیچیده را فراهم کردند و کاربرد آن را در حوزه‌های مالی و بازاریابی به‌طور گسترده توسعه دادند. از اوایل دهه ۱۹۸۰ مفهوم مدیریت ارتباط در حوزه بازاریابی اهمیت پیدا کرده است. به‌منظور مدیریت موثر ارتباط با مشتری، جمع‌آوری اطلاعات درباره ارزش مشتری دارای اهمیت است؛ به‌طوری‌که می‌توان گفت قوی‌ترین ابزارهای کاربردی برای بازاریابی، پیش‌بینی رفتار خرید و بخش‌بندی مشتریان است که امکان تفکیک خریداران از غیرخریداران و شناسایی گروه‌های مشتریان را فراهم کرده آن‌ها را از یکدیگر متمایز می‌کند. بخش‌بندی مشتریان به شناسایی مشتریان با مشخصه‌های مشابه اطلاق می‌شود و بازاریابان برای هدف‌گذاری موثر و تخصیص بهینه منابع از آن استفاده می‌کنند. در بسیاری از پژوهش‌های پیشین در حوزه بخش‌بندی مشتریان، فرض بر این بوده است که بازار پایدار و رفتار مشتری در طول زمان تغییر نمی‌کند؛ بر همین اساس، بخش‌های مشتریان ثابت در نظر گرفته شده و تعلق افراد به این بخش‌ها تغییرناپذیر فرض شده است. با این حال، در شرایطی مانند اقتصاد ایران که ثبات و پایداری بازار همواره با چالش‌های جدی مواجه است، این فرض نمی‌تواند بازتاب‌دهنده واقعیت باشد؛ زیرا نیازها، ترجیحات و رفتار مشتریان تحت تاثیر عوامل روانی-اجتماعی و محیطی در طول زمان تغییر می‌کند. برای رفع این محدودیت، در این پژوهش راهکار ما مبتنی بر تحلیل پویا بوده است؛ به‌گونه‌ای که به‌جای در نظر گرفتن بخش‌های ثابت، الگوهای جابه‌جایی و انتقال مشتریان بین بخش‌های مختلف در طول زمان استخراج و بررسی شده است. این رویکرد باعث می‌شود بتوان تغییرات رفتاری مشتریان را در شرایط بازار ناپایدار نیز ردیابی و مدل‌سازی کرد و در نتیجه نتایج تحقیق از انطباق بیشتری با واقعیت برخوردار باشد. بررسی سایر مطالعات، نشان می‌دهد پژوهش‌های انجام‌شده در این حوزه هر یک در نوع خود به بررسی جوانب مختلفی پرداخته و چندان چارچوب مفهومی جامعی در حوزه بخش‌بندی پویای مشتریان وجود ندارد؛ به‌طوری‌که در بخش مدل‌سازی و پیاده‌سازی تجربی، خلا تحقیقاتی زیادی وجود دارد [17–11] جابه‌جایی و انتقالات مشتریان بین بخش‌های مختلف و استخراج الگوهای غالب آن‌ها را بررسی و مطالعه کرده‌اند. به این ترتیب تعداد بسیار اندکی هر دو رویکرد را هم‌زمان مدنظر قرار داده، اما بررسی تغییرات مشتریان را فقط به تعداد مشتریانی که در طول زمان در همان بخش باقی ماندند، معطوف کرده است. ازآنجاکه تمرکز این پژوهش بر استخراج الگوهای جابه‌جایی و انتقالات مشتریان بین بخش‌های مختلف در طول زمان است، در ادامه مطالعات انجام‌شده در این زمینه بررسی می‌شود. در زمینه پیش‌بینی انتقالات و جابه‌جایی مشتری بین بخش‌های مختلف نیز مطالعاتی انجام شده است که بیشتر آن‌ها از زنجیره مارکوف برای مدل‌سازی و پیش‌بینی استفاده کرده‌اند. مطالعه همبرگ و تتزک [18] یکی از موارد مهم در این زمینه است که در بهینه‌سازی سبد مشتری از زنجیره مارکوف برای پیش‌بینی بخش‌های مشتریان استفاده کرده‌اند. همچنین در مطالعات لمنس، کروکس و استریمرسچ [14] از زنجیره مارکوف در این زمینه استفاده شده است.

[1] Artificial Neural Networks (ANN)

در پژوهش ژانگ و همکاران [9]، از دو الگوریتم داده‌کاوی به‌منظور توسعه یک مدل پیش‌بینی ریزش مشتریان استفاده‌کننده از کارت‌های اعتباری بهره گرفته شد. در این تحقیق که از اطلاعات پایگاه داده یک بانک چینی استفاده شده است، از ۴ دسته متغیر اطلاعات مشتری، اطلاعات کارت، داده‌های مربوط به ریسک مشتری و اطلاعات مربوط به تراکنش‌ها استفاده شده است که شامل ۱۳۵ متغیر مختلف می‌شوند. به‌جای استفاده از تمام این متغیرها، تعداد ۹۵ متغیر با توجه به همبستگی بین آن‌ها برای انجام مراحل بعدی انتخاب شدند [19]. زی و همکاران [20] با استفاده از مدل توسعه‌یافته متوازن به دسته‌بندی مشتریان مربوط به یک بانک بزرگ چینی پرداختند. گرچه مدل‌های پیش‌بینی‌کننده بر پایه داده‌های تاریخی توسعه می‌یابند و از نظر الگوریتمی در صورت مشابهت آماری داده‌ها می‌توانند در مقیاس‌های بزرگ‌تر نیز عملکرد مشابهی داشته باشند، پیاده‌سازی آن‌ها در سامانه‌های زنده بانکی تنها به صحت مدل محدود نمی‌شود. در محیط عملیاتی، چالش‌هایی مانند کیفیت و به‌روز بودن داده‌ها، تأخیر در جریان داده‌های واقعی، الزامات امنیتی، محدودیت‌های زیرساختی و نیاز به تفسیرپذیری برای تصمیم‌گیرندگان مطرح می‌شود که خارج از دامنه مدل‌سازی صرف است. ازاین‌رو، تمرکز این پژوهش بر توسعه و ارزیابی مدل با داده‌های تاریخی است، اما به‌کارگیری عملی آن در سامانه‌های زنده بانک‌ها نیازمند ملاحظات فنی و سازمانی تکمیلی مانند یکپارچه‌سازی با سامانه‌های عملیاتی، کنترل دسترسی و نظارت بر عملکرد مدل در زمان اجرا است. بر این اساس، پژوهش حاضر با هدف طراحی یک مدل داده‌محور برای پیش‌بینی روی‌گردانی مشتریان یک بانک دولتی انجام شده است. در این پژوهش، با استفاده از داده‌های موجود مشتریان شامل اطلاعات تراکنشی، دموگرافیک و رفتاری، ابتدا وضعیت فعلی مشتریان تحلیل شده و سپس مدل یادگیری ماشین *XGBoost* جهت پیش‌بینی رفتار آینده آن‌ها آموزش داده شده است. در ادامه، با بهره‌گیری از چارچوب *SHAP*، مهم‌ترین ویژگی‌های موثر در ریزش مشتری شناسایی شده‌اند تا زمینه برای تدوین راهبردهای هدفمند بازاریابی و حفظ مشتری فراهم گردد.

