University of Qom

Engineering Management and Soft Computing

JEMSC.QOM
Online ISSN: 2538-2675

A model based on random forest algorithm and Jaya optimization to predict bank customer churn

Sepideh Chehreh[1] **and Ali Sarabadani**[2]

1. Corresponding author, Ph.D. Student in information technology engineering specializing in multimedia systems, Faculty of Engineering and Technology, Qom University, Qom, Iran. Email: chehreh.sepideh1@gmail.com
2. Ph.D. Student in information technology engineering specializing in Electronic commerce, Faculty of Engineering and Technology, Qom University, Qom, Iran. Email: alisarabadani14@gmail.com

Article Info

Article type:
Research Article

Article history:
Received 2023 August 31
Received in revised form 2023 October21
Accepted 2024 December 19
Published online 2024 March 15

Keywords:
customer churn,
machine learning,
random forest algorithm,
site optimization.

ABSTRACT

Customer churn is a financial term that refers to the loss of a customer. Due to many banks, losing customers from one bank to another has become a severe concern for different banks. Therefore, in this article, which has been compiled for the customers of a bank, it is possible to identify customers who have a high probability of falling by considering the behavior and characteristics of the customers before the fall occurs and to keep them by providing suggestions. In marketing, everyone agrees that keeping a customer is much less expensive than attracting a new customer, so this article introduces the different phases of predicting customer churn with the help of machine learning. The proposed method combines random forest algorithms and Jaya optimization, and customer dropout is based on different characteristics. Customers like age, Gender, c graphs, and cases. It predicts more. The results of the proposed model in the article are 91.41%, 95.66%, and 93.35%, respectively, in Precision, Recall, and Accuracy criteria.

Cite this article: Chehreh, S. & Sarabadani, A. (2023). A model based on random forest algorithm and Jaya optimization to predict bank customer churn. *Enginiiring Management and Soft Computing, 9 (2)*. 132-148. **DOI**: https://doi.org/

DOI: https://doi.org/

Publisher: University of Qom

JEMSC.QOM
شاپا الکترونیکی: ۲۶۷۵-۲۵۳۸

مدیریت مهندسی و رایانش نرم

دانشگاه قم

ارائه مدلی مبتنی بر الگوریتم جنگل تصادفی و بهینه‌سازی جایا برای پیش‌بینی ریزش مشتریان بانکی

سپیده چهره[1] ✉ **و علی سرآبادانی**[2]

۱. نویسنده مسئول، نویسنده مسئول، دانشجوی دکترا رشته مهندسی فناوری اطلاعات گرایش سیستم‌های چند رسانه‌ای، دانشکده فنی و مهندسی، دانشگاه قم، قم، ایران.
رایانامه: chehreh.sepideh1@gmail.com

۲. دانشجوی دکترا رشته مهندسی فناوری اطلاعات گرایش تجارت الکترونیک، دانشکده فنی و مهندسی، دانشگاه قم، قم، ایران. رایانامه: alisarabadani14@gmail.com

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۲/۰۶/۰۹

تاریخ بازنگری: ۱۴۰۲/۰۷/۲۹

تاریخ پذیرش: ۱۴۰۲/۰۹/۲۸

تاریخ انتشار: ۱۴۰۲/۱۲/۲۵

کلیدواژه‌ها:
مدیریت عدم‌قطعیت،
تحلیل احساسات،
نظریه امکان،
وردنت فازی.

چکیده

ریزش مشتری، یک اصطلاح مالی است که به از دست دادن مشتری اشاره دارد. امروزه با توجه به تعداد زیاد بانک‌ها، ریزش مشتریان از یک بانک به بانک دیگر تبدیل به دغدغه جدی برای بانک-های مختلف شده‌است. بنابراین در این مقاله که برای مشتریان یک بانک گردآوری شده‌است، می-توان با توجه به رفتار و ویژگی‌های مشتریان قبل از وقوع ریزش، به شناسایی مشتریانی که احتمال ریزش بالایی دارند، پرداخت و با ارائه پیشنهادهایی آنها را حفظ نمود. در بازاریابی همه بر این امر توافق دارند که حفظ یک مشتری از جذب یک مشتری جدید بسیار کم‌هزینه‌تر است. از این رو این مقاله به معرفی فازهای مختلف رویکرد پیش‌بینی مشتری ریزشی با کمک یادگیری ماشین پرداخته‌است. روش پیشنهادی ترکیبی از الگوریتم‌های جنگل تصادفی و بهینه‌سازی جایا می‌باشد و ریزش مشتری را براساس ویژگی‌های مختلف مشتری مانند سن، جنسیت، جغرافیا و موارد دیگر پیش‌بینی می‌کند. نتایج حاصل از مدل پیشنهادی در مقاله به ترتیب در معیارهای Precision، Recall و Accuracy برابر مقادیر ۹۱.۴۱ درصد، ۹۵.۶۶ درصد و ۹۳.۳۵ درصد می‌باشد.

استناد: 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، دوره ۹ (۲)*. صص: ۱۴۸-۱۳۲. https://doi.org/

ناشر: دانشگاه قم

۱) مقدمه

در کسب‌وکار امروز، تبلیغ‌ها و آگهی‌های زیادی در همه جا ازجمله تلویزیون، اینترنت ظاهر می‌شوند و مشتری یک خدمت و یا محصول را در معرض بسیاری از پیشنهادهای اغواکننده قرار می‌دهد. بنابراین بسیاری از مشتریان، شرکت ارائه‌دهنده خدمات، محصول خود را تغییر می‌دهند تا از مزایای این پیشنهادها بهره ببرند(امین، ۲۰۱۷). نقشه سفر مشتری، مجموعه کاملی از تجربه‌هایی است که مشتری هنگام تعامل با یک سازمان و برند به‌دست می‌آورد. این نقشه کمک می‌کند تا تجربه مشتری با آن سازمان از یک خدمت مقطعی و گذرا به یک رابطه طولانی‌مدت تبدیل شود.

مرحله‌های مختلف سفر مشتری شامل فاز آگاهی و اطلاع، علاقه‌مندی، بررسی و ارزیابی، خرید، حفظ و نگهداشت و وفاداری است. مرحله حفظ و نگهداشت زمانی است که مشتری خدمات موردنیاز را از سازمان مربوطه دریافت کرده‌است و ممکن است به دلایل مختلف در معرض جدایی از سازمان قرار بگیرد. از این رو در این مرحله سازمان بایستی با بکارگیری روش‌های موثر، مشتریانی که احتمال ریزش دارند را شناسایی کرده و برای حفظ و نگهداشت آنها تلاش کند.

امروزه اصطلاح ریزش[88] مشتری به خوبی در کسب‌وکارهای مختلف شناخته شده‌است. مشتریانی ریزشی یا ریزش‌کننده[89] محسوب می‌شوند که احتمالا پس از استفاده از خدمت‌های ارائه‌شده علاقه خود را به ادامه کار با آن کسب‌وکار از دست می‌دهند و برای استفاده از خدمات آتی به سراغ سازمان رقیب می‌روند(کندل، ۲۰۱۹). درواقع احتمالا سفر مشتری نیمه‌تمام باقی می‌ماند و این مشتری به دسته مشتری‌های وفادار نمی‌پیوندد.

مسئله ریزش مشتری تاثیر زیادی بر سود سازمان مربوطه دارد و یکی از مهمترین موضوعات در کسب وکارها با بازار اشباع‌شده می‌باشد (احمد و خان و بسیط و لی، ۲۰۲۲). بسیاری از کسب‌وکارها در حوزه‌های خدمات مالی، خدمات فروش بلیط خطوط هوایی، خدمات بیمه، بازی‌های آنلاین و حوزه بانکی تلاش می‌کنند تا هرچه بیشتر روابط بلندمدت خود را با مشتریان موجود حفظ کنند. مشتریان پایدار علاوه بر آنکه به‌عنوان مشتریان بلندمدت برای سازمان سودمند هستند، سفیران تبلیغاتی موثر کسب‌وکارها در بازار محسوب می‌شوند(امین و عبدیات و ادمان و لو و انوار، ۲۰۱۹). از این رو روش‌های پیش‌بینی ریزش مشتریان (CCP[90]) در این حوزه از کسب‌وکارها بسیار سودمند است. بنابراین فرآیند پیش‌بینی ریزش باید زودتر از ریزش انجام شود تا سازمان زمان کافی برای پیشگیری از این رفتار را داشته باشد.

۲) پیشینه پژوهش

۲-۱) پیشینه نظری

در این بخش به تئوری‌ها، دیدگاه‌ها و رویکردهای موجود درمورد مسئله پیش‌بینی ریزش مشتری، مدیریت ارتباط با مشتری و یادگیری ماشین توضیح‌هایی مطرح شده‌است.

