"Research Article"

10.30495/QJOPM.2021.1939511.3221

Presenting a Hybrid Model based on the Machine Learning for the Classification of Banking and Insurance Industry Common Customers

Hamidreza Amirhasankhan[1], *Abbas Toloie Eshlaghy*[*2], *Reza Radfar*[3], *Alireza Pourebrahimi*[4]

(Received:2022.09.03; Accepted:2023.02.08)

Abstract

Global competition, dynamic markets, and rapidly shrinking innovation and technology cycles, all have imposed significant challenges on the financial, banking, and insurance industries and the need to data analysis for improving decision-making processes in these organizations has become increasingly important. In this regard, the data stored in the databases of these organizations are considered as valuable sources of information and knowledge needed for organizational decisions. In the present research, the researchers focus on the common customers of the bank and insurance industry. The purpose is to provide a methodology to predict the performance of new customers based on the behavior of previous customers. To this end, a hybrid model based on support vector machine and genetic algorithm is used. The support vector machine is responsible for modeling the relationship between customer performance and their identity information and the genetic algorithm is responsible for tuning and optimizing the parameters of the support vector machine. The results obtained from customer classification using the proposed model in this research led to customer classification with a high accuracy of 99%.

Key Words: support vector machine, genetic algorithm, classification, banking, insurance.

1.Ph.D. Candidate of Information Technology Management Group, UAE Branch, Islamic Azad University, Dubai, UAE
2.Professor, Department of Industrial Management, Science and Research Unit, Islamic Azad University, Tehran, Iran
*.Corresponding Author:toloie@srbiau.ac.ir
3.Professor, Department of Industrial Management, Science and Research Unit, Islamic Azad University, Tehran, Iran
4.Assistant Professor, Department of Management, Karaj Branch, Islamic Azad University, Karaj, Iran

1. Introduction

In this research, the researchers aim to present an efficient model based on support vector machine and genetic algorithm for classifying and predicting the performance of new common customers of banking and insurance industry. The purpose of this research is to enable investment holdings that are common

shareholders of banks and insurance companies to achieve the highest level of customization in decision making for customers and adopt diverse and efficient decisions in accordance with their customers' characteristics and strengthen interactions with customers, better meet customer needs and improve customer satisfaction and loyalty. Accordingly, these holdings can achieve significant results in each of the above-mentioned areas by strengthening databases, communication links of information companies and increasing accuracy in entering and registering initial information and relying on machine learning methods.

2. Literature Review

Among the studies that have been conducted in recent years in the field of banking industry customer classification, the study of Jamshidi et.al. (2019) is included. They presented a multi-objective approach based on adaptive neuro-diffusion inference system for detecting bank money laundering and currency exchange. Magomedov et al. (2018), Dorofeev et al. (2018) and Plaksiy et al. (2018) have used machine learning methods based on artificial intelligence to design and monitor anti-money laundering systems. Leite et al. (2019) and Tiwari et al. (2020) have compiled a rich collection of researches based on machine learning and artificial intelligence to deal with money laundering and other banking crimes in their review papers.

3. Methodology

In this study, the researchers aim to model the classification of common customers of banking and insurance industry using a hybrid method based on support vector machine network and optimization using genetic algorithm. For this purpose, first the independent and dependent variables are determined. In this regard, the identity information of customers is defined as the independent variables and the class that each customer is placed in as the dependent variable. In the next step, the customer set is divided into two groups of training and testing data. The data is randomly divided into two groups of training and testing, such

that 90 percent of the data is used in the training phase and the rest in the testing phase.

4. Result

The criteria of accuracy, recall and precision are used to evaluate the methods of predicting the class of common insurance and bank customers in this research. The most important criterion for determining the efficiency of classification techniques is the Accuracy criterion. This measure calculates the overall accuracy of a classifier. It indicates the fact that the designed classifier has correctly classified what percentage of the entire set of test records. The results obtained in this research show that the support vector machine set by the genetic algorithm for customer classification has correctly recognized 99.98% of the test data. Considering the desired amount of the three criteria of accuracy, recall and precision of this combined method, it is found that this method is able to efficiently classify common bank and insurance customers.

5. Discussion

In this research, the researchers implemented a support vector machine for classifying common customers of banking and insurance and examined the obtained results. After going through the training process and obtaining the optimal parameters of the support vector machines using the genetic algorithm, the performance of this method was evaluated in the testing phase with 6060 customers whose information was not given to the support vector machines in the training phase. The comparison of the output of the support vector machine network with the actual class of customers indicates the appropriate fit of the outputs obtained from the support vector machine network with the real data. Based on the results obtained, the classification error of the proposed model is 0.0003. These results mean that the accuracy of the performance of the support vector machine is about 99.97 percent, which can be considered as an acceptable accuracy. Nowadays, in most organizations, data is rapidly being collected and stored. However, it can be argued that despite the existence of a large volume of data, organizations generally face a lack of knowledge in decision-making. Although using various conventional reporting tools, information can be provided to users so that they can draw conclusions about the data and the logical relationships between them, when a huge volume of data is involved, even experienced and professional users cannot detect useful patterns in the abundance of data. Nowadays, machine learning techniques have been considered to

meet the needs of various organizations and companies in discovering knowledge from a large volume of data. Data mining is the process of extracting information and knowledge and discovering hidden patterns from a very large database. Telecommunication companies, banks, insurance companies, advertising companies and all companies that have large databases can use data mining to improve their decision-making processes. Data mining causes organizations to reach higher levels of knowledge and unknown patterns from the data level. The extracted patterns can be a relationship between the features and characteristics of the system such as the type of demand and the type of customer, future predictions based on the system characteristics, rules (if-then) between the system variables, classifications and clustering of objects and records similar to each other in a system, and the like.

((مقاله پژوهشی))

10.30495/QJOPM.2021.1939511.3221

ارائۀ یک مدل ترکیبی مبتنی بر یادگیری ماشینی برای طبقه‌بندی مشتریان مشترک صنعت بانکداری و بیمه

حمیدرضا امیرحسنخانی[۱]، عباس طلوعی اشلقی*[۲]،رضا رادفر[۳]، علیرضا پورابراهیمی[۴]

(دریافت:۱۴۰۱/۰۶/۱۲- پذیرش نهایی:۱۴۰۱/۱۱/۱۹)

چکیده

رقابت‌های جهانی، صنایع پویا و چرخه‌های نوآوری و فناوری که به سرعت در حال کوتاه شدن هستند همگی چالش‌های مهمی را برای صنعت مالی، بانکداری و بیمه ایجاد کرده‌اند و نیاز به تجزیه و تحلیل داده‌ها جهت بهبود فرآیندهای تصمیم‌گیری در این سازمان‌ها بیش از پیش اهمیت پیدا کرده است؛ در این میان، داده هایی که در پایگاه‌های اطلاعاتی این سازمان‌ها نگهداری می‌شوند به عنوان منابع ارزشمند اطلاعات و دانش مورد نیاز جهت تصمیم‌گیری‌های سازمانی مطرح می‌باشند؛ در این تحقیق بر روی مشتریان مشترک صنعت بانکداری و بیمه تمرکز شده است. هدف از این تحقیق، ارائۀ روشی جهت پیش‌بینی عملکرد مشتریان جدیدالورود بر مبنای رفتار مشتریان پیشین است؛ برای این منظور، از یک مدل ترکیبی مبتنی بر ماشین بردار پشتیبان و الگوریتم ژنتیک استفاده شده است؛ بدین ترتیب که ماشین بردار پشتیبان، وظیفه مدل‌سازی رابطه بین عملکرد مشتریان و اطلاعات هویتی آنها را بر عهده دارد و الگوریتم ژنتیک، وظیفه تنظیم و بهینه‌سازی پارامترهای ماشین بردار پشتیبان را عهده‌دار است. نتایج به دست آمده از طبقه بندی مشتریان با استفاده از مدل پیشنهادی در این تحقیق طبقه بندی مشتریان با دقت بالای ۹۹ درصد است.

واژه‌های کلیدی: ماشین بردار پشتیبان، الگوریتم ژنتیک، طبقه‌بندی، بانک، بیمه

۱. دانشجوی دکتری گروه مدیریت فناوری اطلاعات، واحد امارات، دانشگاه آزاد اسلامی، دبی، امارات متحده عربی amirhasankhani@ut.ac.ir
۲ .استاد گروه مدیریت صنعتی،واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران (نویسنده مسؤول) toloie@srbiau.ac.ir
۳. استادگروه مدیریت صنعتی، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران r.radfar@srbiau.ac.ir
۴ .استادیار گروه مدیریت،واحد کرج، دانشگاه آزاد اسلامی، کرج، ایران

مقدمه

به طور کلی، یکی از مهمترین مقولات صنعت بیمه و بانکداری نوین امروز ارزیابی مشتریان بر مبنای ریسک اعتباری آنهاست؛ رایج‌ترین رویکردی که در این چارچوب امروزه در ابعاد گسترده مورد استفاده قرار می‌گیرد، در نظر گرفتن یک حد آستانه برای طبقه بندی مشتریان بر مبنای ریسک اعتباری آنهاست؛ به طوری که ابتدا بر مبنای تراکنش‌ها و عملکرد آن، به هر مشتری یک امتیاز که نشان دهندهٔ ریسک اعتباری آن است اختصاص داده می‌شود سپس با مقایسهٔ امتیاز به دست آمده توسط هر مشتری با حدود آستانه طبقه بندی کننده می‌توان مشتریان را به لحاظ ریسک اعتباری در طبقات مختلف از پیش تعریف شده دسته بندی نمود. در این رویکرد مشتریان عمدتاً در دو دسته مشتریان پر ریسک و کم ریسک قرار می گیرند. همین طبقه‌بندی می‌تواند به عنوان مبنایی برای آموزش الگوریتم‌های یادگیرنده در فرایندهای یادگیری ماشین مورد استفاده قرار گیرد؛ به طوری که با برچسب‌گذاری مشتریان فعلی بر اساس این چارچوب می‌توان الگوریتم‌های یادگیرنده موجود در یادگیری ماشینی را به منظور پیش‌بینی میزان ریسک اعتباری مشتریان جدید الورود مورد استفاده قرار داد.