جدول ۱- مقایسه مطالعات پیشین.
Table 1- Comparison of previous studies.

ردیف	منابع	نوع داده‌ها	الگوریتم مورد استفاده	ابزار تفسیرپذیری	دقت مدل(%)	شاخص‌های ریزش مشتری
1	[7]	تراکنشی، تعداد خدمات، سابقه کاربری	خوشه‌بندی + زنجیره مارکوف + درخت تصمیم	ندارد	-	کاهش تراکنش، کاهش تعداد خدمات، افت فعالیت
2	[8]	تراکنشی، کانال‌های دیجیتال، دموگرافیک	شبکه عصبی مصنوعی (ANN)	SHAP	88.3	افت استفاده از خدمات دیجیتال، کاهش موجودی و تنوع خدمات
3	[9]	تراکنشی، زمان‌بندی تعامل، موجودی	XGBoost	SHAP	91.4	افت ناگهانی تراکنش‌ها، کاهش تعامل دیجیتال، افت موجودی
4	[10]	داده‌های عمومی آزمایشی	مدل‌های مختلف (عمومی)	SHAP	-	تحلیل تئوریک اهمیت ویژگی‌ها در خروجی مدل‌ها

۲- روش تحقیق

در این پژوهش، چند الگوریتم یادگیری ماشین برای پیش‌بینی رویگردانی مشتریان بانک مورد استفاده قرار گرفتند. هر الگوریتم با توجه به ویژگی‌های داده و توانایی آن در شناسایی الگوهای رفتاری مشتریان انتخاب شد. در ادامه، گام‌های اصلی هر الگوریتم به‌طور مختصر توضیح داده شده است.

۲-۱- درخت تصمیم (Decision Tree)

یکی از روش‌های پرکاربرد در یادگیری ماشین و داده‌کاوی است که برای انجام وظایف طبقه‌بندی و رگرسیون به‌کار می‌رود. در این روش، فرآیند تصمیم‌گیری به‌صورت سلسله‌مراتبی و بر پایه‌ی مجموعه‌ای از آزمون‌ها بر روی ویژگی‌های داده‌ها مدل‌سازی می‌شود. ساختار درخت تصمیم شامل یک گره ریشه (نقطه شروع تصمیم‌گیری)، گره‌های داخلی (نمایانگر آزمون‌های منطقی بر روی ویژگی‌ها) و گره‌های برگ (نشان‌دهنده‌ی برچسب یا مقدار پیش‌بینی‌شده) است. این مدل با تقسیم تکراری داده‌ها به زیرمجموعه‌های کوچک‌تر تا زمانی که معیار توقف برآورده شود، ساخته می‌شود. از مزایای درخت تصمیم می‌توان به‌سادگی تفسیر، قابلیت نمایش گرافیکی و توانایی کار با داده‌های عددی و کیفی اشاره کرد. در مقابل، حساسیت به نویز و احتمال بیش‌برازش از مهم‌ترین محدودیت‌های آن محسوب می‌شوند.

فرآیند ایجاد درخت تصمیم را می‌توان به‌صورت گام‌های زیر در قالب شبه کد خلاصه کرد:

Input: Training data X_train, y_train; Test data X_test

Output: Predicted labels y_pred

1. Initialize root node with all training samples

2. For each node:

   a. For each feature, compute best split threshold with highest information gain

   b. Choose feature & threshold that maximizes impurity reduction

   c. Split node into left/right children

   d. Repeat the process recursively for each child node until the maximum depth is reached or no further split improves purity.

3. Assign a class label to each leaf node using the majority class of its samples.

4. Predict labels y_pred for X_test by traversing the tree from root to leaf based on feature values,

که

X_train: داده‌های آموزشی (متغیرهای ورودی)

y_train: برچسب‌های آموزشی (مقادیر هدف)

X_test: داده‌های آزمون

y_pred: برچسب‌ها یا مقادیر پیش‌بینی‌شده توسط مدل برای داده‌های آزمون

۲-۲- جنگل تصادفی (RandomForest)

جنگل تصادفی یکی از روش‌های قدرتمند در یادگیری ماشین است که بر پایه‌ی تجمیع چندین درخت تصمیم عمل می‌کند. در این روش، تعداد زیادی درخت تصمیم با استفاده از نمونه‌گیری تصادفی از داده‌ها و انتخاب تصادفی زیرمجموعه‌ای از ویژگی‌ها در هر گره ساخته می‌شود. نتیجه نهایی با تجمیع پاسخ درخت‌ها به‌دست می‌آید. این رویکرد با کاهش واریانس مدل و جلوگیری از بیش‌برازش، عملکردی پایدار و دقیق‌تر نسبت به یک درخت تصمیم منفرد ارایه می‌دهد. همچنین جنگل تصادفی نسبت به نویز داده‌ها مقاوم‌تر است و قابلیت تخمین اهمیت ویژگی‌ها را دارد. شبه‌کد زیر مراحل اصلی ساخت و تجمیع درخت‌ها در مدل جنگل تصادفی را نشان می‌دهد.

Input: Training data X_train, y_train; Test data X_test

Output: Predicted labels y_pred

For t=1 to T (number of trees):
a. Draw a bootstrap sample from X_train .
b. Build a Decision Tree using a random subset of features at each split.
c. Store the trained tree ht(x)

For prediction:

I. For each test sample xxx, collect all tree predictions ht(x).
II. Aggregate predictions by majority vote:
III. Ypred=mode{h1(x), h2(x),..., hT(x)}

که

X_train: داده‌های آموزشی (متغیرهای ورودی)

y_train: برچسب‌های آموزشی (مقادیر هدف)

X_test: داده‌های آزمون

y_pred: برچسب‌ها یا مقادیر پیش‌بینی‌شده توسط مدل برای داده‌های آزمون

ht(x): درخت تصمیم شماره‌ی t در یک جنگل تصادفی

۳-۲- ماشین بردار پشتیبان[1]

ماشین بردار پشتیبان یکی از الگوریتم‌های نظارت‌شده در یادگیری ماشین است که برای طبقه‌بندی و رگرسیون به‌کار می‌رود. هدف اصلی این الگوریتم یافتن یک ابرصفحه است که داده‌ها را با بیشترین حاشیه از هم جدا کند. بردارهای نزدیک به مرز تصمیم که نقش کلیدی در تعیین موقعیت ابرصفحه دارند، بردارهای پشتیبان نامیده می‌شوند. در مسایل غیرقابل تفکیک خطی، با استفاده از تابع کرنل، داده‌ها به فضای ویژگی با بعد بالاتر نگاشت می‌شوند تا بتوان جداسازی خطی را در آن فضا انجام داد. از مهم‌ترین کرنل‌ها می‌توان به کرنل‌های خطی، چندجمله‌ای و تابع پایه شعاعی اشاره کرد. از مزایای *SVM* می‌توان به کارایی بالا در فضاهای بعد بالا، پایداری در برابر بیش‌برازش و تعمیم‌پذیری مناسب اشاره کرد. با این حال، حساسیت به انتخاب پارامترها و نوع کرنل از چالش‌های اصلی آن است. شبه‌کد استاندارد الگوریتم *SVM* به‌صورت زیر خلاصه می‌شود:

Input: Training data X_train, y_train; Test data X_test

Output: Predicted labels y_pred

1. Define linear kernel function
2. Optimize the following objective:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

s. t.