۳) پیش‌بینی ریزش مشتری

امروزه با توجه به افزایش رقابت بین کسب‌وکارها، بحث ریزش مشتریان اهمیت زیادی یافته‌است. ریزش مشتری یک

---

88. Churn
89. Churner
90. Customer Churn Prediction.

واقعیت دردناک است که همه کسب‌وکارها به نحوی با آن سروکار دارند حتی بزرگ‌ترین و موفق‌ترین شرکت‌ها هم از ریزش مشتری رنج می‌برند.

از دست دادن مشتری یکی از عواملی است که منجر به کاهش سودآوری شرکت شده و حتی ممکن است ضررهایی نیز از جنبه‌های مختلف مالی، اجتماعی و مواردی از این قبیل را به همراه داشته باشد. ازآنجاکه حفظ و نگهداشت مشتریان با ارزش فعلی سازمان نسبت به جذب مشتریان جدید هزینه بسیار کمتری دربر دارد، برای صاحبان کسب‌وکار ضروری است نگاهی دقیق به استراتژی حفظ مشتریان داشته باشند و برنامه دقیقی برای این منظور تدوین نمایند. شرکت‌های مشتری‌محور می‌بایست به ایجاد روابط بلندمدت با مشتریان خود توجه نمایند و علاوه بر تلاش برای به‌دست‌آوردن مشتریان جدید، بر روی اتخاذ رویکردهای مناسب برای حفظ مشتریان فعلی خود تمرکز کنند. با تدوین استراتژی‌های مؤثر حفظ مشتری، شرکت‌ها می‌توانند از هزینه‌های خدمات پایین‌تر توأم با درآمد بالاتر بهره‌مند گردند. امروزه هزینه جذب یک مشتری جدید به مراتب بیشتر از حفظ یک مشتری موجود است (بریتو و گوبیناف، ۲۰۲۱). پیش‌بینی ریزش مشتریان بطورکلی شامل مزایای زیر است:

- افزایش سود: حفظ مشتریان موجود آسان‌تر و مقرون به صرفه‌تر از جذب مشتریان جدید است.
- جلوگیری از اتلاف درآمد: اثر از دست دادن مشتری طولانی‌مدت است و می‌تواند همه چیز را از درآمد تا فرصت‌های رقابتی تحت تأثیر قرار دهد. از این رو کسب‌وکارها می‌خواهند که درآمد تقلیل‌یافته به سبب مشتری از دست رفته را به حداقل برسانند.
- کاهش هزینه‌های بازاریابی و فروش: هزینه نگهداری مشتری معمولا ۵ تا ۱۵ برابر کمتر از هزینه جذب مشتری جدید است.
- بهبود کیفیت خدمات به مشتری: با دانستن دلایل ریزش می‌توان کیفیت خدمات به مشتریان را ارتقا داد.
- حفظ مشتریان بیشتر: راه‌اندازی فعالانه کارزار تبلیغاتی و راهبردهایی که سبب می‌شود که ریزش مشتریان را به مقدار نزدیک به صفر رساند.
- برگرداندن مشتریان ریزشی: با شناسایی علل ریزش مشتریان قبلی و بکارگیری استراتژی‌های جذب مجدد، سعی برای برگرداندن مشتریان گذشته می‌شود.

با توجه به مزایای ذکر شده برای پیش‌بینی ریزش مشتریان، شرکت‌ها نیازمند یک سیستم پیش‌بینی‌کننده موثری هستند که بتوانند مشتریانی را که احتمال ریزش دارند را به‌طور خودکار تشخیص دهند. مدل‌های پیش‌بینی معمولا براساس الگوهای استفاده مشتری از سرویس و سایر عوامل که به‌طور مستقیم یا غیرمستقیم بر نظر او تاثیر می‌گذارد، ساخته می‌شوند (امین، ۲۰۲۰). در پیش‌بینی ریزش مشتریان درصد احتمال ترک سازمان برای هر مشتری محاسبه شده و در مجموعه داده‌های شرکت ثبت می‌شود. این فرآیند، سازمان را قادر می‌سازد که با تعریف آستانه احتمالی برای ریزش و همچنین خوشه‌بندی مشتریان، برنامه‌های استراتژیک متفاوتی را برای هر خوشه بکار گیرد تا بتواند ریزش‌کنندگان را حفظ کند (کستیک و سینیک، ۲۰۲۰).

۴) مدیریت ارتباط با مشتری

در چند سال اخیر مدیریت ارتباط با مشتریان در حوزه بازاریابی، فناوری اطلاعات و موارد مشابه توجه فراوانی را به خود جلب نموده است. مدیریت ارتباط با مشتری راهبردی تجاری برای بهبود منافع، مزایا و رضایت مشتری با سازماندهی بر اساس مشتری، پرورش رفتارهای رضایت‌بخش مشتری و اجرای فرایندهای مشتری محور می باشد(باران و استانک، ۲۰۱۹).

مدیریت ارتباط با مشتری مجموعه‌ای از فرآیندهای تعاملی است که با هدف رسیدن به تعامل مطلوب بین سرمایه‌گذاری‌های صنعت و تحقق نیازهای مشتری به‌منظور رسیدن به حداکثر سود است. (بالیوا و کاناوا و ساندر و کومار، ۲۰۲۱). بی شک می‌توان گفت مهمترین دارایی اغلب سازمان‌ها مشتریان آن هستند و مشتریان اغلب به خاطر ارتباط مستقیمی که با اقدام‌های یک سازمان دارند، منبع ارزشمندی برای فرصت‌ها و تهدیدهای مرتبط با آن سازمان محسوب می‌شوند(ونگ و شن و شو،۲۰۲۱).

۵) یادگیری ماشین

یادگیری ماشین زمینه نسبتا جدیدی از هوش مصنوعی است که درحال‌حاضر دوران رشد و تکامل خود را می‌گذراند و یادگیری ماشین یک زمینه تحقیقاتی بسیارفعال در علوم کامپیوتر است. علوم مختلفی ازقبیل هوش مصنوعی، روانشناسی، فلسفه، تئوری اطلاعات، آمار و اطلاعات، تئوری کنترل با یادگیری ماشین در ارتباط هستند. یادگیری ماشین عبارت است از اینکه چگونه می‌توان برنامه‌ای نوشت که از طریق تجربه یادگیری کرده و عملکرد خود را بهتر کند (احمد و افضل و صدیق و اجمد و خورشید، ۲۰۲۰).

در یادگیری ماشین با استفاده از تئوری اطلاعات، مدل‌های ریاضی ساخته می‌شود که می‌تواند برای استنتاج استفاده شوند. تکنیک‌های یادگیری ماشین در شرایطی مناسب است که هیچگونه دانش اولیه درمورد الگوهای داده‌ها وجود ندارد. به همین دلیل گاهی به این روش‌ها پایین به بالا می‌گویند (تکونبو و چاپار و تلونی و چیف و سیکان، ۲۰۲۲).

اهمیت مدیریت ریزش مشتری روزبه‌روز برای کسب‌وکارها بیشتر می‌گردد. پس با اعمال فرآیند یادگیری ماشین، می‌توان طی کاوش در تعامل‌های ثبت‌شده سازمان با مشتریان، به استخراج الگوهایی برای پیش‌بینی رفتار ریزش مشتری نائل آمد و مدیران را در اخذ تصمیم‌های لازم برای حفظ این مشتریان و کاهش روند ریزش آنها یاری داد (والوری و پاتیل، ۲۰۲۱). یکی از معیارهای پیش‌بینی ریزش مشتریان براساس بررسی روند رفتار آنها در مدت ارتباط با سازمان، مقایسه تعداد روز سپری‌شده از آخرین سفارش نسبت به بزرگ‌ترین فاصله زمانی ثبت سفارش توسط وی است. این پارامتر می‌تواند به‌عنوان معیاری برای پیش‌بینی ریزش در یادگیری ماشین مورداستفاده قرار گیرد. ریزش و از دست دادن مشتری به‌عنوان یک مسئله مهم و بحرانی در مدیریت ارتباط با مشتری پدیدار شده‌است. (مونیر و علقمدی و تیب و غلب ، ۲۰۲۲).