در این تحقیق قصد داریم یک مدل کارا مبتنی بر ماشین بردار پشتیبان و الگوریتم ژنتیک جهت طبقه بندی و پیش‌بینی عملکرد مشتریان جدیدالورود برای مشتریان مشترک صنعت بانکداری و بیمه ارائه دهیم. هدف این تحقیق، توانمندسازی هلدینگ‌های سرمایه‌گذاری که سهامداران توام بانک و بیمه هستند در بالاترین سطح سفارشی‌سازی در تصمیم گیری برای مشتریان و اتخاذ تصمیمات متنوع و کارا طبق خصوصیات مشتریان‌شان و قوی‌تر کردن تعاملات با مشتری، رفع بهتر نیازهای مشتری و ارتقای سطح رضایت مشتریان و در نتیجه وفادار کردن آنهاست. بر این اساس، این هلدینگ‌ها می‌توانند از طریق تقویت پایگاه داده ها، لینک های ارتباطی شرکت های اطلاعاتی و افزایش دقت در ورود و ثبت اطلاعات اولیه و با اتکا به روش های یادگیری ماشینی، در هریک از زمینه های مذکور به نتایج قابل توجه دست یابد.

در ادامه این بخش به مرور ادبیات موضوع می پردازیم؛ طبقه بندی پژوهش‌های انجام شده از مهم‌ترین کارها در زمینه تحلیل رویگردانی مشتریان بانک است؛ در این حوزه تکنیکهایی نظیر مدل توسعه یافته متوازن از جنگل‌های تصادفی (IBRF) (سای و همکاران[1]، ۲۰۰۹)، تئوری مجموعه‌های ناهموار[2] (لین و همکاران[1]، ۲۰۱۱) ، رگرسیون لاجیت و درخت تصمیم (نای و همکاران[2]، ۲۰۱۱) و درخت تصمیم C5.0 (چو و همکاران[3]، ۲۰۰۷) مورد استفاده قرار گرفته است.

1.Xie et al
2.Rough Set Theory (RST)

یکی دیگر از حوزه های پرکاربرد استفاده از یادگیری ماشین در مؤسسات مالی نظیر بانک و بیمه، مقوله مقابله با انواع تقلب‌های مالی است؛ نمونه‌ای از تکنیک های مورد استفاده در کشف تقلب و سوء استفاده‌های مالی، شامل ترکیب الگوریتم ژنتیک و شبکه‌های عصبی (پاسچ[۴]، ۲۰۰۸) ، شبکه های عصبی خود سازمانده[۵] (کوا و اسریگانش[۶]، ۲۰۰۸) ، قوانین وابستگی به‌دست آمده از منطق فازی (سانچز و همکاران[۷]، ۲۰۰۹) است. در تحقیق لی و همکاران[۸] (۲۰۱۰) برای رسیدن به این هدف از داده‌هایی نظیر مبلغ پیشنهاد اولیه، قیمت بسته شده مزایده، طول مزایده، سود مزایده و میزان اعتبار فروشنده استفاده شده است. در این تحقیق از رگرسیون لاجیت جهت طبقه‌بندی بهره برده شد؛ عملکرد رگرسیون لاجیت، ماشین‌بردار پشتیبان و جنگل‌های تصادفی در مطالعه بهاتاچاریا و همکاران[۹] (۲۰۱۱) براساس داده‌های مورد استفاده در مطالعه لی و همکاران[۱۰](۲۰۱۰) مورد مقایسه قرار گرفته است. در مقاله دومان و اوزکلیک[۱۱](۲۰۱۱) از الگوریتم جستجوی پراکنده[۱۲] بهره برده شده است.

همچنین در ادامه برخی از مطالعات صورت گرفته در مورد کاربرد یادگیری ماشین را جهت پیش‌بینی عملکرد و مقابله با ورشکستگی اقتصادی مورد تجزیه و تحلیل قرار می‌دهیم؛ چرا که اساساً ارزیابی عملکرد مؤسسات مالی نظیر بانک و بیمه به منظور تبیین سیاست‌های کلی و اتخاذ تصمیمات بلندمدت توسط مدیران از اهمیت بسیار زیادی برخوردار است. نمونه‌ای از تکنیکهای مورد استفاده در پژوهش‌های صورت گرفته در ارزیابی عملکرد بانک شامل ماشین‌بردار پشتیبان و درخت تصمیم در تحقیق لین و همکاران[۱۳] (۲۰۰۹)، شبکه‌های عصبی، ماشین‌بردار پشتیبان، تحلیل خوشه‌ای K-means و رگرسیون لاجیت در مطالعه بویاسیوگلو و همکاران[۱۴] (۲۰۰۹) ، رگرسیون

1.Lin et al
2.Nie et al
3. Chu et al
4. Paasch
5. Self – organization maps (SOM)
6. Quah and Sriganesh
7. Sánchez et al
8. Lee et al
9. Bhattacharyya et al
10. Lee et al
11. Duman and Ozcelik
12. Scatter Search (SS)
13. Lin et al
14. Boyacioglu et al

لاجیت، درخت تصمیم و شبکه‌های عصبی در مطالعه ژائو و شینها[1] (۲۰۰۹) و شبکه‌های پس انتشار متوازن در مقاله کیرکاس و همکاران[2] (۲۰۰۷) است.

یکی از موضوعاتی که در سال‌های اخیر توسط پژوهشگران بسیار مورد توجه بوده است مسئلهٔ بازپرداخت وام های اعطا شده توسط بانکها به مشتریان است؛ در این راستا تخمین احتمال عدم بازپرداخت وام توسط وام گیرنده بر مبنای داده های مستخرج از عملکرد گذشته او برای مدیران بانک‌ها از اهمیت بسیار زیادی برخوردار است؛ به طوری که الگوریتم‌های یادگیری ماشین بر مبنای تجزیه و تحلیل داده‌های تاریخی حاصل از عملکرد مشتریان می‌توانند آنها را در دسته‌های مختلفی –که نشان دهندهٔ ریسک اعتباری آنهاست طبقه‌بندی نماید و از این مدلها برای پیش‌بینی احتمال عدم بازگرداندن وام توسط وام‌گیرنده‌های جدید بهره برداری کند.

تحقیقات زیادی در زمینه استفاده از یادگیری ماشینی در رتبه‌بندی اعتبار مشتریان بانک‌ها اانجام شده است که تکنیک‌های مختلف طبقه‌بندی را مورد بهره برداری قرار داده‌اند؛ از جملهٔ این تکنیک‌ها، شبکه‌های عصبی با پایه محرک شعاعی[3] (هوانگ و همکاران، ۲۰۰۶)[4]، درخت دسته‌بندی و رگرسیون CARD (لی و همکاران، ۲۰۰۶)[5]، ماشین‌بردار پشتیبان (هوانگ و همکاران، ۲۰۰۷)[6]، استفادهٔ توأمان شبکه‌های عصبی احتمالی و چندلایه پیشخور، تحلیل براساس حداقل انحراف از میزان متوسط و رگرسیون لاجیت (ابدو و همکاران[7]، ۲۰۰۸) ، شبکه‌های عصبی پس انتشار (ساسترسیک و همکاران[8]، ۲۰۰۹) ، ماشین بردار پشتیبان و روش جدید CLC (لو و همکاران[9]، ۲۰۰۹) ، ترکیب روش ماشین‌بردار پشتیبان با روش‌های انتخاب ویژگی (چن و لی[10]، ۲۰۱۰) ، گره اعتبارسنجی در نرم‌افزار SAS و درخت تصمیم (یاپ و همکاران[11]، ۲۰۱۱) است.