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0.$$

3. Compute support vectors and parameters w,b.
4. Predict each test sample x∈Xtest:

$$y_{\text{pred}} = \text{sign}(w^T\phi(x) + b).$$

[1] Support Vector Machine (SVM)

که

X_train: داده‌های آموزشی (متغیرهای ورودی)

y_train: برچسب‌های آموزشی (مقادیر هدف)

X_test: داده‌های آزمون

y_pred: برچسب‌ها یا مقادیر پیش‌بینی‌شده توسط مدل برای داده‌های آزمون

W: بردار وزن‌ها

b: جمله‌ی بایاس (انحراف از مبدا) که به تابع تصمیم مدل افزوده می‌شود.

ξi: متغیر مجاز در *SVM* که اجازه می‌دهد برخی نمونه‌ها محدودیت حاشیه را نقض کنند.

C: پارامتر منظم‌سازی در *SVM* که موازنه‌ای بین بیشینه‌سازی حاشیه و کمینه‌سازی خطا ایجاد می‌کند.

$\phi(x)$: تبدیل ویژگی‌ها (نگاشت کرنل) در *SVM* که داده‌ها را به فضای با ابعاد بالاتر نگاشت می‌کند.

xi: نمونه i_{th} از داده‌ها

yi: برچسب واقعی متناظر با نمونه *xi*

۴-۲- شبکه عصبی چندلایه[1]

شبکه عصبی چندلایه یکی از مدل‌های پرکاربرد یادگیری عمیق است که از چندین لایه شامل یک لایه ورودی، یک یا چند لایه پنهان و یک لایه خروجی تشکیل می‌شود. هر لایه شامل تعدادی نورون است که با ضرایب وزنی به نورون‌های لایه قبل و بعد متصل می‌شوند. در این مطالعه، داده‌های ورودی پس از نرمال‌سازی به شبکه وارد شدند. برای هر نورون، ترکیب خطی ورودی‌ها از طریق تابع فعال‌سازی غیرخطی تبدیل شد تا توانایی مدل در تقریب روابط غیرخطی افزایش یابد. آموزش شبکه با استفاده از الگوریتم پس‌انتشار خطا و به‌کارگیری بهینه‌ساز انجام گرفت. معیار توقف آموزش بر اساس حداقل شدن تابع هزینه (مانند میانگین مربع خطا یا آنتروپی متقاطع) و بهبود عملکرد روی داده‌های اعتبارسنجی تعیین شد. شبه‌کد زیر مراحل اصلی آموزش و ارزیابی شبکه عصبی چندلایه را نشان می‌دهد.

Input: Training data X_train, y_train; Test data X_test

Output: Predicted labels y_pred

1. Initialize network architecture with input, hidden, and output layers.
2. Randomly initialize weights and biases.
3. For each epoch:

[1] Multi Layer Perceptron (MLP)

a. Perform forward propagation to compute output:

$$a^{(l)} = f\left(W^{(l)}a^{(l-1)} + b^{(l)}\right).$$

b. Compute loss between predicted and true labels.
c. Perform backpropagation to compute gradients of weights.
d. Update weights and biases using gradient descent.

4. Predict y_pred for X_test using forward propagation,

که

X_train: داده‌های آموزشی (متغیرهای ورودی)

y_train: برچسب‌های آموزشی (مقادیر هدف)

X_test: داده‌های آزمون

y_pred: برچسب‌ها یا مقادیر پیش‌بینی‌شده توسط مدل برای داده‌های آزمون

al: بردار خروجی (فعال‌سازی) لایه ۱ در شبکه عصبی چندلایه

Wl ,bl: ماتریس وزن‌ها و بردار بایاس لایه ۱ در *MLP*

f(.): تابع فعال‌سازی در *MLP* مانند *ReLU*، سیگموید یا تانژانت هایپربولیک

۵-۲- شبکه بیزین (Bayesian Network)

شبکه بیزی یک مدل گرافیکی احتمالاتی است که روابط شرطی میان متغیرها را با یک گراف جهت‌دار بدون حلقه نمایش می‌دهد. هر گره در شبکه نشان‌دهنده یک متغیر تصادفی و یال‌ها نشان‌دهنده وابستگی‌های شرطی بین آن‌ها هستند. پارامترهای شبکه شامل توزیع‌های احتمال شرطی برای هر گره نسبت به والدین آن تعیین می‌شوند. آموزش شبکه بیزی شامل برآورد ساختار گراف و پارامترها از داده‌های مشاهده‌شده است و برای استنتاج از الگوریتم‌هایی مانند *Variable Elimination* یا *Belief Propagation* استفاده می‌شود.

شبه‌کد مراحل اجرای شبکه بیزی در الگوریتم زیر ارایه شده است.

Input: Training data X_train, y_train; Test data X_test

Output: Predicted labels y_pred

1. Define network structure G=(V, E), where each node represents a variable.
2. Estimate conditional probability tables (CPTs) P(Xi|Parents (Xi)) from X_train.
3. For each test sample:
   a. Compute posterior probability for each class c:

$$P(c \mid X) \propto P(c)\prod_i P\left(X_i \mid Parents(X_i)\right).$$

   b. Assign class with the highest posterior probability.

4. Output y_pred for all samples in X_test

که

X_train: داده‌های آموزشی (متغیرهای ورودی)

y_train: برچسب‌های آموزشی (مقادیر هدف)

X_test: داده‌های آزمون

y_pred: برچسب‌ها یا مقادیر پیش‌بینی‌شده توسط مدل برای داده‌های آزمون

G=(V,E): گراف بدون دور جهت‌دار که ساختار شبکه بیزی را نمایش می‌دهد؛ V مجموعه‌ی گره‌ها (متغیرها) و E مجموعه یال‌ها (وابستگی‌ها) است.