پیش‌بینی ریزش مشتریان یک مسئله بسیار پیچیده است که چالش‌هایی نظیر داده کثیف، نرخ ریزش کم، پنهان ماندن رویداد ریزش و غیره را به همراه دارد. روش‌های پیش‌بینی ریزش مشتریان به‌طور گسترده بر پایه روش‌های طبقه‌بندی[91] یادگیری ماشین انجام شده‌اند. در عملکرد این نوع طبقه‌بندی‌ها، با مشکل‌هایی نظیر ابعاد گسترده مجموعه داده‌ها[92] و نامتوازن بودن آنها که مانعی جهت پیش‌بینی دقیق است، روبرو هستیم (لئون و سیلوا و تباک، ۲۰۲۲).

---

[91]. Classification
[92]. Dataset

در پیش‌بینی ریزش معمولا طبقه‌بندی دودویی [93]انجام می‌شود که ریزش‌کننده‌ها یک گروه کوچک و اقلیت را تشکیل می‌دهند و غیرریزش‌کننده‌ها اکثریت را تشکیل می‌دهند. بنابراین معمولا مدل‌های یادگیری ماشین به علت کم‌بودن تعداد ریزش‌کننده‌ها تمایل دارند که همه مشتریان را غیرریزش‌کننده تشخیص دهند. اما این کار ارزشی ندارد و در عمل نتیجه خوبی حاصل نمی‌شود (اختر، ۲۰۲۲).

روش‌های گروهی[94] نتایج بسیار خوبی را در مدیریت مجموعه داده‌های متعادل در پیش‌بینی ریزش نشان می‌دهد. علاوه بر این، روش‌های گروهی به‌طور گسترده در مدل‌های پیش‌بینی ریزش استفاده می‌شود. استفاده از این روش‌ها دقیق و ساده است و برای بهره‌برداری از مدل‌های پیچیده با قدرت محاسباتی کم، بسیار مناسب هستند.

۶) پیشینه تجربی

در این پژوهش ۱۴ مقاله در دامنه موضوعی مورد مطالعه، بررسی شده و پس از بررسی مجموعه داده‌ها و شاخص‌های ارزیابی استفاده شده در آنها، به یک جمع‌بندی برای مدل پیشنهادی رسیده است. جدول ۱ کارهای پیشین بررسی‌شده را معرفی می‌کند.

جدول ۱. خلاصه مقالات بررسی شده

منابع	روش‌های ارزیابی	دقت	صنعت	روش‌های پیش‌بینی	سال
(هدیباس و ماتو و پیلای و رزول، ۲۰۱۷)	Accuracy	۷۰ درصد	شرکت اشتراک دهی	Deep Feed-Forward Networks	۲۰۱۷
(ژانگ و لی و مو و تانگ، ۲۰۱۸)	Accuracy, Precision, Recall, F1-Score, and AUC	۷۹.۱ درصد	مخابرات	Deep ANN, Machine Learning Algorithms	۲۰۱۸
(مالدونادو و لوپز و ورتی، ۲۰۲۰)	Accuracy, AUC, Precision, Recall, F1-Score	۸۲ درصد	بیمه	Deep and Shallow Learning	۲۰۲۰
(کالزدا و اکراسدوتیر و باسنز، ۲۰۲۰)	Area Under Curve of ROC(AUC) and Complexity	۹۲ درصد	مخابرات	Transfer Learning of Ensemble	۲۰۲۰
(سی و یانگ و نیو و سو و تاسی و ژنگ، ۲۰۲۰)	Area Under Curve of ROC(AUC)	۸۵ درصد	مخابرات	Ensemble Algorithm	۲۰۲۰
(دومیتروسو، ۲۰۲۱)	Accuracy and Precision of Classification	۸۱ درصد	مخابرات	Begging and Neural Network	۲۰۲۱
(کریستی، ۲۰۲۱)	Accuracy, Cost, Profit	۸۵ درصد	مخابرات	Profit Tree	۲۰۲۱
(لی و چونگ و فنگ و مو، ۲۰۲۱)	AUC, EMPC	۹۰ درصد	مخابرات	Minimax Probability Machines	۲۰۲۱
(سلوریا و پیانهریت و جونیور، ۲۰۲۱)	AUC, Tenlift AUPR	۸٤ درصد	مخابرات	Similarity Forests	۲۰۲۱
(جین و یاداو و مانو، ۲۰۲۲)	Accuracy, Kappa, Recall, etc.	۵۹.۲٤ درصد	بانک	Cross-Company Just-in-time Approach	۲۰۲۲
(حسینی و شرنجنی و اکبرآبادی، ۲۰۲۲)	AUC	۸۲ درصد	بانک	Multi-objective–ant Colony Optimization	۲۰۲۲
(مستوزلا، ۲۰۲۲)	Top Decile Lift	۷۵ درصد	بانک	Graph Theory	۲۰۲۲
(هارادا، ۲۰۲۲)	G means and AUC	٦۰ درصد	بانک	Boosted-Stacked Learners and Bagged-Stacked Learners	۲۰۲۲
(جین و خومتتا و استریستوا، ۲۰۲۲)	Accuracy, Kappa, MAE, Coverage Case, RMSE	۷۱ درصد	بانک	Logistic Regression and Logit Boost	۲۰۲۲

93. Binary
94. Ensemble Methods

همانطورکه بررسی پژوهش‌های پیشین نشان می‌دهد، تأکید مقالات در یک یا دو روش از روش‌های استخراج ویژگی است، آنها همچنین همزمان روش‌های یادگیری ماشین و روش یادگیری عمیق را در استخراج ویژگی درنظر نگرفته‌اند. مقالات اخیر نشان می‌دهد که معیارهای ارزیابی برای ایجاد یک سیستم تشخیص ریزش کافی نیستند. در طبقه-بندی ریزش مشتریان، همیشه باید مدل‌ها را با عملکرد خود نسبت به ماهیت نامتعادل مجموعه داده‌ها مقایسه کرد (لی و ژانگ و مائو، ۲۰۱۸). از آنجاکه مشکل ریزش به دلیل عدم دسترسی به سابقه داده‌های ریزش‌گر بسیار نامتعادل است، بنابراین لازم است مدل‌های طبقه‌بندی ریزش با کلیه معیارهای مرتبط ارزیابی شوند.

۷) طرح مسئله و راه‌حل پیشنهادی

زبان پایتون با مجموعه‌ای از کتابخانه‌های علم داده و یادگیری ماشین عرضه می‌شود که می‌توان از آنها برای پیش‌بینی براساس ویژگی یا ویژگی‌های مختلف یک مجموعه داده استفاده کرد. این زبان یکی از پرکاربردترین زبان‌های برنامه‌نویسی برای تجزیه و تحلیل داده‌های مالی است که دارای کتابخانه‌های مفید و عملکرد داخلی فراوانی است. در این مقاله به بررسی مشتریان یک بانک پرداخته شده‌است و نتایج حاصل از بررسی نشان می‌دهد که چگونه یک بانک می‌تواند ریزش مشتری را بر اساس ویژگی‌های مختلف مشتری مانند سن، جنسیت، جغرافیا و موارد دیگر پیش‌بینی کند. کتابخانه SCIKIT-LEARNپایتون یکی از این ابزارها می‌باشد و در این مقاله از این کتابخانه برای پیش‌بینی ریزش مشتری استفاده شده‌است.

۸) فاز اول (تحلیل کسب‌وکار)

فاز اول شناخت کسب‌وکار و مطالعه بر روی عواملی است که در آن نوع کسب‌وکار مورد بررسی به فرآیند پیش‌بینی ریزش کمک می‌کند. لازم به ذکر است شاخص‌هایی که در یک بانک بر روی ریزش تاثیرگذار هستند با شاخص‌های یک سازمان مالی و مخابراتی متفاوت هستند. از این رو با استفاده از جلسه‌های مصاحبه و روش‌های طوفان فکری و موجودیت انواع داده در سازمان، به جمع‌آوری شاخص‌های موثر و داده‌های مربوط به آنها پرداخته خواهد شد. به علاوه در این فاز آستانه تحمل ریزش شرکت یعنی درصد احتمالی که شرکت برای یک مشتری ریزش‌کننده بالقوه تعیین می‌کند، تعریف می‌شود.