از جملهٔ مطالعاتی که در سالهای اخیر در حوزهٔ طبقه بندی مشتریان صنعت بانکداری صورت گرفته است: مقاله جمشیدی و همکاران[12] (۲۰۱۹) است؛ آنها یک رویکرد چند هدفه مبتنی بر سیستم

1. Zhao and Sinha
2. Kirkos et al
3.Radial Basis Function (RBF Network)
4.Huang et al
5.Lee et al
6.Huang et al
7.Abdou et al
8.Šušteršič et al
9.Luo et al
10.Chen and Li
11.Yap et al
12.Jamshidi et al

استنتاج نورو دیفیوژن سازگار برای شناسایی پول‌شویی بانکی و مبادله ارز ارائه دادند. مگومدو[1]و و همکاران، (۲۰۱۸)، دورفیو و همکاران[2] (۲۰۱۸) و پلاکسی و همکاران[3] (۲۰۱۸) از روشهای یادگیری ماشین مبتنی بر هوش مصنوعی جهت طراحی و نظارت بر سیستمهای ضدپول‌شویی بهره برده‌اند. لیت و همکاران[4] (۲۰۱۹)، و تیاواری و همکاران[5] (۲۰۲۰) در مقالات مروری، مجموعه‌ای غنی از تحقیقات صورت گرفته مبتنی بر یادگیری ماشینی و هوش مصنوعی جهت مقابله با پول‌شویی و سایر جرائم بانکی را گردآوری کرده‌اند.

علی‌رغم حجم انبوه تحقیقات و پژوهش‌هایی که در کشورهای پیشرفته در زمینه توسعه و به‌کارگیری تکنیک‌های داده‌کاوی در صنعت بانکداری و بیمه به عمل آمده و نتیجه آن بهبود فرایندهای بانکی و بیمه بوده است، زمینه‌های بالقوه بسیاری در به‌کارگیری این دانش در بانک‌ها و بیمه ها درکشورمان وجود دارد؛ بنابراین زمینه‌سازی جهت آشنایی کارشناسان و متخصصان امور بانکی و بیمه با تکنیک‌های داده‌کاوی و کاربردهای آن و همچنین برگزاری دوره‌های آموزشی در این زمینه و به‌کارگیری عملی این علم در بانک‌ها و بیمه کشور از اهمیت بسزایی برخوردار است.

برای شناخت بهتر و آگاهی بیش‌تر از مشتریان باید به مفهوم بخش‌بندی مشتریان پرداخت که طی آن تلاش می‌شود گروه‌های مشتری با نیازها و الگوهای رفتاری مشابه مشخص شوند؛ از جمله الزامات یک بخش‌بندی موفق، انتخاب متغیر مناسب است. از سوی دیگر با ظهور تکنولوژی های جدید و امکان رقابت در سطح جهانی، بسیاری از سازمانها به منظور خدمت‌دهی بهتر و نزدیک‌تر با مشتریان، به مدیریت ارتباط با مشتری روی آورده‌اند. برنامه‌ریزی برای ارتباط با مشتری، بدون تسهیل و ایجاد رابطهای مربوط به بخش‌بندی مشتریان امکان پذیر نبوده و یکی از پیش‌نیازهای شناخت و کشف رفتار آتی مشتریان می باشد.

مشتریان در فرآیند بخش بندی به گونه‌ای تقسیم می‌شوند که افراد شبیه به یکدیگر در یک گروه قرار گرفته و گروه های مختلف کمترین شباهت را به یکدیگر داشته باشند. سپس با توجه به خصوصیات هر گروه، برنامه های خاصی جهت تولید محصولات جدید، تبلیغات و بازاریابی در نظر گرفته می شود. یکی از ابزارهایی که امروزه جهت بخش بندی مورد توجه قرار گرفته، ابزارهای داده-کاوی و خوشه بندی می باشد؛ داده‌کاوی، فرآیند اکتشاف و تحلیل الگوهای معنی‌دار و قواعد، در بین مقادیر زیاد داده‌ها بوسیله ابزارهای خودکار و نیمه‌خودکار می باشد.

1.Magomedov et al
2.Dorofeev et al
3.Plaksiy et al
4.Leite et al
5.Tiwari et al

هدف از این تحقیق، ارائۀ یک روش طبقه‌بندی مشتریان مشترک صنعت بانکداری و بیمه و نیز ایجاد مدلی جهت پیش‌بینی اعتبار مشتریان جدیدالورود می‌باشد. این تحقیق گروه متجانسی از مشتریان را مورد تحلیل قرار داده تا به نمایندگی از کل مشتریان در ساخت مدل به کار روند.

معرفی مورد مطالعاتی:

بانک ایران زمین (سهامی عام) به موجب مجوز شماره ۲۸۳۵۹۲/۸۹ مورخ ۱۳۸۹/۱۲/۲۱ صادره از بانک مرکزی جمهوری اسلامی ایران در تاریخ ۱۳۸۹/۱۲/۲۴ تحت شماره ۳۹۹۲۷۹ در ادارۀ ثبت شرکت ها و مؤسسات غیر تجاری در تهران به ثبت رسیده است؛ سهام بانک در تاریخ ۱۳۸۹/۱۲/۲۵ در فرابورس ایران پذیرفته شده است و از تاریخ ۱۳۹۰/۷/۵ درفهرست تابلوی قیمت های بورس قرار گرفته است. بانک ایران زمین، پدیده‌ای از کارآفرینی در بخش های مختلف اقتصادی است. این بانک در پی آن است تا با ارائه خدمات با کیفیت و تأکید بر مشتری مداری، بانکی پیشرو در اقتصاد آینده ایران باشد. نیز این بانک بر آن است که با رعایت کامل قوانین و مقررات پولی و مالی کشور، تجربیات برتر بین المللی را به کار بندد و خدمات خود را در بالاترین استانداردهای کیفی ارائه نماید و با استفاده از نوآوری‌ها و فن‌آوری‌های روز دنیا به لحاظ عملیات و کارآیی، پیشرو بانکهای خصوصی کشور باشد و راهکارهای بانکی و مالی کامل را به همه مشتریان ایران زمین ارائه کند. ویژگی متمایز بانک ایران زمین رویکرد دانش محوری است که فراتر از قلمرو سنتی بانکداری است و کمک می‌کند تا بانک، ضمن حفظ حقوق صاحبان سهام و سپرده گذاران، نسبت به تحولات محیطی، به بهترین نحو و هوشمندانه عمل کند و بهترین ساختاربندی را درباره محصولات و خدمات خود ارائه دهد.

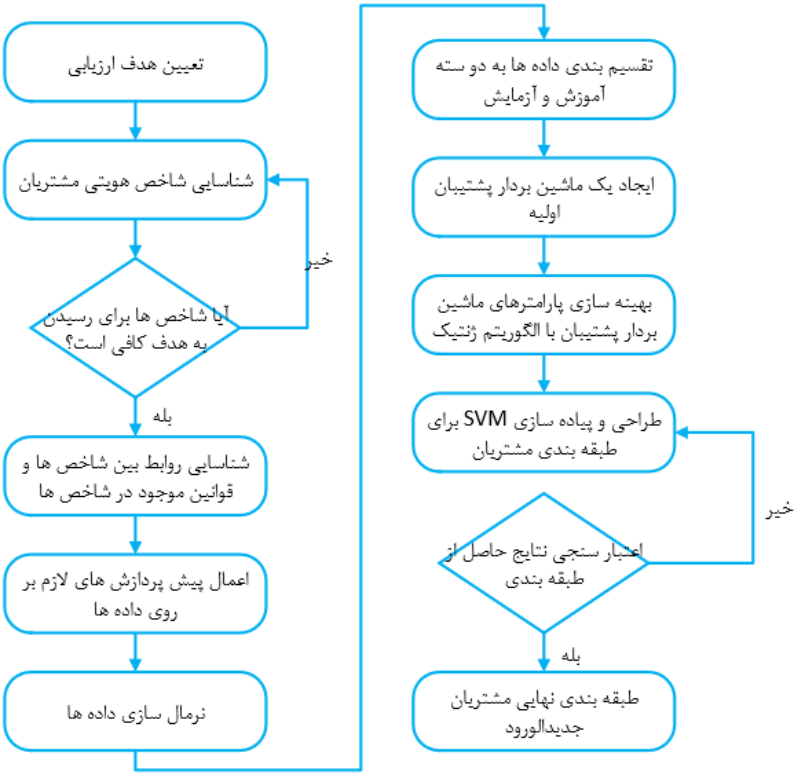
شرکت بیمه آرمان (سهامی عام) به موجب قانون تأسیس بیمه‌گری و با مجوز شماره ۴۱۹۸۰۱ بیمه مرکزی جمهوری اسلامی ایـران در تاریخ ۱۳۹۰/۱۰/۷ در ادارۀ ثبت شرکت ها و مؤسسات غیر تجاری، به ثبت رسید و پروانه فعالیت بیمه‌گری در انواع رشته های بیمه‌ای را تحت شماره ۴۳۵۴۱/۹۰ مورخ ۱۳۹۰/۱۱/۱۸ از بیمه مرکزی جمهوری اسلامی ایران دریافت کرد.

چشم انداز بیمۀ آرمان، فعالیت در یک صنعت خدماتی مبتنی بر رقابت آزاد، مستلزم تدوین و تألیف نقشۀ راه و ترسیم افق فعالیت های سازمان است؛ افزایش چشمگیر تعداد شرکت های بیمۀ فعال در سطح کشور و نیز حضور شرکت‌های با سابقه و توانمند در این صنعت، نشان از اهمیت و لزوم برنامه‌ریزی، سازماندهی، مدیریت و اتخاذ تدابیر مناسب برای شرکت‌های نوپاست.