Parents(Xi): مجموعه‌ی والدهای گره Xi در شبکه بیزی (متغیرهایی که بر Xi تاثیر دارند)

P(Xi/Parents(Xi): توزیع احتمال شرطی متغیر Xi با توجه به والدهایش در شبکه بیزی

P(c/X): احتمال پسین کلاس c با توجه به داده‌ی ورودی X، طبق قضیه‌ی بیز

۵-۲- تقویت گرادیان (XGBoost)

الگوریتم *XGBoost* یکی از روش‌های پیشرفته در خانواده مدل‌های تقویت گرادیان است که با هدف بهبود دقت، کارایی و سرعت یادگیری طراحی شده است. این الگوریتم مجموعه‌ای از درخت‌های تصمیم را به‌صورت ترتیبی آموزش می‌دهد، به‌گونه‌ای که هر درخت جدید خطاهای پیش‌بینی درخت‌های قبلی را اصلاح می‌کند. در نتیجه، مدل نهایی ترکیبی از چندین درخت ضعیف است که با هم یک پیش‌بین قوی را تشکیل می‌دهند.

XGBoost از بهینه‌سازی گرادیانی برای کمینه‌سازی تابع خطا استفاده می‌کند و در مقایسه با نسخه‌های سنتی گرادیان بوستینگ، شامل بهبودهایی در زمینه‌ی منظم‌سازی، کنترل پیچیدگی مدل و مدیریت داده‌های ناقص است. این ویژگی‌ها موجب کاهش بیش‌برازش و افزایش قابلیت تعمیم مدل می‌شوند.

علاوه بر این، *XGBoost* از محاسبات موازی و بهینه‌سازی حافظه بهره می‌برد که باعث افزایش قابل توجه سرعت آموزش در مجموعه داده‌های بزرگ می‌شود. به دلیل دقت بالا، پایداری و کارایی مناسب، *XGBoost* در بسیاری از مسایل یادگیری ماشین از جمله پیش‌بینی، طبقه‌بندی و رگرسیون به‌عنوان یکی از بهترین الگوریتم‌ها شناخته می‌شود. شبه‌کد الگوریتم *XGBoost* برای تبیین مراحل اجرای روش ارایه شده است.

Input: Training data X_train, y_train; Test data X_test

Output: Predicted labels y_pred

1. Initialize model with a constant prediction (e.g., mean of ytrainy_{train}ytrain).
2. For t=1 to T (number of boosting rounds):
   a. Compute pseudo-residuals:

$$r_i = -\frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i}.$$

b. Fit a regression tree ft(x) to predict residuals ri
c. Compute leaf weights to minimize loss with regularization.
d. Update model:

$$\hat{y}_i \leftarrow \hat{y}_i + \eta f_t(x_i).$$

3. Predict y_pred for X_test using the sum of all trees:

$$y_{\text{pred}} = \sum_{t=1}^{T} \eta f_t(x),$$

که

X_train: داده‌های آموزشی (متغیرهای ورودی)

y_train: برچسب‌های آموزشی (مقادیر هدف)

X_test: داده‌های آزمون

y_pred: برچسب‌ها یا مقادیر پیش‌بینی‌شده توسط مدل برای داده‌های آزمون

r_i: گرادیان منفی تابع خطا نسبت به پیش‌بینی

y_i: برچسب واقعی مربوط به نمونه i

$\hat{y}_i$: مقدار پیش‌بینی‌شده برای نمونه i در زمان آموزش مدل

η: نرخ یادگیری که میزان تاثیر هر درخت جدید در *XGBoost* را کنترل می‌کند.

$f_t(x)$: تابع پیش‌بینی (درخت) در مرحله‌ی t از فرآیند *Boosting*

۷-۲- مجموعه داده

بدین منظور داده‌های مربوط به مشتریان یکی از بانک‌های دولتی که در سال‌های ۱۴۰۰ تا ۱۴۰۳ فعال بودند بررسی و جمع‌آوری شد. در این پژوهش، ۱۲ ویژگی کلیدی مرتبط با رفتار مشتریان برای پیش‌بینی رویگردانی انتخاب شد. این تعداد بر اساس مرور ادبیات پیشین، محدودیت داده‌های واقعی بانک و تحلیل اولیه اهمیت متغیرها بر اساس نظر خبرگان تعیین گردید. انتخاب تعداد بیشتری از ویژگی‌ها می‌توانست باعث افزایش پیچیدگی مدل و کاهش قابلیت تفسیر شود و کاهش تعداد ویژگی‌ها نیز ممکن بود اطلاعات رفتاری مهم مشتریان از دست برود؛ بنابراین، تعداد ۱۲ داده، تعادل مناسبی بین دقت پیش‌بینی و قابلیت عملی پیاده‌سازی مدل‌ها ایجاد می‌کند.

برای هر مورد ۱۲ داده شامل سن مشتری، مدت زمان رابطه با بانک، تعداد سپرده‌های مشتری، تعداد ماه‌های غیرفعال (فاقد تراکنش) در ۱۲ ماه گذشته، میانگین تعداد تراکنش روزانه در ۶ ماه گذشته، حداقل میانگین ماهانه سپرده‌ها در ۱۲ ماه گذشته (میلیون ریال)، میانگین سپرده‌ها در ۱۲ ماه گذشته، تغییر مقدار تراکنش‌های سه ماهه چهارم به سه ماهه اول، میانگین تراکنش‌ها در ۱۲ ماه گذشته (میلیون ریال)، میانگین ماهانه تعداد تراکنش‌ها در ۱۲ ماه گذشته، تغییر تعداد تراکنش‌های سه ماهه چهارم به سه ماهه اول و میانگین سپرده‌ها در ۳ ماه گذشته (میلیون ریال) جمع‌آوری شد.

جدول ۲- ویژگی‌های مشتریان.
Table 1- Customer features.

ردیف	ویژگی منتخب	توضیح مختصر	منابع
1	سن مشتری	شاخص جمعیت‌شناختی مرتبط با وفاداری و ریزش	[21]
2	مدت زمان رابطه با بانک	طول عمر مشتری؛ مشتریان قدیمی احتمال ریزش کمتر	[22]، [23]
3	تعداد سپرده‌های مشتری	عمق رابطه مشتری با بانک	[24]
4	تعداد ماه‌های غیرفعال در ۱۲ ماه گذشته	شاخص کاهش تعامل مشتری	[25]
5	میانگین تعداد تراکنش روزانه در ۶ ماه گذشته	میزان فعالیت مالی و تعامل مشتری	[26]
6	حداقل میانگین ماهانه سپرده‌ها در ۱۲ ماه گذشته (میلیون ریال)	شاخص توان مالی و پایداری حساب	[27]
7	میانگین سپرده‌ها در ۱۲ ماه گذشته	سطح تعامل و اعتماد مشتری	[28]
8	تغییر مقدار تراکنش‌های سه ماهه چهارم به سه ماهه اول	روند افزایش یا کاهش فعالیت مالی	[29]
9	میانگین تراکنش‌ها در ۱۲ ماه گذشته (میلیون ریال)	سطح تعامل مالی و عمق رابطه	[30]
10	میانگین ماهانه تعداد تراکنش‌ها در ۱۲ ماه گذشته	شاخص فعالیت مداوم مشتری	[31]
11	تغییر تعداد تراکنش‌های سه ماهه چهارم به سه ماهه اول	روند رفتاری و تغییر در تعامل	[32]
12	میانگین سپرده‌ها در ۳ ماه گذشته میلیون ریال	شاخص فعالیت مالی اخیر مشتری	[33]