مجموعه داده‌ای که برای توسعه یک مدل پیش‌بینی ریزش مشتری برای یک بانک درنظر گرفته شده‌است، دارای۱۴ستون است که به‌عنوان ویژگی یا متغیر شناخته می‌شود. با توجه به شکل ۱ ، ۱۳ ستون اول متغیر مستقل هستند، درحالیکه ستون آخر، متغیر وابسته است که دارای مقدار باینری ۱ یا ۰ است.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	RowNumb	Customer	Surname	CreditSco	Geograph	Gender	Age	Tenure	Balance	NumOfPr	HasCrCard	IsActiveM	Estimated	Exited	
2	1	15634602	Hargrave	619	France	Female	42	2	0	1	1	1	101348.9	1	
3	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.6	0	
4	3	15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.6	1	
5	4	15701354	Boni	699	France	Female	39	1	0	2	0	0	93826.63	0	
6	5	15737888	Mitchell	850	Spain	Female	43	2	125510.8	1	1	1	79084.1	0	
7	6	15574012	Chu	645	Spain	Male	44	8	113755.8	2	1	0	149756.7	1	
8	7	15592531	Bartlett	822	France	Male	50	7	0	2	1	1	10062.8	0	
9	8	15656148	Obinna	376	Germany	Female	29	4	115046.7	4	1	0	119346.9	1	
10	9	15792365	He	501	France	Male	44	4	142051.1	2	0	1	74940.5	0	
11	10	15592389	H?	684	France	Male	27	2	134603.9	1	1	1	71725.73	0	
12	11	15767821	Bearce	528	France	Male	31	6	102016.7	2	0	0	80181.12	0	
13	12	15737173	Andrews	497	Spain	Male	24	3	0	2	1	0	76390.01	0	
14	13	15632264	Kay	476	France	Female	34	10	0	2	1	0	26260.98	0	
15	14	15691483	Chin	549	France	Female	25	5	0	2	0	0	190857.8	0	
16	15	15600882	Scott	635	Spain	Female	35	7	0	2	1	1	65951.65	0	
17	16	15643966	Goforth	616	Germany	Male	45	3	143129.4	2	0	1	64327.26	0	
18	17	15737452	Romeo	653	Germany	Male	58	1	132602.9	1	1	0	5097.67	1	
19	18	15788218	Henderso	549	Spain	Female	24	9	0	2	1	1	14406.41	0	
20	19	15661507	Muldrow	587	Spain	Male	45	6	0	1	0	0	158684.8	0	
21	20	15568982	Hao	726	France	Female	24	6	0	2	1	1	54724.03	0	

Churn_Modelling

شکل ۱. نموداری برای نمونه

در شکل ۱ ، مقدار ۱ به حالتی اشاره دارد که مشتری پس از ۶ ماه بانک را ترک کرده‌است و مقدار ۰ موردی است که مشتری بعد از ۶ ماه بانک را ترک نکرده‌است. این به‌عنوان یک مشکل طبقه‌بندی باینری شناخته می‌شود که در آن شما فقط دو مقدار ممکن برای متغیر وابسته دارید. در این مورد، مشتری یا پس از ۶ ماه بانک را ترک می‌کند یا ترک نمی‌کند. ذکر این نکته حائز اهمیت است که داده‌های متغیرهای مستقل ۶ ماه قبل از داده‌های متغیر وابسته جمع‌آوری‌شده‌است زیرا وظیفه توسعه یک مدل یادگیری ماشینی است که بتواند پیش‌بینی کند که آیا مشتری پس از ۶ ماه بانک را ترک می‌کند یا خیر. برای حل این مشکل می‌توان از الگوریتم‌های طبقه‌بندی مدل‌های یادگیری ماشین استفاده کرد.

۹) فاز دوم (وارد کردن کتابخانه‌ها)

در این فاز کتابخانه‌های Numpy، pandas ، Matplotlib موردنیاز است.

۱۰) فاز سوم (انتخاب ویژگی)

بعد از جمع‌آوری مجموعه داده‌های موردنیاز براساس ویژگی‌های کلیدی و نوع صنعت، در این فاز به انجام پیش‌پردازش‌های اولیه ازجمله حذف نویزها و داده‌های فراموش‌شده و تمیزکردن مجموعه داده‌ها پرداخته می‌شود. با توجه به شکل ۱ دارای مجموعه ویژگی‌های زیر هستیم:

- Row Number: مربوط به شماره رکورد یا ردیف است و تأثیری بر خروجی ندارد و این ستون حذف خواهد شد.
- Customer ID: حاوی مقادیر تصادفی است و تأثیری بر خروج مشتری از بانک ندارد و این ستون حذف خواهد شد.
- Surname: نام خانوادگی مشتری تأثیری بر تصمیم او برای ترک بانک ندارد و این ستون نیز حذف خواهد شد.

- Credit Score: میزان امتیاز مشتری در اعتبارات می‌تواند بر روی ریزش مشتری تأثیر بگذارد زیرا مشتری با امتیاز اعتباری کمتر بانک را ترک می‌کند.
- Geography: موقعیت یک مشتری می‌تواند بر تصمیم او برای ترک بانک تأثیر بگذارد.
- Gender: تاثیر جنسیت در خروج مشتری از بانک تاثیرگذار است.
- Age: این ویژگی با ریزش مشتری مرتبط است زیرا مشتریان مسن‌تر، کمتر از افراد جوان‌تر بانک خود را ترک یا عوض می‌کنند.
- Tenure: این ویژگی به تعداد سال‌هایی اشاره دارد که یک فرد مشتری بانک بوده‌است. به‌طورمعمول، مشتریان مسن‌تر وفادارتر هستند و کمتر بانک را ترک می‌کنند.
- Balance: یک شاخص بسیار خوب از ریزش مشتری است زیرا افرادی که موجودی بیشتری در حساب‌های خود دارند، در مقایسه با افرادی که موجودی کمتری دارند، کمتر بانک را ترک می‌کنند.
- NumOfProducts: به تعداد محصولاتی که مشتری از طریق بانک خریداری کرده‌است، اشاره دارد.
- HasCrCard: این ویژگی بیانگر این موضوع است که آیا مشتری کارت اعتباری دارد یا خیر. این ستون نیز مرتبط است زیرا افرادی که کارت اعتباری دارند، کمتر بانک را ترک می‌کنند.
- IsActiveMember: مشتریان فعال کمتر بانک را ترک می‌کنند بنابراین این ویژگی در تعیین ریزش مشتری کمک‌کننده است.
- Estimated Salary: این ویژگی همانند موجودی، افرادی که حقوق کمتری دارند در مقایسه با افرادی که حقوق بالاتری دارند، احتمال بیشتری دارد بانک را ترک کنند.
- Exited: در ارزیابی ریزش مشتری این ویژگی لحاظ خواهد شد.

پس از مشاهده دقیق ویژگی‌ها، ستون‌های Row Number، Customer ID و Surname از مجموعه ویژگی‌ها حذف می‌شوند و تمام ستون‌های باقی‌مانده به یک شکل به ریزش مشتری کمک می‌کنند.

۱۱) فاز چهارم (تبدیل داده‌ها)

الگوریتم‌های یادگیری ماشین با داده‌های عددی، کارآیی بسیار بهتری را ارائه می‌دهند. درمجموع داده‌های مورد مطالعه، دو ستون دسته‌بندی داریم: جغرافیا و جنسیت. این دو ستون حاوی داده‌ها در قالب متنی هستند و باید آنها به ستون‌های عددی تبدیل شوند. یکی از راه‌های تبدیل ستون‌های دسته‌بندی به ستون‌های عددی این است که می‌توان هر دسته را با یک عدد جایگزین کرد. به‌عنوان‌مثال: در ستون جنسیت، زن را می‌توان با مقدار ۰ و مرد را با مقدار ۱ جایگزین کرد یا بالعکس. این برای ستون‌هایی با دو دسته کار می‌کند.

برای ستونی مانند جغرافیا با سه دسته یا بیشتر، می‌توان از مقادیر ۰ و ۱ و ۲ برای سه کشور فرانسه، آلمان و اسپانیا استفاده کرد. لذا اگر این کار انجام گیرد، الگوریتم‌های یادگیری ماشین فرض می‌کنند که یک رابطه ترتیبی بین این سه کشور وجود دارد. به‌عبارت‌دیگر، الگوریتم فرض می‌کند که ۲ بزرگ‌تر از ۱ و ۰ است که درواقع از نظر کشورهای

زیربنایی که اعداد نشان می‌دهند، این‌طور نیست. یک راه بهتر برای تبدیل ستون‌های دسته‌بندی به ستون‌های عددی، استفاده از one-hot encoding است. در این فرآیند دسته‌ها فرانسه، آلمان، اسپانیا هستند و آنها به‌عنوان یک ستون مجزا درنظر گرفته شده‌اند و در هر ستون از ۱ برای تعیین اینکه دسته برای ردیف فعلی وجود دارد، استفاده شده‌است و در غیر این صورت از ۰ به‌عنوان جایگزین استفاده شده‌است.