هدف از این تحقیق، ارائۀ یک روش طبقه‌بندی مشتریان مشترک صنعت بانکداری و بیمه و نیز ایجاد مدلی جهت پیش‌بینی اعتبار مشتریان جدیدالورود می‌باشد؛ این تحقیق، گروه متجانسی از مشتریان را مورد تحلیل قرار داده تا به نمایندگی از کل مشتریان در ساخت مدل به کار روند.

مدل تحقیق

ساختار کلی مدل ارائه شده به صورت شکل ۱ نشان داده می‌شود.



شکل ۱- ساختار کلی مدل ارائه شده در این تحقیق

Figure 1: The general structure of the model presented in this research

ابزار و روش:

در این بخش، ابتدا رفتار و فرمولاسیون ریاضی حاکم بر ماشین‌های بردار پشتیبان تشریح شده سپس به توضیح الگوریتم ژنتیک جهت بهینه‌سازی پارامترهای موجود در شبکه ماشین‌های بردار پشتیبان پرداخته شده و در نهایت، طراحی و پیاده‌سازی شبکهٔ ماشین‌های بردار پشتیبان جهت الگویابی و طبقه‌بندی مشتریان شرح داده شده است.

ماشین بردار پشتیبان

مسئلهٔ دسته‌بندی یکی از مسائل اصلی مطرح شده در یادگیری ماشین است؛ به گونه‌ای که بسیاری از مسائل را می‌توان به صورت یک مسئلهٔ دسته‌بندی مطرح کرده و حل کرد. از طرفی در

یادگیری ماشین نیز روش‌های مختلفی برای حل مسئلهٔ دسته‌بندی صورت گرفته است؛ یکی از روش هایی که در حال حاضر به صورت گسترده برای مسئلهٔ دسته‌بندی مورد استفاده قرار می گیرد، روش ماشین بردار پشتیبان است. شاید به گونه‌ای بتوان محبوبیت کنونی روش ماشین بردار پشتیبان را با محبوبیت شبکه های عصبی در دههٔ گذشته مقایسه کرد؛ علت این قضیه نیز قابلیت استفاده این روش در حل مسائل گوناگون می‌باشد، در حالی‌که روشهایی مانند درخت تصمیم‌گیری را نمی‌توان به راحتی در مسائل مختلف به کار برد. در حوزه مبانی مرتبط با یادگیری ماشین، ماشینهای بردار پشتیبان یک مدل با نظارت مبتنی بر الگوریتم‌های یادگیرنده است که داده‌ها را برای طبقه‌بندی و همچنین تحلیل رگرسیون تجزیه و تحلیل می‌کند. ماشینهای بردار پشتیبان یکی از قوی‌ترین روش-های پیش‌بینی‌کننده هستند که بر اساس چارچوب‌های آماری و برنامه‌ریزی ریاضی ارائه شده‌اند.

این الگوریتم بر اساس یک فرایند آموزشی مشخص، داده‌هایی را که هر کدام به یک دسته معین تعلق دارند مدل‌سازی و طبقه‌بندی می‌کند؛ به طور کلی ماشین بردار پشتیبان، مدلی را ایجاد می‌کند که نمونه‌های جدیدی را به یک دستهٔ مشخص اختصاص دهد و آن را به یک دسته‌بندی خطی دودویی غیر احتمالی تبدیل می‌نماید. نسخه‌های غیر خطی این دسته‌بندی در مطالعات بعدی توسعه داده شده است. مکانیزم عملکرد یک ماشین بردار پشتیبان به این صورت است که نمونه‌های آموزشی را به نقاطی در فضا نگاشت می‌دهد تا فاصله بین دو دسته را به حداکثر برساند. سپس نمونه‌های جدید در همان فضا طبقه‌بندی می‌شوند. علاوه بر انجام طبقه‌بندی خطی، ماشین‌های بردار پشتیبان می‌توانند با استفاده از مفهومی که حقه کرنل نامیده می شود، یک طبقه‌بندی غیر خطی را به طور مؤثر انجام دهند و به طور ضمنی ورودی های خود را به فضاهایی با تعداد ابعاد زیاد و پیچیدگی بالا نگاشت دهند.

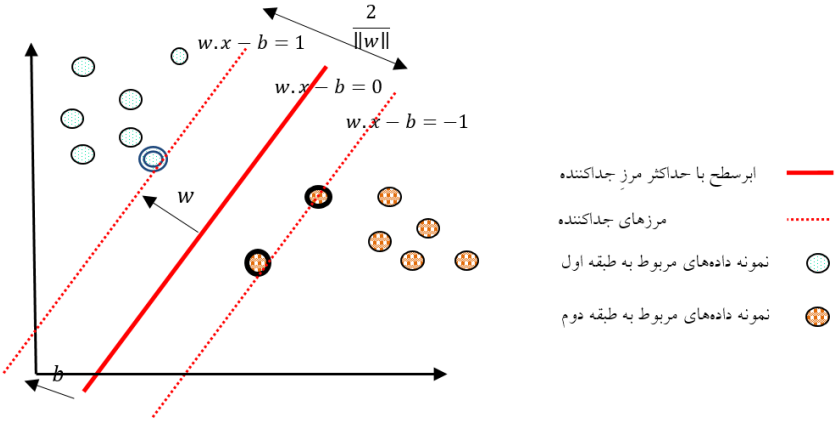
وقتی داده‌ها فاقد برچسب هستند، یادگیری با نظارت امکان‌پذیر نیست و یک روش یادگیری بدون نظارت لازم است که سعی می‌کند داده‌ها را خوشه بندی کند و سپس داده‌های جدید را به این خوشه‌های تشکیل شده اختصاص دهد. الگوریتم خوشه‌بندی بردار پشتیبان، از آماره‌های بردارهای پشتیبان توسعه یافته در الگوریتم ماشین‌های بردار پشتیبان، برای خوشه‌بندی داده‌های بدون برچسب استفاده می‌کند و یکی از پرکاربردترین الگوریتم‌های خوشه‌بندی در کاربردهای مختلف است.

اصولا طبقه‌بندی داده‌ها یک کار معمول در یادگیری ماشین است؛ در این چارچوب اینگونه فرض می‌شود که برخی از نقاط داده شده که هر یک به یکی از دو کلاس تعلق دارند در اختیار است. هدف این است که تصمیم بگیریم یک دادهٔ جدید در کدام کلاس قرار می‌گیرد. ابرصفحه‌های زیادی وجود دارند که ممکن است داده‌ها را طبقه‌بندی کنند. یک انتخاب منطقی به عنوان بهترین ابرصفحه، آن است که بزرگترین فاصله یا حاشیه را بین دو کلاس ایجاد کند. بنابراین ابرصفحه‌ای

مطلوب طوری انتخاب می‌شود که فاصله آن تا نزدیکترین داده در هر طرف حداکثر شود. اگر چنین ابرصفحه‌ای وجود داشته باشد، به عنوان ابرصفحه با حداکثر حاشیه شناخته می‌شود و طبقه‌بندی کننده خطی که تعریف می‌کند به عنوان طبقه‌بندی کننده با حداکثر حاشیه شناخته می شود. معادل چنین ابرصفحه‌ای، یک پرسپترون با پایداری مطلوب در شبکه‌های عصبی است.

به طور کلی، یک ماشین بردار پشتیبان یک ابرصفحه یا مجموعه‌ای از ابرصفحه‌ها را در فضایی با ابعاد زیاد یا بی نهایت می سازد، که می‌تواند برای طبقه‌بندی، رگرسیون یا سایر امور مانند outliers detection استفاده شود. به طور شهودی، یک تفکیک خوب توسط ابرصفحه‌هایی انجام می شود که بیشترین فاصله را تا نزدیکترین داده آموزش داده شده در هر کلاس (به اصطلاح حاشیه عملکردی) دارد؛ زیرا به طور کلی هرچه حاشیه بیشتر باشد، خطای تعمیم طبقه‌بندی کننده کمتر است.

در شکل ۲ تصویری از یک مجموعه دادهٔ متعلق به دو طبقه نشان داده شده که ماشین بردار پشتیبان بهترین ابرسطح را برای جداسازی آن‌ها انتخاب می‌کند. در این شکل داده‌ها دو بعدی هستند یعنی هر داده تنها از دو متغیر تشکیل شده است.



شکل ۲- ابرسطح با حداکثر مرزِ جداکننده به همراه مرزهای جداکننده برای طبقه‌بندی
Figure 2: Hyperplane with maximum separator boundary by considering separator boundaries for classification

حل معادله یافتن خط بهینه برای داده‌ها بوسیله روش‌های برنامه ریزی درجه دو که روش‌های شناخته شده‌ای در حل مسائل محدودیت‌دار هستند صورت می‌گیرد. قبل از تقسیمِ خطی برای اینکه ماشین بتواند داده‌های با پیچیدگی بالا را دسته‌بندی کند داده‌ها به فضایی با ابعاد خیلی بالاتر منتقل

می‌شوند. برای اینکه بتوان مسئلهٔ ابعاد خیلی بالا را با استفاده از این روش‌ها حل کرد از قضیه دوگان لاگرانژ برای تبدیلِ مسئلهٔ مینیمم‌سازی مورد نظر به فرم دوگان آن که در آن به جای تابعی پیچیده که داده‌ها را به فضایی با ابعاد بالا می‌برد تابعِ ساده‌تری به نامِ تابع کرنل استفاده می‌کنیم. از توابع کرنل مختلفی از جمله کرنل‌های نمایی، چندجمله‌ای و سیگموید برای این امر می‌توان استفاده نمود؛ در این تحقیق از یک کرنل گوسی بهره برده شده است.