همان‌گونه که در **جدول** مشاهده می‌شود، ویژگی‌های انتخابی با یافته‌های پژوهش‌های پیشین هم‌راستا هستند و درعین‌حال با محدودیت داده‌های واقعی بانک و تحلیل اولیه اهمیت متغیرها تطبیق یافته‌اند.

۲-۸- پیش‌پردازش داده‌ها

در این پژوهش، تمام ۱۲ متغیر مورد استفاده کمی بودند. برای مشتریانی که یکی از مقادیر داده مفقود داشتند، این مقدار با میانگین همان متغیر در بین سایر مشتریان جایگزین شد. استفاده از میانگین به‌عنوان برآوردگر ناریب امکان می‌دهد که رفتار مشتری بدون ایجاد سوگیری سیستماتیک مدل‌سازی شود، درحالی‌که حذف کامل مشتریان با داده ناقص ممکن است منجر به کاهش حجم نمونه و از دست رفتن اطلاعات ارزشمند گردد. با توجه به این نکته و نسبت بسیار کم داده‌های مفقود (کمتر از ۱%)، این روش منطقی و متداول در مطالعات پیش‌بینی رفتار مشتری محسوب می‌شود. سپس همه داده‌های کمی با روش *z-score* نرمالایز شدند.

۲-۹- پیاده‌سازی مدل‌های یادگیری ماشین

الگوریتم‌های مورد استفاده در این پژوهش با نرم‌افزار *R* پیاده‌سازی شدند. برای همه الگوریتم‌ها از کتابخانه‌های استاندارد *R* استفاده شد تا صحت عملکرد آن‌ها تضمین شود. استفاده از کتابخانه‌های معتبر، صحت پیاده‌سازی الگوریتم‌ها را تضمین می‌کند. علاوه بر این، نتایج بخش‌بندی و پیش‌بینی با مقایسه میانگین و پراکندگی، به‌صورت منطقی و نزدیک به هم بودند که نشان‌دهنده صحت اجرای الگوریتم‌ها است.

۲-۱۰- آنالیز نتایج

ابتدا آمار توصیفی داده‌ها به همراه بررسی توزیع متغیرها در دو گروه مشتریان وفادار و مشتریان رویگردان گزارش شد. سپس آزمون مقایسه میانگین دوجامعه (گروه مشتریان وفادار و رویگردان) در هر ۱۲ متغیر با سطح اطمینان ۹۵% انجام شد. تمامی تحلیل‌های آماری با استفاده از نرم‌افزار *SPSS 20* انجام شد. سپس مدل‌های *ML* مختلف شامل جنگل تصادفی، درخت تصمیم، *SVM*، *MLP*، شبکه بیزین و *XGBoost* اجرا شد تا بهترین مدل برای پیش‌بینی رفتار مشتریان تعیین شود. مدل‌های استفاده‌شده بر اساس معیارهای صحت، حساسیت و ویژگی و با استفاده از روش *10-fold cross validation* ارزیابی شدند. کارایی میزان طبقه‌بندی مدل‌های مختلف بر اساس ماتریس آشفتگی که در **جدول ۲** بیان گردیده موردبررسی قرار گرفت.

جدول ۳- ماتریس آشفتگی.
Table 3- Turbulence matrix.

		Predicted	
		Positive	Negative
Real	Positive	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

در این پژوهش معیارهای *Positive* (*TP*) و *True Negative* (*TN*) به ترتیب تعداد مشتریان رویگردان و وفادار است که به‌درستی توسط الگوریتم‌ها طبقه‌بندی شده بودند، *False Positive* (*FP*) تعداد مشتریان وفادار می‌باشد که به اشتباه توسط مدل رویگردان تشخیص داده شده بودند و *False Negative* (*FN*) تعداد مشتریان رویگردان بوده است که توسط الگوریتم‌ها به‌عنوان وفادار در نظر گرفته شده بودند. در **جدول ۳** نحوه محاسبه معیارهای عملکردی مورد استفاده در پژوهش نمایش داده شده است. پیاده‌سازی و ارزیابی تکنیک‌های یادگیری ماشین در محیط *R* *4.4.2* انجام شد.

جدول ۴- محاسبه معیارهای عملکردی مورد استفاده.
Table 4- Calculation of performance metrics used.

Performance criteria	Calculation
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$
Sensitivity/ Recall	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{FP + TN}$

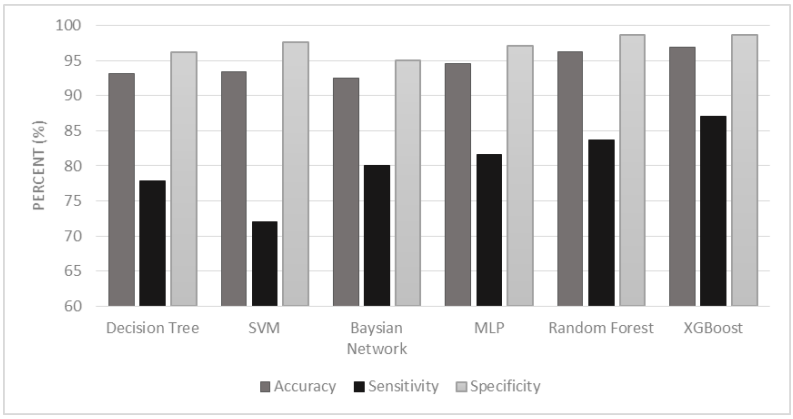
۳- یافته‌ها

در این مطالعه داده‌های مربوط به مشتریان یکی از بانک‌های دولتی خراسان شمالی در سال ۱۴۰۰ تا ۱۴۰۳ جمع‌آوری شد. از ۲۰۲۵ مشتری تعداد ۳۲۵ (۱۶%) نفر از خدمات بانک رویگردانی کرده‌اند. **جدول ۴** آمار توصیفی متغیرهای مطالعه را نشان می‌دهد.

جدول ۵- توصیف متغیرهای مطالعه
Table 5- Description of study variables.