۱۲) فاز پنجم (پیش‌پردازش داده‌ها)

در فرآیندهای داده‌کاوی مانند طبقه‌بندی و خوشه‌بندی، نیاز داریم تا داده‌ها برای الگوریتم آماده شوند. زیرا معمولا نمی‌توان داده‌ها را به‌صورت خام به الگوریتم‌های داده‌کاوی و یادگیری ماشین تزریق کرد. از آنجاییکه داده‌ها معمولا از منابعی تهیه می‌شوند که این منابع بدون توجه به فرآیندهای داده‌کاوی، داده‌ها را تولید یا نگهداری کرده‌اند، نیاز است داده‌ها با توجه به شرایط و مسئله، به داده‌های مناسب جهت تزریق به الگوریتم‌های داده‌کاوی آماده شوند. برای آماده‌سازی داده‌ها، نیاز است تا آنها را از شکل و حالت اولیه خارج کرده و به شکلی که برای الگوریتم مناسب باشد تبدیل کنیم.

همچنین داده‌های موجود معمولا دارای زواید مختلفی هستند که ممکن است الگوریتم را دچار خطا کنند. در داده‌کاوی نیاز است تا داده‌های اضافی که به مسئله و الگوریتم کمکی نمی‌کنند، حذف گردند. در این مرحله از پیش‌بینی، داده‌ها آماده است و می‌توان مدل یادگیری ماشینی را برای آموزش آن درنظر بگیریم. در ابتدا باید متغیری که پیش‌بینی می‌شود از مجموعه داده‌ها جدا گردد. این فرآیند شامل تمام ستون‌ها به‌جز ستونی است که باید پیش‌بینی شود. ستون Exited برچسب داده‌ها را مشخص می‌کند بنابراین در ادامه عملکرد مدل یادگیری ماشینی با این ستون ارزیابی می‌گردد. برای پیاده‌سازی مدل، داده‌ها به یک مجموعه آموزشی و آزمایشی تقسیم شده‌است. مجموعه آموزشی حاوی داده‌هایی است که برای آموزش مدل یادگیری ماشین استفاده می‌شود و مجموعه آزمایش برای ارزیابی اینکه مدل معرفی‌شده چقدر خوب است، استفاده خواهد شد. ۲۰ درصد از داده‌ها را برای مجموعه آزمایش و از ۸۰ درصد باقیمانده برای مجموعه آموزشی استفاده شده‌است.

۱۳) فاز ششم (آموزش الگوریتم یادگیری ماشین)

در این مرحله از یک الگوریتم یادگیری ماشینی برای پیش‌بینی، استفاده شده‌است که الگوها یا روندها را در داده‌های آموزشی شناسایی می‌کند. این مرحله به آموزش الگوریتم معروف است.

ویژگی‌ها و خروجی صحیح به الگوریتم داده می‌شود و براساس آن داده‌ها، الگوریتم یاد خواهد گرفت که ارتباط بین ویژگی‌ها و خروجی‌ها را بیابد. پس از آموزش الگوریتم، از آن برای پیش‌بینی داده‌های جدید استفاده شده‌است. چندین الگوریتم یادگیری ماشین وجود دارد که می‌توان از آنها برای انجام چنین پیش‌بینی‌هایی استفاده کرد. در این مقاله از الگوریتم جنگل تصادفی[۹۵] استفاده شده‌است که در ادامه به آن پرداخته خواهدشد.

۱۴) الگوریتم جنگل تصادفی[۹۶]

جنگل تصادفی یک روش یادگیری ترکیبی برای طبقه‌بندی داده‌ها می‌باشد که از تعداد زیادی درخت تصمیم در زمان آموزش

[95]. Random forest
[96]. Random forest

تشکیل شده‌است و خروجی این الگوریتم برای طبقه‌بندی مجموعه‌ای از داده‌ها مورداستفاده قرار می‌گیرد. در الگوریتم جنگل تصادفی درواقع مجموعه‌ای از درخت‌های تصمیم، با هم یک جنگل را تولید می‌کنند و این جنگل می‌تواند تصمیم‌های بهتری را نسبت به یک درخت اتخاذ نماید. در الگوریتم جنگل تصادفی به هر کدام از درخت‌ها، یک زیرمجموعه‌ای از داده‌ها تزریق می‌شود. درخت‌ها می‌توانند تصمیم بگیرند و مدل طبقه‌بندی خود را بسازند. در هنگام پیش‌بینی، هر کدام از این الگوریتم‌های یادگرفته‌شده، یک نتیجه را پیش‌بینی می‌کنند. در نهایت الگوریتمِ جنگلِ تصادفی، می‌تواند با استفاده از رای‌گیری، آن طبقه‌ای را که بیشترین رای را آورده‌است، انتخاب کرده و به‌عنوان طبقه نهایی جهت انجام عملیات طبقه‌بندی قرار دهد. همچنین جنگل‌های تصادفی برای درختان تصمیمی که در فرآیند آموزش دچار بیش‌برازش می‌شوند، مناسب هستند و جنگل تصادفی معمولا بهتر از درخت تصمیم است اما این بهبود عملکرد تا حدی به نوع داده هم بستگی دارد (رائو، ۲۰۱۶).

۱۵) نحوه عملکرد الگوریتم جنگل تصادفی

نحوه کارکردن الگوریتم جنگل تصادفی به شرح زیر است:

۱- ابتدا با انتخاب نمونه‌های تصادفی از یک مجموعه داده الگوریتم شروع می‌شود.

۲- این الگوریتم برای هر نمونه، یک درخت تصمیم‌گیری ایجاد می‌کند. سپس نتیجه پیش‌بینی از هر درخت تصمیم به‌دست می‌آید.

۳- رای‌گیری برای هر نتیجه پیش‌بینی‌شده انجام می‌گیرد.

۴- بیشترین نتیجه پیش‌بینی، به‌عنوان پیش‌بینی نهایی انتخاب می‌گردد.

در شکل ۲ کد الگوریتم جنگل تصادفی نیز قابل مشاهده است.

```
Precondition: A training set S := (x_1, y_1), ..., (x_n, y_n), features F, and number
  of trees in forest B.
1 function RandomForest(S, F)
2     H ← ∅
3     for i ∈ 1, ..., B do
4         S^(i) ← A bootstrap sample from S
5         h_i ← RandomizedTreeLearn(S^(i), F)
6         H ← H ∪ {h_i}
7     end for
8     return H
9 end function
10 function RandomizedTreeLearn(S, F)
11     At each node:
12         f ← very small subset of F
13         Split on best feature in f
14     return The learned tree
15 end function
```

شکل۲. کد الگوریتم جنگل تصادفی(لی و چونگ و فنگ و مو، ۲۰۲۱)

۱۶) الگوریتم جایا [۹۷]

یکی از الگوریتم‌های فراابتکاری مناسبی که برای حل مسئله‌های بهینه‌سازی محدود و نامحدود استفاده می‌شود، الگوریتم بهینه‌سازی جایا است. الگوریتم جایا، در فرآیند جستجو سعی می‌کند با فرار از بدترین راه‌حل، با اجتناب از شکست، برای

---

97. Jaya

یافتن راه‌حل‌های بهینه به بهینه‌ترین راه‌حل ممکن، نزدیک‌تر شود. الگوریتم جایا روشی ساده و جدید به‌منظور حل مسائل بهینه‌سازی می‌باشد. این الگوریتم بر این اصل استوار است که برای رسیدن به بهترین جواب باید راه‌حل‌های موجود در هر تکرار به سمت بهترین راه‌حل ارائه‌شده حرکت کرده و از بدترین راه‌حل در همان تکرار نیز دور شوند. از مزایای الگوریتم جایا می‌توان به سادگی، عدم نیاز به هیچ پارامتر کنترلی و سرعت آن اشاره کرد. به بیان دیگر الگوریتم جایا، الگوریتم فر ابتکاری است که با استفاده از مناسب‌ترین مفاهیم تحریک می‌شود و با اجتناب از راه‌حل غیربهینه، به سمت راه‌حل‌های بهینه می‌روند. در الگوریتم جایا تابع F(x) تابع هدف مسئله بوده و الگوریتم جایا قصد یافتن مینیمم (ماکزیمم) آن را دارد. در هر تکرار k، الگوریتم جایا دارای n متغیر تصمیم‌گیری و N_param راه‌حل پیشنهادی است.

مراحل الگوریتم جایا به شریح زیر است:

۱- پارامترهای الگوریتم جایا مانند اندازه جمعیت N و حداکثر تعداد تکرار Kmax باید در تمام مسائل بهینه‌سازی مقداردهی اولیه شوند. داده‌های تعریف شده باید از مجموعه داده‌های مسئله استخراج شوند و همچنین تابع هدف و راه‌حل پیش‌بینی‌شده معمولا در این مرحله بایستی مشخص گردد. تعریف داده‌های ورودی و مشخص‌شدن حد بالا و حد پایین مجموعه داده‌ها باید در این مرحله صورت پذیرد.