الگوریتم ژنتیک

در این تحقیق از ماشین‌های بردار پشتیبان برای طبقه‌بندی مشتریان استفاده می‌شود؛ با توجه به حساسیت ماشین‌های بردار پشتیبان به پارامترهای ورودی که بر اساس آن‌ها خطوط تفکیک کننده را ایجاد می‌کنند، تنظیم و مقداردهی دقیق این پارامترها تا حد بسیار زیادی می‌تواند در عملکرد مثبت این روش‌ها مؤثر باشد. در پژوهش‌های مختلف، مقداردهی دقیق پارامترهای روش‌های طبقه‌بندی کننده، همواره تحت عنوان یک مسئلهٔ بهینه‌سازی مستقل مورد تجزیه‌وتحلیل قرار می‌گیرد. به این صورت که هدف این مسائل کمینه‌سازی خطای طبقه‌بندی به ازای مقادیر مختلف پارامترهای ورودی الگوریتم‌های مورد استفاده می‌باشد. در این تحقیق به منظور بهینه‌سازی پارامترهای شبکه ماشین‌های بردار پشتیبان توسعه داده شده جهت طبقه‌بندی مشتریان از الگوریتم ژنتیک استفاده می‌شود؛ در ادامه کلیات الگوریتم ژنتیک ارائه می‌شود.

اصول کاری الگوریتم ژنتیک، در ساختار الگوریتمی شکل ۳ نمایش داده شده است؛ مهم‌ترین گام لازم برای پیاده‌سازی الگوریتم ژنتیک و انواع مختلف آن عبارتند از: تولید جمعیت (اولیه) از جواب‌های یک مسأله، مشخص کردن تابع هدف، تابع برازندگی[1] و به کار گرفتن عملگرهای ژنتیک جهت ایجاد تغییرات در جمعیت جواب‌های مسأله. عملگرهای ژنتیک قابل تعریف در الگوریتم ژنتیک، در ادامه معرفی خواهند شد؛ اصول کاری الگوریتم ژنتیک عبارتند از:

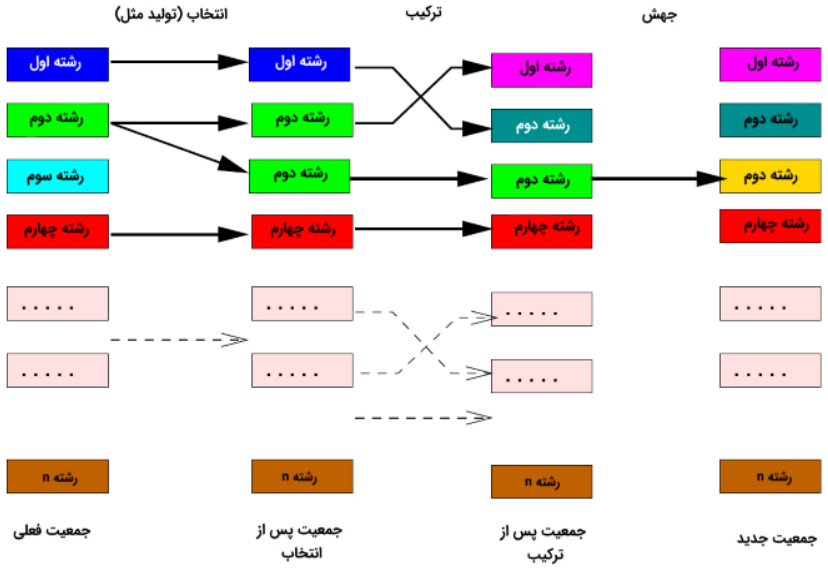
- ✓ فرموله کردن جمعیت ابتدایی متشکل از جواب‌های مسأله
- ✓ مقداردهی اولیه و تصادفی جمعیت ابتدایی متشکل از جواب‌های مسأله

حلقه تکرار:

- ✓ ارزیابی تابع هدف مسأله
- ✓ پیدا کردن تابع برازندگی مناسب
- ✓ انجام عملیات روی جمعیت متشکل از جواب‌های مسأله با استفاده از عملگرهای ژنتیک

1.Fitness

- ✓ عملگر تولید مثل
- ✓ عملگر ترکیب
- ✓ عملگر جهش
- ✓ تا زمانی که : شرط توقف انجام شود.



شکل ۳- نمای کلی از فرایند تکاملی در الگوریتم ژنتیک پس از تولید مثل، ترکیب و جهش
Figure 3: Overview of the evolutionary process in the genetic algorithm after reproduction, crossover and mutation

ما دراین تحقیق برای سیاست انتخاب از مکانیزم چرخهٔ رولت بهره برده ایم؛ برای انجام عملگر تقاطع، ابتدا والدین انتخاب گشته، سپس فرزندان با استفاده از عملگر تقاطع یکنواخت تولید می‌شوند. عملیات جهش نیز بر روی هر درآیه از ماتریس موجود در کرموزوم انجام می‌شود. دراین عملگر، پس از انتخاب والد مورد نظر، به‌ازای هر ژن در کروموزوم والد، عددی تصادفی بین صفر و یک تولید می‌شود و با نرخ جهش مشخص، مقادیر ژن‌های کروموزوم والد مورد جهش قرارمی‌گیرد.

طبقه بندی مشتریان با استفاده از مدل ترکیبی مبتنی بر ماشین بردار پشتیبان و الگوریتم ژنتیک:

با توجه به توضیحات ارائه شده در مورد ماشین‌های بردار پشتیبان و الگوریتم ژنتیک، در این بخش بنا داریم از یک روش ترکیبی مبتنی بر شبکهٔ ماشین‌های بردار پشتیبان و بهینه‌سازی با

استفاده از الگوریتم ژنتیک طبقه‌بندی مشتریان مشترک صنعت بانکداری و بیمه را مدل‌سازی نماییم؛ برای این منظور ابتدا متغیرهای مستقل و وابسته مشخص می‌شود. در این تحقیق، اطلاعات هویتی مشتریان تحت عنوان متغیرهای مستقل و طبقه‌ای که هر مشتری در آن قرار گرفته است به عنوان متغیر وابسته تعریف می‌شود. در مرحلهٔ بعد مجموعهٔ مشتریان به دو دسته داده‌های آموزش و داده-های آزمایش تقسیم‌بندی می‌شوند؛ تقسیم‌بندی داده‌ها به دو گروه آموزش و آزمایش به صورت تصادفی صورت می‌گیرد؛ به این صورت که ۹۰ درصد از داده‌ها در فاز آموزش و مابقی در فاز آزمایش به کار گرفته می‌شود.

با توجه به اینکه عملکرد دسته‌بندی در ماشین‌های بردار پشتیبان به پارامترهای آن بستگی دارد باید هر ماشین بردار پشتیبان با مجموعه پارامترهای کارا مورد استفاده قرار گیرد. در این تحقیق پارامترهای قابل تنظیم برای ماشین‌های بردار پشتیبان عبارتند از:

- ضریب جریمه (C)
- خطای قابل پذیرش (ε)
- میزان انحراف معیار تابع کرنل گوسی مورد استفاده (δ)

پارامترهای فوق به صورت یکپارچه و هم‌زمان برای ماشین بردار پشتیبان –با استفاده از الگوریتم ژنتیک– تنظیم می‌شود میزان برازش فرض شده برای هر پاسخ در الگوریتم ژنتیک خطای کلی طبقه‌بندی می‌باشد.

یافته‌ها

در بخش‌های قبل، مدل پیشنهادی برای طبقه بندی مورد بحث قرار گرفت. همچنین روش و چارچوب تحقیق و روش پیشنهادی برای ارزیابی و بخش‌بندی مشتریان مشترک صنعت بانکداری و بیمه بر اساس ارزش مشتریان نیز به‌طور خلاصه شرح داده شد. در این فصل به پیاده‌سازی روش پیشنهادی که در بخش قبل به تفصیل مورد بحث و بررسی قرار گرفت، پرداخته شده است. این روش با توجه به مورد مطالعاتی که برای این تحقیق در نظر گرفته‌شده، پیشنهادشده است. بنابراین عوامل مؤثر بر ارزش مشتریان نیز با شرایط اطلاعات اخذ شده از این مؤسسات منطبق شده است.

اطلاعات هویتی و شاخص‌های عملکری مشتریان

در این تحقیق با توجه به اهداف ترسیم شده، اطلاعات مشتریان مشترک صنعت بانکداری و بیمه مورد نیاز بوده است. با توجه به اینکه هر یک از مؤسسات بانک ایران زمینه و بیمه آرمان اساسا تنها اطلاعات مشتریان خود را در پایگاه‌های اطلاعاتی ثبت می کردند، جمع آوری این اطلاعات با

پیچیدگی‌های بسیار زیادی همراه بود. در نهایت این اطلاعات با انطباق دادن پایگاه داده‌های این دو سازمان و همکاری بخشهای IT آنها جمع‌آوری شد.