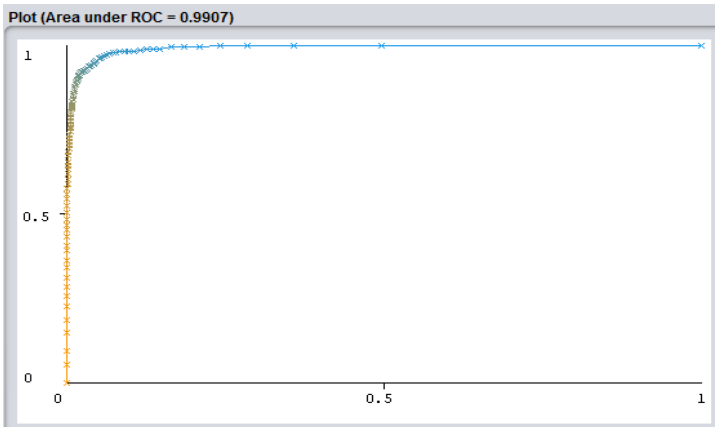
نام متغیر	مشتریان وفادارمیانگین (انحراف معیار)	مشتریان رویگردان میانگین (انحراف معیار)	P-Value
سن مشتری	45.12 ±8.08	45.74 ±7.67	0.058
مدت‌زمان رابطه با بانک (بر اساس تاریخ افتتاح سپرده به ماه)	36.65 ± 8.02	36.27 ± 7.8	0.168
تعداد سپرده‌های مشتری	3.87 ± 1.12	3.39 ± 1.14	<0.001
تعداد ماه‌های غیرفعال (فاقد تراکنش) در ۱۲ ماه گذشته	2.36 ± 1.04	2.75 ± 1.02	<0.001
میانگین تعداد تراکنش روزانه در ۶ ماه گذشته	2.28 ± 1.07	2.79 ± 1.08	<0.001
حداقل میانگین ماهانه سپرده‌ها در ۱۲ ماه گذشته (میلیون ریال)	1256.6 ± 757.74	672.82 ± 921.39	<0.001
میانگین سپرده‌ها در ۱۲ ماه گذشته	747.273 ± 908.67	746.216 ± 910.21	0.977
تغییر مقدار تراکنش‌های سه ماهه چهارم به سه ماهه اول	0.77 ± 0.22	0.69 ± 0.21	<0.001
میانگین تراکنش‌ها در ۱۲ ماه گذشته (میلیون ریال)	665.66 ± 351.78	509.03 ± 230.23	<0.001
میانگین ماهانه تعداد تراکنش‌ها در ۱۲ ماه گذشته	68.67 ± 22.92	44.93 ± 14.57	<0.001
تغییر تعداد تراکنش‌های سه ماهه چهارم به سه ماهه اول	0.74 ± 0.23	0.55 ± 0.23	<0.001
میانگین سپرده‌ها در ۳ ماه گذشته (میلیون ریال)	234 ± 0.27	123 ± 0.26	<0.001

با توجه به اینکه مقادیر *p-value* برای متغیرهای سن مشتری، مدت زمان رابطه با بانک و میانگین سپرده‌ها در ۱۲ ماه گذشته بیشتر از ۰/۰۵ هستند، می‌توان نتیجه گرفت که تفاوت آماری معناداری بین گروه‌ها برای این متغیرها وجود ندارد. لذا این متغیرها از مطالعه خارج شدند. در مورد سایر متغیرها به دلیل وجود *p-value* کمتر از ۰/۰۵، این نشان‌دهنده تفاوت معنادار آماری است و به‌عنوان ورودی‌های مدل‌های یادگیری ماشین در نظر گرفته شد. **شکل ۱** عملکرد مدل‌های یادگیری ماشین را در دسته‌بندی مشتریان به دو گروه وفادار و رویگردان نشان می‌دهد. از نظر هر سه معیار عملکردی، *XGBoost* با صحت ۹۶/۸۹، حساسیت ۸۷/۱۱ و ویژگی ۹۸/۷۱% بهترین الگوریتم است. **شکل ۲** نمودار *ROC* مربوط به الگوریتم *XGBoost* را نشان می‌دهد. مقدار سطح زیر منحنی[1] برابر با ۰/۹۹۰۷ و بسیار نزدیک به ۱ است. *AUC* معیار توانایی مدل در تمایز بین مشتریان ریزش‌کننده و غیرریزش‌کننده است؛ هرچه این مقدار به ۱ نزدیک‌تر باشد، قدرت تشخیص مدل بالاتر و عملکرد آن بهتر است. به‌طور معکوس، *AUC* نزدیک به ۰/۵ نشان‌دهنده عملکرد تقریبا شانسی مدل خواهد بود؛ بنابراین مقدار ۰/۹۹۰۷ نشان‌دهنده کیفیت بسیار بالای پیش‌بینی مدل است.



شکل ۱- مقایسه عملکرد الگوریتم‌های یادگیری ماشین.

Figure 1- Comparing the performance of machine learning algorithms.



شکل ۲- نمودار ROC برای الگوریتم Xgboost.

Figure 2- ROC chart for the Xgboos algorithm.

۴- بحث و نتیجه‌گیری

نتایج این پژوهش نشان داد که استفاده از الگوریتم‌های یادگیری ماشین می‌تواند رویکردی کارآمد و اثربخش برای شناسایی و پیش‌بینی ریزش مشتریان در صنعت بانکداری باشد. تحلیل داده‌ها حاکی از آن است که برخی مدل‌های یادگیری ماشین، به‌ویژه الگوریتم‌هایی مانند *XGBoost*،

[1] Area Under the Curve (AUC)

عملکرد بهتری نسبت به سایر روش‌ها از جمله رگرسیون لجستیک یا درخت تصمیم ساده دارند. این یافته‌ها بیانگر آن است که مدل‌هایی که قابلیت پردازش روابط غیرخطی و تعامل میان متغیرها را دارند، توانایی بیشتری در پیش‌بینی رفتار مشتریان از خود نشان می‌دهند. مطالعه حاضر همسو با پژوهش‌های پیشین در حوزه تحلیل داده‌های بانکی است که اهمیت یادگیری ماشین در ارتقای کیفیت تصمیم‌گیری و مدیریت ارتباط با مشتریان را برجسته ساخته‌اند. با این حال، وجه تمایز تحقیق حاضر آن است که با تمرکز بر داده‌های واقعی مشتریان بانک و مقایسه چندین الگوریتم مختلف، تصویری جامع از نقاط قوت و ضعف هر یک از روش‌ها در شرایط واقعی ارایه کرده است. یکی از یافته‌های مهم پژوهش این بود که متغیرهایی مانند مدت زمان همکاری مشتری با بانک، تعداد تراکنش‌ها، تعداد ماه‌های غیرفعال، تعداد سپرده‌ها و میانگین موجودی حساب بیشترین تاثیر را در احتمال ریزش مشتریان دارند. این موضوع نشان می‌دهد که علاوه بر شاخص‌های مالی سنتی، شاخص‌های رفتاری و الگوهای تعاملی مشتریان با بانک نیز نقش بسیار مهمی در پیش‌بینی ریزش ایفا می‌کنند. تبیین نتایج نشان می‌دهد که مشتریانی که تعامل کمتری با بانک دارند یا از خدمات متنوع بانک کمتر استفاده می‌کنند، بیشتر در معرض ریزش قرار می‌گیرند؛ بنابراین، بانک‌ها می‌توانند با شناسایی این گروه از مشتریان، اقدامات پیشگیرانه‌ای مانند ارایه پیشنهادهای شخصی‌سازی‌شده، طراحی بسته‌های تشویقی یا بهبود تجربه کاربری خدمات دیجیتال را در دستور کار قرار دهند.