۲- در این مرحله تشکیل جمعیت اولیه صورت می‌پذیرد. براساس داده‌های موجود در مرحله اول، یک جمعیت اولیه به‌صورت تصادفی به‌گونه‌ای که محدودیت‌ها رعایت شود، تشکیل می‌گردد که جامعه تشکیل‌شده شامل N راه d بعدی است.

۳- محدودیت‌های مسئله در این مرحله مورد بررسی قرار می‌گیرد و بهترین و بدترین راه‌حل‌ها در این مرحله مشخص می‌شود.

۴- الگوریتم جایا به‌صورت تصادفی هر متغیر تصمیم موجود در هر راه‌حل را در هر تکرار تغییر می‌دهد. به عبارت دیگر تابع هدف مسئله مشخص می‌شود.

۵- برای استفاده از الگوریتم مورد نظر می‌بایست بدترین و بهترین جواب مشخص شود. بنابراین هدف الگوریتم جایا با استفاده از شرایط فوق تحقق می‌یابد.

۶- الگوریتم جایا روی مجموعه جواب‌های موجود باید اعمال گردد تا در مجموعه جواب‌های محاسبه‌شده، بهبود حاصل گردد.

۷- پس از به‌رو رسانی مجدد جواب‌ها، شماره تکرار یک واحد افزایش می‌یابد و شرط همگرایی بررسی می‌شود. اگر شرط همگرایی در الگوریتم برآورده شده‌بود این قاعده پایان می‌یابد و بهترین جواب چاپ می‌شود در غیر این صورت مجدد مرحله ۵ مورد بررسی قرار می‌گیرد. در شکل ۳ کد الگوریتم جایا نیز قابل مشاهده است.

1: *Initialize input Jaya parameters:* r_1, r_2, X_j^{UB}, X_j^{LB}, N, d, k_{max}
2: --------- *Initializing the initial population*---------
3: **For** $i = 1:1:N$ **Do** /* *Each solution* */
4: **For** $j = 1:1:d$ **Do** /* *Each variable*/
5: **Calculate:** $X_{i,j} = X_j^{LB} + (X_j^{UB} - X_j^{LB}) \times U(0,1)$
6: **Calculate:** $f(X_i)$
7: ---------- *Define the best and worst solutions in the population*----------
8: itr (k)=1
9: **while** $(k < k_{max})$ **do**
10: *Determine the best solution in the population* (X_{best})
11: *Determine the best solution in the population* (X_{worst})
12: **For** $i = 1:1:N$ **Do**
13: **For** $j = 1:1:d$ **Do**
14: Set $r_1 \in (0,1)$
15: Set $r_1 \in (0,1)$
17: **Calculate:** $X'_{i,j} = X_{i,j} + r_1 \times (X_{best,j} - |X_{i,j}|) - r_2 \times (X_{worst,j} - |X_{i,j}|)$
18: **End For**
19: **If** $f(X'_i) \leq f(X_i)$ **then**
20: $X'_i = X_i$ *(update process)*
21: **End If**
22: **End For**
23: $k = k + 1$
24: **End while**

شکل ۳. کد الگوریتم جایا (ویلرت، ۲۰۲۱).

با توجه به الگوریتم جنگل تصادفی نتایج حاصل از ارزیابی رضایت‌بخش نبود بنابراین روش پیشنهادی ترکیبی از الگوریتم جنگل تصادفی و الگوریتم جایا می‌باشد که مراحل آن به شرح زیر است:

مرحله اول مقداردهی اولیه: در این مرحله یک تابع توزیع به‌صورت تصادفی برای تعیین پارامترهای الگوریتم *Jaya* مانند اندازه جمعیت N که بیانگر تعداد راه‌حل‌های ممکن است و حداکثر تعداد تکرار الگوریتم که با $kmax$ نمایش داده می‌شود، تعریف می‌گردد. همچنین مجموعه پارامترهای الگوریتم جنگل تصادفی تعریف می‌گردد.

مرحله دوم مشخص‌سازی تابع هدف مناسب: بهترین مقدار برای تابع هدف در مرحله i^{th} به ازای داده ورودی x_i به‌صورت $f(i)$ تعریف می‌گردد. به دلیل اینکه الگوریتم جایا هر کدام از راه‌حل‌های انتخابی x_i را برای طراحی درست طبقه‌بندی الگوریتم جنگل تصادفی با جایا بررسی می‌کند، انتخاب صحیح تابع هدف مناسب جز ضروریات است، که mتعداد نمونه‌های مشاهده‌شده در فرآیند تست است و $Error(p)$ برای به‌دست آوردن میزان خطا مقدار واقعی از مقدار برآوردشده، مورد استفاده قرار می‌گیرد.

در این مقاله به دقت طبقه‌بندی و تعداد ویژگی‌های انتخاب شده، به‌عنوان دو پارامتر اساسی در تعریف تابع هدف مناسب می‌توان اشاره کرد. بنابراین انتخاب راه‌حل با بالاترین دقت تصمیم‌گیری و کمترین تعداد ابعاد، به‌عنوان دو پارامتر برای برآورد بهترین تابع هدف تعریف می‌شود و میزان مناسب بودن تابع هدف به کمک رابطه (۱) ارزیابی می‌گردد:

$$finess(i) = \frac{\sum_{i=1}^{m}(Error(p))^2}{m}.p == 1.2 \ldots . m \qquad \text{رابطه ۱)}$$

مرحله سوم مشخص‌سازی بهترین و بدترین کاندیدها: بعد از ارزیابی تابع هدف مناسب، بهترین و بدترین کاندیدها که به ترتیب با xbest و xworst تعریف می‌گردد، که این مقدارها از طریق ماکزیمم و مینیمم مقداری که تابع هدف محاسبه کرده‌است، تعریف شده‌است.

مرحله چهارم بهبود راه‌حل: در این مرحله بهترین و بدترین راه‌حل‌های به‌دست‌آمده از مرحله قبل ذخیره می‌شوند و سایر راه‌حل‌های به‌دست‌آمده به‌روزرسانی می‌شوند و مجدد تابع هدف مناسب به‌وسیله مقادیر جدید به‌دست‌آمده، تخمین زده می‌شود.

مرحله پنجم بررسی راه‌حل: اگر مقدار تابع هدف به‌دست‌آمده در مرحله جدید بهتر از مرحله قبلی باشد، مقدار جدید به‌دست‌آمده جایگزین مقدار قبلی می‌شود در غیر این صورت الگوریتم برای بار دیگر تکرار می‌شود.

مرحله ششم: کلیه مرحله‌های ۲ تا ۵ مادامی که تعداد دفعات تکرار الگوریتم برابر *kmax* شود، انجام می‌گردد.

۱۷) فاز هفتم (ارزیابی مدل)

بعد از آموزش، ارزیابی صورت می‌گیرد تا بررسی شود که الگوریتم انتخابی در پیش‌بینی ریزش مشتری چقدر خوب عمل می‌کند. برای ارزیابی عملکرد یک الگوریتم طبقه‌بندی، متداول‌ترین معیارهای مورداستفاده عبارتنداز: دقت[۹۸]، یادآوری[۹۹] و صحت[۱۰۰] . در معیار دقت حداکثر مقدار این معیار یک و یا ۱۰۰ درصد است و حداقل مقدار آن صفر است و هرچه مواردی که برنامه به غلط پیش‌بینی کرده‌است که به آن مثبت کاذب می‌گویند نسبت به پیش‌بینی‌های درست که به آن مثبت واقعی[۱۰۱] می‌گویند بیشتر باشد، مقدار دقت کمتر خواهد شد و طبق رابطه (۲) قابل محاسبه است.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{(رابطه۲)}$$

که در آن TP مثبت واقعی و FP مثبت کاذب است.

در معیار یادآوری حداکثر مقدار این معیار یک و یا ۱۰۰ درصد است و حداقل مقدار آن صفر است و هرچه مواردی که برنامه پیش‌بینی نکرده‌است که به آن منفی کاذب [102] می‌گویند نسبت به پیش‌بینی‌های درست که به آن مثبت واقعی می‌گویند بیشتر باشد، مقدار یادآوری کمتر خواهد شد و طبق رابطه (۳) قابل محاسبه است.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{(رابطه ۳)}$$

که در آن TP مثبت واقعی و FN منفی کاذب است.