اطلاعات هویتی مشتریان مشترک بیمه آرمان و بانک ایران زمین گردآوری شده است؛ این اطلاعات در مورد ۶۰۵۹۶ مشتری بوده است؛ پس از بررسی‌های اولیه و پیش‌پردازش اطلاعات، در نهایت جدول نهایی اطلاعات مشتریان مشترک به صورت جدول ۱ قابل ارائه است. در این جدول نحوۀ کدینگ اطلاعات مشتریان در فرایند پردازش اطلاعات ارائه می‌شود.

جدول ۱- اطلاعات هویتی مشتریان مشترک بانک و بیمه

Table 1: Identity information of bank and insurance common customers

سرفصل Headline	نام شاخص در پایگاه داده The index name in the database	نام فارسی شاخص The Persian name of the indicator	نحوه‌ی کدینگ coding way
اطلاعات هویتی Identity information	Cust_Brith_Date	سن	
	Sex_Desc	جنسیت	زن =۰، مرد = ۱، نامشخص = ۲
	Cust_Married_Desc	وضعیت تاهل	مجرد = ۰، متاهل = ۱، نامشخص = ۲
	Cust_Grad_Desc	وضعیت تحصیل	بیسواد = ۰، زیردیپلم =۱، دیپلم =۲، فوق دیپلم = ۳، لیسانس = ۴، فوق لیسانس = ۵، دکتری = ۶، نامشخص = ۷
	Cust_Group_Desc	حقیقی یا حقوقی بودن	حقوقی = ۱، حقیقی = ۰
	count_card	تعداد کارت های مشتری	
	has_longterm_deposit	افتتاح سپرده بلند مدت	ندارد = ۰، دارد = ۱
	Dpst_Group_Desc	نوع سپرده	سپرده سرمایه گذاری بلند مدت = ۱، سپرده سرمایه گذاری کوتاه مدت = ۲، سپرده قرض الحسنه پس انداز = ۳، سپرده قرض الحسنه جاری = ۴، غیره =۵
	Count_Acc	تعداد حساب در بانک	

پیش‌پردازش‌های اعمال شده بر روی داده‌ها:

داده های مستخرج از پایگاه داده های بانک و بیمه به صورت خام قابل استفاده نیست؛ در این مرحله خطاهای داده‌ها تصحیح می‌شوند و داده‌های اشتباه جایگزین می شوند تا بخش زیادی از زمان داده‌کاوی در این تحقیق را در برگیرد. در این مرحله به دلیل بالا بودن تعداد رکوردهای اطلاعات مشتریان مشترک بانک و بیمه و تراکنش‌های آنان، گزارش‌گیری بسیار زمان‌بر و نیاز به گرفتن گزارش‌ها در چند مرحله بود. پس از دریافت جداول توسط نرم‌افزار SQLSERVER به منظور انجام این پژوهش، بانک اطلاعاتی یکپارچه‌ای، شامل مشتریان مشترک بانک و بیمه و

تراکنش‌های آن‌ها طی یک دوره زمان مشخص استخراج گردیده است. در انتها داده‌ها به فرمت فایل Excel آماده گردید. این اطلاعات با انطباق دادن پایگاه داده‌های این دو سازمان برای ۶۰۵۹۶ مشتری به نمایندگی از مجموعه مشتریان هدف جمع‌آوری شد. این تعداد مشتری برای مطالعهٔ حاضر و ساخت الگوهای مورد نیاز پایلوت مناسبی محسوب می‌شود.

حذف داده‌های ناقص، بی‌کیفیت و مغشوش: در این گام برخی از پارامترهای موجود که دارای اطلاعات مفقوده زیاد بودند یا با سایر اطلاعات سازگاری نداشتند حذف شدند.

نرمال‌سازی داده‌ها جهت نگاشت به بازه‌ی بین صفر و یک

نکته مهمی که باید خاطرنشان کرد این است که به دلیل آنکه اطلاعات هویتی در طیف‌های مختلف جای می‌گیرند، به‌منظور کسب نتایج قابل‌اطمینان‌تر از فرایند خوشه‌بندی بهتر است تمام مقادیر با طیف مشابهی مقیاس‌بندی شوند. ما این کار را با مقیاس‌بندی تمام مقادیر در طیف بین ۰ و ۱ انجام می‌دهیم که بدین معناست که باید تمام مقادیر را به دامنه تغییرات هر پارامتر تقسیم کنیم. در این تحقیق از نرمال‌سازی Max-Min که یک انتقال خطی روی داده‌های اصلی ایجاد می‌کند استفاده شده است.

به دلیل تفاوت در واحد هر یک از شاخص‌ها، لازم است تا مقادیر این شاخص‌ها بر اساس یک واحد یکسان، نرمال‌سازی گردند؛ این شاخص‌ها با استفاده از فرمول‌های زیر، بین اعداد ۰ تا ۱ نرمال شدند:

رابطه نرمال سازی برای شاخص‌های منفی: $x' = \frac{x_{max} - x}{x_{max} - x_{min}}$

رابطه نرمال سازی برای شاخص‌های مثبت: $x' = \frac{x - x_{min}}{x - x_{min}}$,

در رابطه‌های بالا x_{max} نشان‌دهنده بیشترین مقادیر شاخص‌ها هستند و x_{min} بیانگر کمترین مقادیر شاخص‌ها هستند و x نیز مقادیر اصلی شاخص‌ها را نشان می‌دهند. در نهایت x' نیز نشان‌دهنده مقادیر نرمال شده شاخص‌ها می‌باشد؛ در فرایند نرمال‌سازی، شاخص‌های عملکردی منفی ضمن نگاشت در بازه بین صفر و یک، جنبهٔ مثبت پیدا می کنند. این شاخص‌ها شامل مجموع برداشت، تعداد برداشت، تعداد چک‌های برگشتی در طول دوره، خسارت دریافتی توسط مشتری و تعداد خسارت می‌شوند.

پیاده‌سازی متدلوژی ترکیبی شبکه ماشین‌های بردار پشتیبان و الگوریتم ژنتیک برای طبقه‌بندی مشتریان

در این تحقیق مشتریان مشترک بانک و بیمه در دو دسته مشتریان با ریسک کم و مشتریان پر ریسک طبقه بندی می شوند؛ فرایند برچسب‌گذاری مشتریان بر اساس شاخص‌های عملکردی آنها در بانک و بیمه مشخص می گردد؛ این شاخص‌ها به طور کلی شامل موارد زیر می‌شود:

۱) شاخص های عملکردی در حوزه بانک
- ✓ مجموع مبلغ واریز
- ✓ مجموع برداشت
- ✓ تعداد واریز
- ✓ تعداد برداشت
- ✓ تعداد چک های برگشتی در طول دوره
- ✓ میانگین مبالغ تراکنش های مشتری

۲) شاخص های عملکردی در حوزه بیمه
- ✓ حق بیمه پرداختی
- ✓ تعداد بیمه نامه صادره
- ✓ خسارت دریافتی
- ✓ تعداد خسارت

تمرکز اصلی این تحقیق برروی ساخت مدل داده‌محور، جهت طبقه‌بندی مشتریان و بهره-برداری از آن برای پیش‌بینی عملکرد مشتریان جدیدالورود است؛ لذا روال اجرایی مدل پیشنهادی در این تحقیق بعد از فرایند برچسب‌گذاری مشتریان صورت می‌گیرد. دو دسته به دست آمده از فرایند برچسب‌گذاری به ترتیب دارای حدودا ۷۰ و ۳۰ درصد مجموعه مشتریان مورد مطالعه هستند. مشتریان قرار گرفته دردسته ۲ دارای ارزش بیشتری بر اساس مجموع شاخص‌های عملکردی در حوزۀ بانک و بیمه هستند.

در واقع مشتریان دستۀ اول، مشتریان پرریسک و مشتریان دستۀ دوم، مشتریان کم ریسک محسوب می‌شوند.

جدول ۲- برچسب‌گذاری مشتریان بر اساس شاخص های عملکردی

Table 2: Costumer labeling based on the functional indexes

دسته Class	تعداد اعضای هر خوشه The number of members of each cluster	درصد اعضای هر خوشه Percentage of members of each cluster	برچسب Label
1	42216	69.67%	پرریسک
2	18380	30.33%	کم‌ریسک

در این بخش با پیاده‌سازی مدل ترکیبی الگوریتم ژنتیک و ماشین‌های بردار پشتیبان بناست تا طبقه مشتریان جدیدالورود را مدل‌سازی و پیش‌بینی نماییم؛ بنابراین ابتدا مشتریان مشترک بیمه و بانک را به دو گروه داده‌های آموزش و داده‌های آزمایش تقسیم‌بندی می‌کنیم؛ تقسیم‌بندی داده‌ها بین دو گروه به صورت تصادفی صورت گرفته؛ با حفظ این شرط که در هر مرحله، مشتریان هر طبقه وجود داشته باشند. برای این منظور حدود ۱۰ درصد از داده‌ها به صورت تصادفی برای فاز آزمایش و مابقی آن‌ها برای فاز آموزش انتخاب می‌شوند. به طور دقیق‌تر ۶۰۶۰ مشتری برای فاز آزمایش کنار گذاشته شده و در فاز آموزش، اطلاعات آن‌ها به ماشین‌های بردار پشتیبان داده نمی‌شود. در فاز آموزش برای ۵۰۴۳۶ مشتری که اطلاعات آن‌ها جمع‌آوری شده است ماشین بردار پشتیبان طراحی و تنظیم می‌شود. بر اساس آنچه در بخش ۵ گفته شد تنظیم پارامترهای ماشین بردار پشتیبان با استفاده از الگوریتم ژنتیک صورت گرفته است. تعداد تکرارهای الگوریتم ژنتیک مورد استفاده در این بخش ۵۰ تکرار، تعداد جمعیت هر نسل ۵۰ جواب، نرخ تقاطع ۰,۷ و نرخ جهش ۰,۳ در نظر گرفته شده است.