به‌طورکلی، نتایج این پژوهش تایید می‌کند که الگوریتم‌های یادگیری ماشین می‌توانند ابزارهای قدرتمندی در مدیریت ارتباط با مشتریان بانکی باشند. این مدل‌ها با دقت بالای خود قادرند مشتریان در معرض ریزش را شناسایی کنند و بدین ترتیب به مدیران کمک کنند تا اقدامات اصلاحی لازم را قبل از خروج مشتریان انجام دهند. اجرای چنین مدل‌هایی می‌تواند منجر به کاهش نرخ ریزش، افزایش رضایت مشتریان، بهبود تصویر برند و درنهایت ارتقای سودآوری بانک شود. مدیران می‌توانند مشتریانی که احتمال ریزش بالایی دارند را شناسایی کرده و برای آن‌ها برنامه‌های حفظ مشتری اختصاصی طراحی کنند. بر اساس یافته‌ها، متغیرهایی مثل میانگین سپرده‌ها، تعداد تراکنش‌ها و مدت زمان رابطه با بانک نقش مهمی در پیش‌بینی ریزش دارند. مدیران می‌توانند نظارت و تحلیل این متغیرها را در سیستم‌های *CRM* و داشبوردهای مدیریتی تقویت کنند. برای مشتریان در معرض ریزش، اقدامات وفادارسازی و ارایه مشوق‌های مالی یا خدمات ویژه می‌تواند اثرگذار باشد. با استفاده از خروجی مدل، بانک می‌تواند کمپین‌های بازاریابی را بر اساس ریسک ریزش مشتریان هدفمند کند و منابع را به شکل بهینه تخصیص دهد. با توجه به تغییر شرایط بازار و رفتار مشتریان، مدیران باید مدل‌های پیش‌بینی را به‌صورت دوره‌ای بازبینی و به‌روزرسانی کنند تا دقت پیش‌بینی حفظ شود.

کاربردهای عملی نتایج این پژوهش را می‌توان در چند محور اصلی خلاصه کرد:

۱. حفظ مشتریان ارزشمند: شناسایی مشتریان کلیدی که احتمال ریزش بالایی دارند و طراحی برنامه‌های وفاداری ویژه برای آنان
۲. بهبود استراتژی‌های بازاریابی: استفاده از نتایج پیش‌بینی برای طراحی کمپین‌های هدفمند و ارایه خدمات شخصی‌سازی‌شده
۳. مدیریت ریسک و کاهش هزینه‌ها: تمرکز بر روی مشتریانی که احتمال ترک آن‌ها بیشتر است و در نتیجه بهینه‌سازی تخصیص منابع
۴. افزایش بهره‌وری خدمات دیجیتال: استفاده از مدل‌های یادگیری ماشین برای توسعه سیستم‌های هوشمند پیشنهاددهنده و ارتقای تجربه مشتریان در بسترهای آنلاین

در مجموع، پژوهش حاضر با تاکید بر اهمیت یادگیری ماشین در تحلیل داده‌های بانکی و ارایه شواهد تجربی معتبر، گامی مهم در جهت توسعه سیستم‌های هوشمند مدیریت مشتری در صنعت بانکداری محسوب می‌شود. استمرار چنین پژوهش‌هایی می‌تواند زمینه‌ساز ایجاد بانک‌های داده‌محور و هوشمند در آینده باشد؛ بانک‌هایی که با تکیه بر تحلیل پیشرفته داده‌ها، قادر خواهند بود نه‌تنها ریزش مشتریان را کاهش دهند بلکه ارزش طول عمر مشتری را نیز به‌طور چشمگیری افزایش دهند.

۱-۴- محدودیت‌ها

این پژوهش با محدودیت‌هایی نیز همراه بوده است؛ ۱- آنکه داده‌های استفاده‌شده مربوط به یک بانک دولتی بوده و ممکن است الگوهای رفتاری مشتریان در سایر بانک‌ها متفاوت باشد، ۲- کیفیت و کامل بودن داده‌ها نیز می‌تواند بر عملکرد مدل‌ها تاثیرگذار باشد؛ به‌ویژه زمانی که داده‌های ناقص یا نویزی وجود داشته باشد و ۳- برخی عوامل روان‌شناختی و اجتماعی موثر بر ریزش مشتریان در این مطالعه لحاظ نشده‌اند، درحالی‌که می‌توانند در تصمیم مشتریان برای ادامه یا قطع همکاری با بانک نقش‌آفرین باشند. از این رو، پیشنهاد می‌شود پژوهش‌های آتی از داده‌های چند

بانکی، مدل‌های پیشرفته‌تر مانند یادگیری عمیق و همچنین متغیرهای رفتاری و روان‌شناختی مشتریان استفاده کنند تا نتایج جامع‌تر و دقیق‌تری حاصل شود.

تشکر و قدردانی

نویسندگان از همراهی و حمایت تمامی افرادی که به‌طور مستقیم یا غیرمستقیم در تکمیل این پژوهش نقش داشتند، قدردانی می‌کنند.

منابع مالی

این پژوهش هیچ‌گونه حمایت مالی یا کمک‌هزینه تحقیقاتی از نهادها یا سازمان‌های تامین‌کننده مالی دریافت نکرده است.

تعارض با منافع

نویسندگان اعلام می‌کنند که هیچ‌گونه تعارض منافع در این پژوهش وجود ندارد.

منابع

[1] Jafari, M. J., Tarokh, M. J., & Soleimani, P. (2024). A data-driven Agent-based model and framework for Churn prediction in Telecommunication Industry. *Modern research in decision making*, *9*(2), 164–190. (**In Persian**). https://journal.saim.ir/article_714344.html?lang=en

[2] Khajvand, M., & Tarokh, M. J. (2011). Analyzing customer segmentation based on customer value components (Case study: a private bank) (Technical note). *Advances in industrial engineering*, *45*(Special Issue), 79–93. https://aie.ut.ac.ir/article_23328.html

[3] Ha, S., & Bae, S. (2006). *Keeping track of customer life cycle to build customer relationship* (Vol. 4093).