در معیار صحت تعداد مواردی که درست پیش‌بینی شده‌است که به آن مثبت واقعی می‌گویند بر تعداد کل پیش‌بینی‌هایی که انجام شده‌است، تقسیم می‌گردد و طبق رابطه (۴) قابل محاسبه است.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \qquad \text{(رابطه ٤)}$$

که در آن TP مثبت واقعی، TN منفی واقعی[۱۰۳]، FP مثبت کاذب و FN منفی کاذب است.

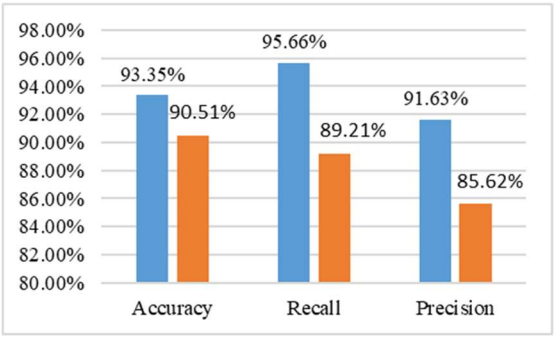
---

98. Precision
99. Recall
100. Accuracy
101. True Positive
102. False Negative
103. True Negative

در کتابخانه scikit-learn پایتون، می‌توانیم از توابع داخلی برای یافتن همه این مقادیر استفاده کنیم و اگر مدل کارآیی لازم را نداشت، دوباره به فاز پنج برمی‌گردیم و تنظیمات پارامترهای اولیه بر روی مدل قبلی برای رسیدن به نتیجه مطلوب تغییر پیدا خواهد کرد. نتایج حاصل از مقاله به ترتیب در جدول ۲ و شکل ۴ آورده شده‌است.

جدول ۲. نتایج به‌دست‌آمده استفاده از الگوریتم جنگل تصادفی و روش پیشنهادی

	Accuracy	Recall	Precision
جنگل تصادفی	٪ ۹۰.۵۱	٪ ۸۹.۲۱	٪ ۸۵.۶۲
مدل پیشنهادی	٪ ۹۳.۳۵	٪ ۹۵.۶۶	٪ ۹۱.۴۱

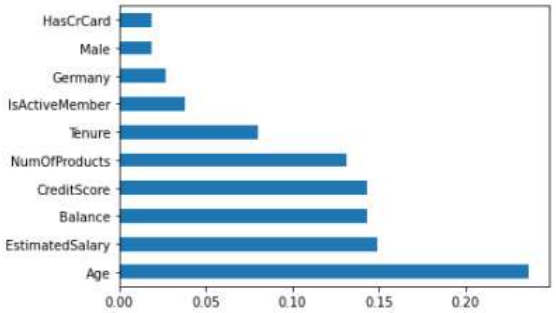


شکل ٤. مقایسه نتایج به‌دست‌آمده از الگوریتم جنگل تصادفی و روش پیشنهادی

همانطورکه قابل ملاحظه است، نتایج حاصل از مدل پیشنهادی در مقاله به ترتیب در معیارهای Precision، Recall و Accuracy حدودا به میزان ۳ درصد، ۶ درصد و ۷ درصد از الگوریتم جنگل تصادفی بالاتر می‌باشد.

۱۸) فاز هشتم (ارزیابی ویژگی)

به‌عنوان آخرین مرحله در پیش‌بینی ریزش مشتری، بررسی می‌شود که کدام ویژگی مهمترین نقش را در شناسایی ریزش مشتری ایفا می‌کند. خوشبختانه طبقه‌بندی جنگل تصادفی[۱۰۴] دارای ویژگی با نام ویژگی‌های مهم[۱۰۵] است که حاوی اطلاعاتی درمورد مهمترین ویژگی‌های یک طبقه‌بندی معین است و خروجی آن به‌صورت شکل ۵ است.



شکل ۵. ارزیابی ویژگی‌های بررسی‌شده

104. Random Forest Classifier
105. Feature Importance

براساس این داده‌ها، ملاحظه می‌شود که ویژگی‌های سن، حقوق، دستمزد تخمینی مشتری و مانده‌حساب، به ترتیب بیشترین تأثیر را بر ریزش مشتری دارد.

۱۹) نتیجه‌گیری و پیشنهادات

با توجه به اهمیت حفظ مشتری در کسب‌وکارهای مختلف اعم از بانک‌ها، سازمان‌های مخابراتی، بیمه و حتی بازی‌های آنلاین، تحقیقات متعددی بر روی پیش‌بینی ریزش مشتریان و عوامل موثر بر روی آنها انجام شده‌است. در این تحقیق‌ها سعی شده‌است از راه‌های گوناگون ازجمله داده موجود در ارتباط با مشتریان و نحوه تعامل مشتریان با سازمان و همچنین تعامل مشتریان با یکدیگر به پیش‌بینی دقیق ریزش مشتریان دست یابند.

نحوه انجام فرآیند پیش‌بینی ریزش مشتریان بسیار اهمیت دارد چرا که انتخاب نوع داده‌ها و نوع مدل یادگیری، به شدت دقت پیش‌بینی را تحت تاثیر قرار می‌دهد. این فرآیند با شناخت کسب‌وکار و عوامل موثر بر ریزش مشتری در آن صنعت آغاز می‌شود. پس از مهندسی ویژگی‌ها، انتخاب مدل یادگیری و ارزیابی آن، بهترین مدل انتخاب می‌شود و براساس آن مشتریان ریزش‌گر سازمان انتخاب خواهد گردید. در نهایت سازمان برای حفظ مشتریان ریزش‌گر، استراتژی‌های مختلفی اتخاذ می‌کند. پیش‌بینی ریزش مشتری برای ثبات مالی بلندمدت یک بانک بسیار مهم است.

در این مقاله، با موفقیت یک مدل یادگیری ماشین ایجاد شد که می‌تواند ریزش مشتری را با دقت ۹۳.۳۵ درصد پیش‌بینی کند. نتایج حاصل از مدل پیشنهادی در مقاله به ترتیب در معیارهای Precision، Recall و Accuracy حدودا به میزان ۳ درصد، ۶ درصد و ۷ درصد از الگوریتم جنگل بالاتر می‌باشد. علاوه بر این، طبق بررسی‌های به عمل آمده در سازمان‌هایی که تعداد مشتریان زیادتراست وداده‌های مختلفی را دراختیار دارند، توانسته‌اند در پیش‌بینی ریزش مشتریان موفق‌تر عمل نمایند.

پیشنهاد می‌شود که برای توسعه تکنیک‌های داده‌کاوی در آینده از روش‌های مبتنی بر مشکل استفاده شود و بر پیش‌بینی نوع ریزش تمرکز گردد. علاوه بر این، می‌توان از مدل‌های ترکیبی استفاده نمود و آنها را با مدل‌های موجود مقایسه کرد. این امر به طراحی چندبعدی مجموعه داده‌های مشتری کمک می‌کند و به ابداع تکنیک‌های جدید مدیریت ریزش برای مجموعه داده‌های مختلف و مکان‌های جغرافیایی مختلف می‌انجامد. همچنین تصمیم‌گیری براساس تحلیل‌های ناشی از داده‌کاوی باعث می‌شود پیش‌بینی ریزش دقیق‌تر شود و بینشی ارزشمند به وجود آید.

منابع

A. Ahmed, A. Khan, S. H. Khan, A. Basit, I. U. Haq, and Y. S. Lee, "*Transfer Learning and Meta Classification Based Deep Churn Prediction System for Telecom Industry,"* pp. 1–10, 2022. https://doi.org/10.1016/j.apmrv.30214.02.003

A. Amin., "Customer churn prediction in the telecommunication sector using a rough set approach," Neurocomputing, vol. 237, 2017. https://doi.org/10.1016/j.apmrv.2018.02.056

A. H. A. Kandel, "A comparative study of tree-based models for churn prediction: a case study in the telecommunication sector." 2019. DOI:10.1007/s00170-013-5021-y

B. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, "*Customer churn prediction in telecommunication industry using data certainty*," J. Bus. Res., vol. 94, pp. 290–301, 2019. https://doi.org/10.1016/j.amc.2005.01.081

Baran, R.R. and Strunk, D.P. *Principles of Customer Relationship Management. Australia*: Thomson Southwest., PP: 131-134, 2019. DOI: https://10.1016/S0305-0548(03)00095-9

Baliga, A. J., Chawla, V., Sunder M, V., & Kumar, R. *Barriers to service recovery in B2B markets: a TISM approach in the context of IT-based services*. Journal of Business & Industrial Marketing. 1(11), 202-226, 2021. DOI: https://10.1016/S0305-0548(03)00095-9