خروجی الگوریتم ژنتیک، پارامترهای تنظیم شده ماشین بردار پشتیبان جهت طبقه‌بندی مشتریان است؛ بر این اساس الگوریتم ژنتیک پارامترهای ماشین بردار پشتیبان را تنظیم نموده و در خروجی گزارش می‌کند. در ادامه پارامترهای ماشین‌های بردار پشتیبان مستخرج از الگوریتم ژنتیک ارائه داده می‌شود.

جدول ۳- پارامترهای بهینه شده ماشین بردار پشتیبان با استفاده از الگوریتم ژنتیک

Table 3: Optimized parameters of support vector machine using genetic algorithm

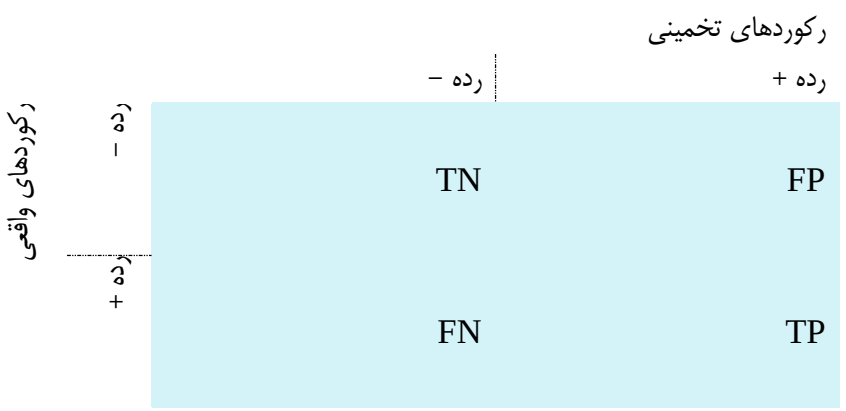
پارامتر Parameter	مقدار Value
C	4.99
ε	0.51
δ	0.26

ارزیابی نتایج

به منظور ارزیابی مدل پیشنهادی در این تحقیق، از مفاهیم دقت، بازخوانی و صحت[1] مدل استفاده شده است؛ پیش از برشمردن معیارهای ارزیابی، باید مفهوم ماتریس درهم‌ریختگی روشن شود؛ این ماتریس، چگونگی عملکرد الگوریتم رده‌بندی را با توجه به مجموعهٔ دادهٔ ورودی به تفکیک

1.Accuracy- Recall-Precision

انواع رده‌های مسئله نشان می‌دهد. شکل ۴ یک ماتریس درهم‌ریختگی را برای مسئله‌ای نشان می‌دهد که دارای دو رده "+" و "−" است. هدف مسئله، تشخیص رکوردهای با رده مثبت از داده‌هایی است که تاکنون دیده نشده است.



شکل ٤: ماتریس درهم‌ریختگی
Figure 4: Confusion matrix

مفاهیم ماتریس درهم‌ریختگی به شرح زیر تعریف می‌شوند:

- ✓ تعداد منفی‌های صحیح[1] (TN): تعداد رکوردهایی که ردۀ واقعی آن‌ها منفی بوده و الگوریتم رده‌بندی نیز آن‌ها را به‌درستی منفی تشخیص داده است.
- ✓ تعداد مثبت‌های ناصحیح[2] (FP): تعداد رکوردهایی که ردۀ واقعی آن‌ها منفی بوده ولی الگوریتم رده‌بندی آن‌ها را به‌اشتباه مثبت تشخیص داده است.
- ✓ تعداد منفی‌های ناصحیح[3] (FN): تعداد رکوردهایی که ردۀ واقعی آن‌ها مثبت بوده ولی الگوریتم رده‌بندی آن‌ها را به‌اشتباه منفی تشخیص داده است.
- ✓ تعداد مثبت‌های صحیح[4] (TP): تعداد رکوردهایی که ردۀ واقعی آن‌ها مثبت بوده ولی الگوریتم رده‌بندی آن‌ها را به‌درستی مثبت تشخیص داده است.

مهم‌ترین معیار برای تعیین کارایی تکنیک دسته‌بندی معیار دقت[1] است؛ این معیار، دقت کل یک دسته‌بندی را محاسبه می‌نماید و نشان‌دهندۀ این حقیقت است که دسته‌بندی طراحی‌شده چند

1.True Negative
2.False Positive
3.False Negative
4.True Positive

درصد از کل مجموعه رکوردهای آزمایشی را به‌درستی دسته‌بندی کرده است. دقت دسته‌بندی با استفاده از رابطهٔ زیر به‌دست می‌آید که بیان می‌کند دو مقدار TP و TN مهم‌ترین مقادیری هستند که در یک مسئلهٔ دودسته‌ای باید بیشینه شوند. مشکل اصلی هم، نامتعادل بودن داده‌ها و تفاوت معنی‌دار تعداد نمونه‌های هر دسته است که باعث می‌شود یک مدل متمایل به دستهٔ پرتعداد، دقت کلی را بالا نشان دهد؛ بنابراین نیاز به معیاری دقیق‌تر برای سنجش دقت و کارایی الگوریتم‌های پیشنهادی دسته‌بندی هستیم، که در رابطه زیرنمایش آمده است. گاهی بازخوانی ما به خاطر ضعیف بودن مدل پیشنهادی، بالاست. این ضعیف بودن را با معیار دیگری باید اندازه بگیریم. برای حل این مشکل، در کنار معیار بازخوانی، معیار دیگری را به نام صحت[۲]، برابر تعداد نمونه‌های تشخیصی درست مثبت به‌کل نمونه‌های مثبت اعلام‌شده به‌صورت رابطه زیر تعریف می‌کنیم تا میزان مثبت‌های اشتباه را هم در نظر گرفته باشیم.

$$\text{Accuracy}=\frac{tp+tn}{tp+tn+fp+fn}$$

$$\text{Recall}=\frac{tp}{tp+fn}$$

$$\text{Precision}=\frac{tp}{tp+fp}$$

در این بخش ماشین بردار پشتیبان با استفاده از نرم‌افزار MATLAB پیاده‌سازی شده است. خطای فاز آموزش با ماشین بردار پشتیبان ۰,۰۰۰۰۵ می باشد. در جدول ۴، دقت و خطای پیش‌بینی طبقه‌بندی مشتریان توسط ماشین بردار پشتیبان در قالب ماتریس درهم‌ریختگی، به تفکیک دسته‌ها مشاهده می‌شود.

1.Accuracy
2.Precision

جدول ٤- دقت و خطای پیش‌بینی خوشه مشتریان توسط ماشین بردار پشتیبان در داده‌های فاز آزمایش
Table 4: Accuracy and error of customer cluster prediction by support vector machine in test phase data

پیش‌بینی Prediction / واقعی Real	1	2
1	4245	2
2	0	1813

همانطور که در ابتدای این بخش از مقاله مطرح شد، معیارهای دقت، بازخوانی و صحت برای ارزیابی روش‌های پیش‌بینی طبقۀ مشتریان مشترک بیمه و بانک در این تحقیق مورد استفاده قرار می‌گیرند؛ مهم‌ترین معیار برای تعیین کارایی تکنیک‌های دسته‌بندی معیار Accuracy است؛ این معیار، دقت کل یک دسته‌بندی را محاسبه می‌نماید و نشان‌دهنده این حقیقت است که دسته‌بندی طراحی‌شده، چند درصد از کل مجموعه رکوردهای آزمایشی را به‌درستی دسته‌بندی کرده است. نتایج جدول ۵ نشان می‌دهد که ماشین بردار پشتیبان تنظیم‌شده به وسیله الگوریتم ژنتیک برای طبقه‌بندی مشتریان، ۹۹,۹۸ درصد داده‌های آزمایشی را به درستی تشخیص داده است و با توجه به بالا بودن درصد سه معیار دقت، بازخوانی و صحت این روش ترکیبی نتیجه می‌گیریم این روش به طور کارآمدی قادر به طبقه بندی مشتریان مشترک بانک و بیمه است.