[4] Keaveney, S. M. (1995). Customer switching behavior in service industries: An exploratory study. *Journal of marketing*, *59*, 71–82. https://doi.org/10.1177/002224299505900206

[5] Clemes, M., Gan, C., & Zhang, D. (2010). Customer switching behaviour in the Chinese retail banking industry. *International journal of bank marketing*, *28*, 519–546. https://doi.org/10.1108/02652321011085185?urlappend=%3Futm_source%3Dresearchgate

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

[7] Asgari, M., Taghva, M., & Taghavifard, M. (2018). Churn prediction in Iran banking industry case of a private Iranian bank. *Public management researches*, *11*(41), 57–82. https://doi.org/10.22111/jmr.2018.4378

[8] Bilal Zorić, A. (2016). Predicting customer churn in banking industry using neural networks. *Interdisciplinary description of complex systems*, *14*, 116–124. https://doi.org/10.7906/indecs.14.2.1

[9] Zhang, T. (2022). Prediction and clustering of bank customer churn based on XGBoost and K-means. *BCP business & management*, *23*, 360–366. https://doi.org/10.54691/bcpbm.v23i.1373

[10] Lundberg, S., & Lee, S. I. (2017). *A unified approach to interpreting model predictions*. https://doi.org/10.48550/arXiv.1705.07874

[11] Hassani, M., Siccha, S., Richter, F., & Seidl, T. (2015). Efficient process discovery from event streams using sequential pattern mining. *2015 IEEE symposium series on computational intelligence (SSCI)*. IEEE. https://doi.org/10.1109/SSCI.2015.195

[12] Böttcher, M., Spott, M., Nauck, D., & Kruse, R. (2009). Mining changing customer segments in dynamic markets. *Expert systems with applications*, *36*(1), 155–164. https://doi.org/10.1016/j.eswa.2007.09.006

[13] Blocker, C. P., & Flint, D. J. (2007). Customer segments as moving targets: Integrating customer value dynamism into segment instability logic. *Industrial marketing management*, *36*(6), 810–822. https://doi.org/10.1016/j.indmarman.2006.05.016

[14] Lemmens, A., Croux, C., & Stremersch, S. (2012). Dynamics in the international market segmentation of new product growth. *International journal of research in marketing*, *29*(1), 81–92. https://doi.org/10.1016/j.ijresmar.2011.06.003

[15] Woo, J., Bae, S., & Park, S. C. (2005). Visualization method for customer targeting using customer map. *Expert syst. appl.*, *28*, 763–772. https://doi.org/10.1016/j.eswa.2004.12.041

[16] Ha, S. H. (2007). Applying knowledge engineering techniques to customer analysis in the service industry. *Advanced engineering informatics*, *21*(3), 293–301. https://doi.org/10.1016/j.aei.2006.12.001

[17] Tan, H., Xu, J., & Zhao, B. (2009). Research on index system of dynamic customer segmentation based on the case study of china telecom. *2009 international conference on information management and engineering* (pp. 441–445). IEEE. https://doi.org/10.1109/ICIME.2009.82

[18] Homburg, C., Steiner, V., & Totzek, D. (2009). Managing dynamics in a customer portfolio. *Journal of marketing- j marketing*, *73*, 70–89. https://doi.org/10.1509/jmkg.73.5.70?urlappend=%3Futm_source%3Dresearchgate

[19] Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert systems with applications*, *38*(12), 15273–15285. https://doi.org/10.1016/j.eswa.2011.06.028

[20] Xie, Y., Li, X., Ngai, E., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert systems with applications*, *7*, 5445–5449. https://doi.org/10.1016/j.eswa.2008.06.121

[21] Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2018). Customer churn prediction in telecommunication industry using data certainty. *Journal of business research*, *94*. https://doi.org/10.1016/j.jbusres.2018.03.003

[22] Ahmad, R., & Buttle, F. (2002). Customer retention management: A reflection of theory and practice. *Marketing intelligence & planning*, *20*, 149–161. https://doi.org/10.1108/02634500210428003?urlappend=%3Futm_source%3Dresearchgate

[23] Awanife, S. (2025). Customer churn prediction in digital banking: a comparative study of xai techniques for interpretable decision-making. *American journal of humanities and social sciences research*, *09*(07), 114–122. https://B2n.ir/nr8717

[24] Coussement, K., & Van den Poel, D. (2008). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert systems with applications*, *36*, 6127–6134. https://doi.org/10.1016/j.eswa.2008.07.021

[25] Ballings, M., & Van den Poel, D. (2012). Customer event history for churn prediction: How long is long enough? *Expert systems with applications*, *39*, 13517–13522. https://doi.org/10.1016/j.eswa.2012.07.006

[26] Li, Y., & Yan, K. (2025). Prediction of bank credit customers churn based on machine learning and interpretability analysis. *Data science in finance and economics*, *5*(1), 19–34. https://doi.org/10.3934/DSFE.2025002

[27] Singh, P. P., Anik, F. I., Senapati, R., Sinha, A., Sakib, N., & Hossain, E. (2024). Investigating customer churn in banking: a machine learning approach and visualization app for data science and management. *Data science and management*, *7*(1), 7–16. https://doi.org/10.1016/j.dsm.2023.09.002

[28] Kosiba, J. P., Boateng, H., Okoe, A., & Hinson, R. (2018). Trust and customer engagement in the retail banking sector. *Service industries journal*, *40*(13). https://doi.org/10.1080/02642069.2018.1520219?urlappend=%3Futm_source%3Dresearchgate

[29] Bhuria, R., Gupta, S., Kaur, U., Bharany, S., Rehman, A., Hussen, S., … & Jangir, P. (2025). Ensemble-based customer churn prediction in banking: a voting classifier approach for improved client retention using demographic and behavioral data. *Discover sustainability*, *6*(1). https://doi.org/10.1007/s43621-025-00807-8

[30] Jones, K., & Leonard, L. N. K. (2008). Trust in consumer-to-consumer electronic commerce. *Information & management*, *45*(2), 88–95. https://doi.org/10.1016/j.im.2007.12.002

[31] Larivière, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert systems with applications*, *29*(2), 472–484. https://doi.org/10.1016/j.eswa.2005.04.043

[32] Bucko, J., Pavlov, B., & Pitka, T. (2025). Evaluating the effectiveness of customer behavior analysis in online sales through financial composite metrics. *Journal of marketing analytics*, 1–7. https://doi.org/10.1057/s41270-025-00430-6

[33] Aliyev, M., Ahmadov, E., Gadirli, H., Mammadova, A., & Alasgarov, E. (2020). *Segmenting bank customers via RFM model and unsupervised machine learning*. ArXiv E-Prints. https://ui.adsabs.harvard.edu/link_gateway/2020arXiv200808662A/doi:10.48550/arXiv.2008.08662