Chauhan, S.; Akhtar, A.; Gupta, A. *Customer experience in digital banking: A review and future research directions*. IJQSS 14, 311–348, 2022. https://doi.org/10.1016/j.jmse.2020.10.001

Christy, A.J.; Umamakeswari, A.; Priyatharsini, L.; Neyaa, *A. RFM ranking-An effective approach to customer segmentation*. J. King. Saud. Univ. Sci., 33, 1251–1257, 2021. DOI: https://doi.org/10.1016/j.apm.2012.04.041

De Lima Lemos, R.A.; Silva, T.C.; Tabak, B.M. *Propension to customer churn in a financial institution*: A machine learning approach. Neural Comput. Appl, 1–18, 2022. (DOI): https://doi.org/10.22059/IMJ.2016.61711

D. Spanoudes and T. Nguyen, "Deep Learning in Customer Churn Prediction: Unsupervised Feature Learning on Abstract Company Independent Feature Vectors," pp. 1–22, 2017 .https://doi.org/10.1016/j.fss.2011.03.003

Elena Dumitrescu et al., *"Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects*", European Journal of Operational Research, 2021. https://doi.org/10.1016/S1874-8651(10)75329-4

E. S. Halibas, A. C. Matthew, I. G. Pillai, J. H. Reazol, E. G. Delvo, and L. B. Reazol, "*Determining the intervening effects of exploratory data analysis and feature engineering in telecoms customer churn modelling,*" in 2018 4th MEC International Conference on Big Data and Smart City (ICBDSC), pp. 1–7, 2018. https://doi.org/10.1016/S1874-8651(10)56951-4

G. Zhang, W. Li, T. Mo, and W. Tan, "*Deep and Shallow Model for Insurance Churn Prediction Service*," 2019, doi: 10.1109/SCC.2017.

Hastie, Trevor. *The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations*. Springer. ISBN 0-387-95284-5. OCLC 46809224, 2022. doi.org/10.1016/S1874-8651(10)74596-4

Hosseini, M.; Shajari, S.; Akbarabadi, M. *Identifying multi-channel value co-creator groups in the banking industry*. J. Retail. Consum. Serv, 5, 102312, 2022. https://doi.org/10.1016/S1874-8651(10)60853-4

iahou, X.; Harada, Y. B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. J. Theor. Appl. Electron. Commer. Res, 17, 458–475,2022. https://doi.org/10.1016/S1874-8651(10)60874-4

Jain, H.; Yadav, G.; Manoov, R. Churn prediction and retention in banking, telecom and IT sectors using machine learning techniques. In Advances in Machine Learning and Computational Intelligence; Springer: Singapore, pp. 137–156, 2022. https://doi.org/10.1016/S1874-8651(10)60001-4

Li, M.; Wang, Q.W.; Shen, Y.Z.; Zhu, T.Y. Customer relationship management analysis of outpatients in a Chinese infectious disease hospital using drug-proportion recency-frequency-monetary model. Int. J. Med. Inform, 147, 104373, 2021. https://doi.org/10.1016/S1874-8651(10)60384-4

Li, Y.; Chu, X.Q.; Tian, D.; Feng, J.Y.; Mu, W.S. *Customer segmentation using K-means clustering and the adaptive*. Appl. Soft Comput, 113, 107924, 2021.Silveira, L.J.; Pinheiro, P.R.; Junior, L.S.D.M. *A Novel Model Structured on Predictive Churn Methods in a Banking Organization*. J. Risk Financ. Manag. 14, 481, 2022. https://doi.org/10.1016/S1874-8651(10)60852-4

Matuszela ´nski, K.; Kopczewska, K. *Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach*. J. Theor. Appl. Electron. Commer. Res, 17, 165–198,2022 . https://doi.org/10.1016/S1874-8651(10)906874-4

Muneer, A.; Ali, R.F.; Alghamdi, A.; Taib, S.M.; Almaghthawi, A.; Ghaleb, E.A.A. *Predicting customers churning in banking industry: A machine learning approach*. Indones. J. Electr. Eng. Comput. Sci, 26, 539–549, 2022. https://doi.org/10.1016/S1874-8651(10)60374-4

Piryonesi, S. M.; El-Diraby, T. E. "*Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index".* Journal of Infrastructure Systems. doi:10.1061/(ASCE)IS.1943- 55X.0000512, 2022. https://doi.org/10.1016/S1874-8651(10)90854-4

R. Rao, *"Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems,"* International Journal of Industrial Engineering Computations, vol. 7, no. 1, pp. 19-34, 2016. https://doi.org/10.1016/S1874-8651(10)60874-4

S. Maldonado, J. López, and C. Vairetti, "*Profit-based churn prediction based on Minimax Probability Machines*," Eur. J. Oper. Res., vol. 284, no. 1, pp. 273–284, doi: 10.1016/j.ejor.2020.12.007, 2020 . https://doi.org/10.1016/S1874-8651(10)60932-4

Sunday, K., Ocheja, P., Hussain, S., Oyelere, S., Samson, B., Agbo, F.*: Analyzing student performance in programming education using classification techniques*. Int. J. Emerg. Technol. Learn. (iJET) 15(2), 127–144, 2020. https://doi.org/10.1016/S1874-8651(10)60037-4

T. Calzada-Infante, M. Óskarsdóttir, and B. Baesens, "*Evaluation of customer behavior with temporal centrality metrics for churn prediction of prepaid contracts,*" Expert Syst. Appl., vol. 160, p. 113553, doi: 10.1016/j.eswa.2020.113553, 2020 . https://doi.org/10.1016/S1874-8651(10)60035-4

Tékouabou, S.C.K.; Chabbar, I.; Toulni, H.; Cherif, W.; Silkan, H. Optimizing the early glaucoma detection from visual fields by combining preprocessing techniques and ensemble classifier with selection strategies. Expert Syst. Appl, 189, 115975, 2022. https://doi.org/10.1016/S1874-8651(10)67841-4

Tien-Yu. Hsu, "*Machine learning applied to stock index performance enhancement*", Journal of Banking and Financial Technology, pp. 1-13, 2021.) https://doi.org/10.1016/S1874-8651(10)61324-4

U. Amin et al., "*Just-in-time customer churn prediction in the telecommunication sector*," J. Supercomput., vol. 76, no. 6, pp. 3924–3948, doi: 10.1007/s11227-017-2149-9, 2020 . https://doi.org/10.1016/S1874-8651(10)62134-4

U. M. Kostić, M. I. Simić, and M. V Kostić, "*Social Network Analysis and Churn Prediction in Telecommunications Using Graph Theory*," Entropy, vol. 22, no. 7, p. 753, 2020. https://doi.org/10.1016/S1874-8651(10)60001-4

U. Özmen, E. K. Aydoğan, Y. Delice, and M. D. Toksarı, "*Churn prediction in Turkey's telecommunications sector: A proposed multiobjective–cost-sensitive ant colony optimization*," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 10, no. 1, p. e1338, doi: 10.1002/widm.1338, 2020. https://doi.org/10.1016/S1874-8651(10)60001-4

Valluri, C.; Raju, S.; Patil, V.H. Customer determinants of used auto loan churn: Comparing predictive performance using machine learning techniques. J. Mark. Anal. 2021. https://doi.org/10.1016/S1874-8651(10)60001-4

Veningston, K.; Rao, P.V.; Selvan, C.; Ronalda, M. Investigation on Customer Churn Prediction Using Machine Learning Techniques. In Proceedings of International Conference on Data Science and Applications; Springer: Singapore, pp. 109–119, 2022. https://doi.org/10.1016/S1874-8651(10)60001-4

W. Ahmed, H. Afzal, I. Siddiqi, M. F. Amjad, and K. Khurshid, "*Exploring nested ensemble learners using overproduction and choose approach for churn prediction in telecom industry*," Neural Comput. Appl., vol. 32, no. 8, pp. 3237–3251, doi: https://10.1007/s00521-018-3678-8, 2020 .

W. Jain, A. Khunteta, and S. Srivastava, "*Churn Prediction in Telecommunication using Logistic Regression and Logit Boost,*" Procedia Comput. Sci., vol. 167, pp. 101–112, doi: https://10.1016/j.procs.2020.03.187,2020.

Wu, J.; Shi, L.; Yang, L.P.; Niu, X.X.; Li, Y.Y.; Cui, X.D.; Tsai, S.B.; Zhang, Y.B. *User Value Identification Based on Improved RFM Model and K-Means++ Algorithm for Complex Data Analysis*. Wirel Commun. Mob.Com, 9982484, 1–8, 2021 . https://doi.org/10.1016/S1874-8651(10)60001-4