جدول ٥: مقادیر دقت، یادآوری و صحت روش ترکیبی الگوریتم ژنتیک و ماشین بردار پشتیبان
Table 5: Accuracy, Recall and Precision values of hybrid method of genetic algorithm and support vector machine

دقت Accuracy	بازخوانی Recall	صحت Precision
99.97%	99.94%	99.98%

بحث و نتیجه‌گیری

در این تحقیق، با پیاده‌سازی‌سازی ماشین بردار پشتیبان برای طبقه‌بندی مشتریان مشترک بانک و بیمه به بررسی نتایج حاصل از آن پرداختیم؛ به طوری که پس از طی فرایند آموزش و دست‌یابی به پارامترهای بهینۀ ماشین‌های بردار پشتیبان با استفاده از الگوریتم ژنتیک عملکرد این روش در فاز آزمایش با ۶۰۶۰ مشتری که اطلاعات آن در فاز آموزش به ماشین‌های بردار پشتیبان

داده نشده است ارزیابی شد. مقایسه خروجی شبکه ماشین‌های بردار پشتیبان با طبقه واقعی مشتریان، حکایت از تناسب مناسب خروجی‌های به دست آمده از شبکه ماشین‌های بردار پشتیبان با داده‌های واقعی دارد. با توجه به نتایج به دست آمده خطای طبقه‌بندی مدل پیشنهادی ۰,۰۰۰۳ می‌باشد. این نتایج بدان معنی است که دقت عملکرد ماشین بردار پشتیبان حدود ۹۹,۹۷ درصد است که به این ترتیب می‌تواند دقت قابل‌قبولی قلمداد شود. امروزه در اکثر سازمان‌ها، داده‌ها به سرعت در حال جمع‌آوری و ذخیره شدن می‌باشند. با وجود این، می‌توان ادعا کرد که علی‌رغم وجود حجم انبوه داده-ها، سازمان‌ها عموماً با فقر دانش در تصمیم‌گیری‌ها روبرو هستند. اگرچه با استفاده از ابزارهای گوناگون گزارش‌گیری معمولی می‌توان اطلاعاتی را در اختیار کاربران قرارداد تا بتوانند به نتیجه‌گیری در مورد داده‌ها و روابط منطقی میان آن‌ها بپردازند؛ اما هنگامی که حجم عظیمی از داده‌ها مطرح باشد، حتی کاربران حرفه‌ای و باتجربه نیز نمی‌توانند الگوهای مفید را در میان انبوه داده‌ها تشخیص دهند. امروزه تکنیک‌های یادگیری ماشینی جهت پاسخگویی به نیازهای سازمان‌ها و شرکت‌های مختلف در کشف دانش از حجم انبوه داده مورد توجه قرار گرفته‌اند. داده‌کاوی فرآیند استخراج اطلاعات و دانش و کشف الگوهای پنهان از یک پایگاه داده بسیار بزرگ می‌باشد. شرکت‌های مخابراتی، بانک‌ها، بیمه‌ها، شرکت‌های تبلیغاتی و کلیه شرکت‌هایی که از بانک‌های اطلاعاتی بزرگی برخوردار هستند با استفاده از داده‌کاوی می‌توانند فرآیندهای تصمیم‌گیری خود را بهبود بخشند. داده-کاوی سبب می‌شود که سازمان‌ها از سطح داده به سطوح بالاتر دانش و الگوهای ناشناخته برسند. الگوهای استخراج شده می‌توانند رابطه‌ای بین ویژگی‌ها و مشخصات سیستم مانند نوع تقاضا و نوع مشتری، پیش‌بینی‌های آینده براساس مشخصات سیستم، قوانین (اگر ــ آنگاه) بین متغیرهای سیستم، دسته‌بندی‌ها و خوشه‌بندی‌های اشیاء و رکوردهای شبیه به هم در یک سیستم و غیره باشند.

تعارض منافع

نویسندگان هیچگونه تعارض منافعی ندارند.

References

Abdou, H., Pointon, J., & El-Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. Expert Systems with Applications, 35(3), 1275-1292. **doi:10.1016/j.eswa.2007.08.030**

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision support systems, 50(3), 602-613. **doi:10.1016/j.dss.2010.08.008**

Boyacioglu, M. A., Kara, Y., & Baykan, Ö. K. (2009). Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. Expert Systems with Applications, 36(2), 3355-3366. **doi:10.1016/j.eswa.2008.01.003**

Chen, F. L., & Li, F. C. (2010). Combination of feature selection approaches with SVM in credit scoring. Expert systems with applications, 37(7), 4902-4909. **doi:10.1016/j.eswa.2009.12.025**

Chu, B. H., Tsai, M. S., & Ho, C. S. (2007). Toward a hybrid data mining model for customer retention. Knowledge-Based Systems, 20(8), 703-718. **do:10.1016/j.knosys.2006.10.003**

Dorofeev, D., Khrestina, M., Usubaliev, T., Dobrotvorskiy, A., & Filatov, S. (2018, May). Application of machine analysis algorithms to automate implementation of tasks of combating criminal money laundering. In International Conference on Digital Transformation and Global Society (pp. 375-385). Springer, Cham.

Duman, E., & Ozcelik, M. H. (2011). Detecting credit card fraud by genetic algorithm and scatter search. Expert Systems with Applications, 38(10), 13057-13063. **doi:10.1016/j.eswa.2011.04.110**

Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. Expert systems with applications, 33(4), 847-856. **doi:10.1016/j.eswa.2006.07.007**

Huang, Y. M., Hung, C. M., & Jiau, H. C. (2006). Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. Nonlinear Analysis: Real World Applications, 7(4), 720-747. **doi:10.1016/j.nonrwa.2005.04.006**

Jamshidi, M. B., Gorjiankhanzad, M., Lalbakhsh, A., & Roshani, S. (2019, May). A novel multiobjective approach for detecting money laundering with a neuro-fuzzy technique. In 2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC) (pp. 454-458). IEEE. **doi:10.1109/ICNSC.2019.8743234**

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert systems with applications*, *32*(4), 995-1003. **doi:10.1016/j.eswa.2006.02.016**

Lee, B., Cho, H., Chae, M., & Shim, S. (2010). Empirical analysis of online auction fraud: Credit card phantom transactions. *Expert Systems with Applications*, *37*(4), 2991-2999. **doi:10.1016/j.eswa.2009.09.034**

Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. Computational Statistics & Data Analysis, 50(4), 1113-1130. **doi:10.1016/j.csda.2004.11.006**

Lin, C. S., Tzeng, G. H., & Chin, Y. C. (2011). Combined rough set theory and flow network graph to predict customer churn in credit card accounts. Expert Systems with Applications, 38(1),8-15. **doi:10.1016/j.eswa.2010.05.039**

Lin, S. W., Shiue, Y. R., Chen, S. C., & Cheng, H. M. (2009). Applying enhanced data mining approaches in predicting bank performance: A case of Taiwanese commercial banks. Expert Systems with Applications, 36(9), 11543-11551. **doi:10.1016/j.eswa.2009.03.029**

Luo, S. T., Cheng, B. W., & Hsieh, C. H. (2009). Prediction model building with clustering-launched classification and support vector machines in credit scoring. Expert Systems with Applications, 36(4), 7562-7566. **doi:10.1016/j.eswa.2008.09.028**

Magomedov, G. S., Dobrotvorsky, A. S., Khrestina, M. P., Pavelyev, S. A., & Yusubaliev, T. R. (2018). Application of Artificial Intelligence Technologies for the Monitoring of Transactions in AML-Systems Using the Example of the Developed Classification Algorithm. Int. J. Eng. Technol, 7, 76-79.

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. **doi:10.1016/j.eswa.2011.06.028**

Paasch, C. A. (2008). Credit card fraud detection using artificial neural networks tuned by genetic algorithms. Hong Kong University of Science and Technology (Hong Kong), 1-1112.

Plaksiy, K., Nikiforov, A., & Miloslavskaya, N. (2018, August). Applying big data technologies to detect cases of money laundering and counter financing of terrorism. In 2018 6th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW) (pp. 70-77). IEEE. **10.1109/W-FiCloud.2018.00017**

Sobreira Leite, G., Bessa Albuquerque, A., & Rogerio Pinheiro, P. (2019). Application of technological solutions in the fight against money laundering—A systematic literature review. *Applied Sciences*, *9*(22), 1-29. **doi:10.3390/app9224800**

Quah, J. T., & Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. *Expert systems with applications*, *35*(4), 1721-1732. **doi:10.1016/j.eswa.2007.08.093**

Sánchez, D., Vila, M. A., Cerda, L., & Serrano, J. M. (2009). Association rules applied to credit card fraud detection. *Expert systems with applications*, *36*(2), 3630-3640. **doi:10.1016/j.eswa.2008.02.001**

Šušteršič, M., Mramor, D., & Zupan, J. (2009). Consumer credit scoring models with limited data. Expert Systems with Applications, 36(3), 4736-4744. **doi:10.1016/j.eswa.2008.06.016**

Tiwari, M., Gepp, A., & Kumar, K. (2020). A review of money laundering literature: the state of research in key areas. *Pacific Accounting Review*, Vol. 32 No. 2, pp. 271-303. **doi:10.1108/PAR-06-2019-0065**

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, *36*(3), 5445-5449. **doi:10.1016/j.eswa.2008.06.121**

Yap, B. W., Ong, S. H., & Husain, N. H. M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications*, *38*(10), 13274-13283. **doi:10.1016/j.eswa.2011.04.147**

Zhao, H., Sinha, A. P., & Ge, W. (2009). Effects of feature construction on classification performance: An empirical study in bank failure prediction. *Expert Systems with Applications*, *36*(2), 2633-2644. **doi:10.1016/j.eswa.2008.01.053**