شناسایی و تعیین عوامل مؤثر بر رفتار اعتباری متقاضیان حقیقی تسهیلات خرد در شهر تهران؛ با نگرش بر کاهش نرخ رشد

مطالبات در یکی از بانک‌های خصوصی

خشایار مقدم[1]
علی ارشدی[2]
کیانوش رضایی[3]

تاریخ دریافت: ۱۳۹۱/۱۲/۲۵ تاریخ پذیرش: ۱۳۹۲/۰۲/۲۰

چکیده

این پژوهش با هدف شناخت عوامل مؤثر بر رفتار اعتباری متقاضیان حقیقی تسهیلات خرد[4] تدوین شده است. در این راستا پژوهشگران اقدام به شناسایی ۳۶ شاخص کرده و سپس با استفاده از روش نمونه‌گیری احتمالی طبقه‌بندی‌شده،[5] ۴۲۹ پرونده اعتباری انتخاب و با بهره‌گیری از تکنیک‌های داده‌کاوی مانند انواع درخت تصمیم (CHAID،[6] CART،[7] QUEST،[8] C5،[9])، شبکه عصبی،[10] ماشین بردار پشتیبان،[11] رگرسیون لجستیک[12] و تحلیل تشخیصی[13] اقدام به ارزیابی میزان اثرگذاری این شاخص‌ها بر احتمال نکول تعهدات کردند. نتایج حاصل از پژوهش نشان می‌دهد که مدل شبکه عصبی در مقایسه با سایر روش‌ها بالاترین قابلیت پیش‌بینی را دارد. علاوه بر این از ۳۵ شاخص مورد بررسی دوازده شاخص مدت (دوره زمانی بازپرداخت)، نوع عقد، تعداد اقساط بازپرداخت، نوع وثیقه، تعداد چک برگشتی قبل از دریافت تسهیلات، مبلغ قسط، مدت سپرده‌گذاری (سپرده کوتاه‌مدت)، بخش اقتصادی، جنسیت، شغل، میانگین گردش شش ماه قبل از دریافت تسهیلات و وضعیت مالکیت به ترتیب بیشترین اهمیت را در توضیح رفتار اعتباری وام‌گیرندگان دارند. در نهایت قواعد به دست آمده از درخت تصمیم برای تصمیم‌سازی (طراحی یک مدل امتیازدهی اعتباری) در بانک استخراج شدند.

کلید واژه‌ها: تسهیلات خرد، بانک خصوصی، شبکه عصبی، SVM، رگرسیون لجستیک، تحلیل تشخیص، درخت تصمیم، احتمال نکول تعهدات و نمونه‌گیری طبقه‌بندی‌شده.

طبقه‌بندی JEL: C1, C4, E1, E5.

۱ـ کارشناس بانکی
۲ـ عضو هیئت علمی پژوهشکده پولی و بانکی بانک مرکزی
۳ـ کارشناس بانکی

۴ـ تسهیلات خرد در پژوهش حاضر عبارت است از تسهیلات با مبالغ یک میلیارد و زیر یک میلیارد ریال

5- Stratified Probability Sampling
6- Decision Tree
7- Chi-Squared Automatic Interaction Detection
8- CART
9- Quick, Unbiased, Efficient & Statistical Tree
10- Artificial Neural Network
11- Support Vector Machines (SVM)
12- Logistic Regression
13- Discriminant Analysis

۱. مقدمه

یکی از ریسک‌هایی که صنعت بانکداری همواره با آن مواجه بوده است عدم تمایل یا ناتوانی گیرندگان تسهیلات در بازپرداخت تسهیلات دریافتی در زمان مقرر است. در راستای حل این معضل، بانکداری سنتی در جهت شناسایی اهلیت و گرفتن وثیقه و تضمینات محکم از مشتریان گام برداشته است. گذر زمان نشان داد که روش پیش‌گفته کارایی چندانی در کاهش مقدار تسهیلات غیرجاری نداشته است. بنابراین ارائه روش‌هایی آینده‌نگر و مبتنی بر پیش‌بینی رفتار اعتباری مشتریان پیش از دریافت تسهیلات ضروری می‌نماید.

بر این اساس بانک‌ها اقدام به طراحی و به کارگیری سیستم امتیازدهی اعتباری مشتریان[1] به عنوان یک روش آینده‌نگر کردند. امتیازدهی اعتباری به عنوان یک روش ارزیابی اعتبار، پیشینه‌ای بیش از ۵۰ سال دارد. در این پژوهش، ابتدا شاخص‌های مؤثر بر احتمال نکول تعهدات شناسایی شده، سپس از طریق مکاتبه با شعب اطلاعات مورد نظر جمع‌آوری شد، در ادامه ضمن آماده‌سازی و پالایش داده‌ها از انواع تکنیک‌های داده‌کاوی و آماری همچون درخت تصمیم (CHAID، CART، QUEST، C5)، شبکه عصبی مصنوعی، SVM، رگرسیون لجستیک و تحلیل تشخیصی برای ارزیابی احتمال نکول تعهدات اعتباری مشتریان حقیقی استفاده شده است. در این پژوهش متغیرهای مستقل، شاخص‌های مؤثر بر احتمال نکول تعهدات و متغیر وابسته دو وجهی، مشتریان خوش‌حساب و بدحساب هستند.

۲. ضرورت و بیان مسئله

طی سال‌های اخیر با اجرای نارسای طرح هدفمندکردن یارانه‌ها همزمان با فراگیرشدن و عمق‌یافتن رکود اقتصادی در کشور و تبدیل آن به بحران، فعالان اقتصادی دچار مشکلات عدیده مالی شده‌اند. این مسئله اثرات خود را در قالب کمبود نقدینگی، زیان‌دهی فعالیت‌های اقتصادی، ناتوانی در انجام تعهدات و در نهایت عدم بازپرداخت تمام یا بخشی از تسهیلات دریافتی از بانک‌ها، نشان داده است.

بانک خصوصی مورد مطالعه به عنوان یک بنگاه انتفاعی در راستای عمل به تعهدات خود در قبال مشتریان و سهامداران؛ حفظ منابع و بیشینه‌کردن سود، مکلف به طراحی ابزارهایی کارآمد بوده که با بهره‌گیری از آنها بتواند ضمن اعطای انواع تسهیلات به مشتریان و متقاضیان نسبت به

1- Credit Scoring

بازگشت اصل و سود مورد انتظار نیز تا حد زیادی مطمئن باشد. در چنین شرایطی بانک کمتر با مشکل بلوکه‌شدن بخشی از منابع و سود پیش‌بینی‌شده در اثر عدم بازپرداخت اقساط توسط تسهیلات‌گیرندگان مواجه می‌شود. این پژوهش درصدد پاسخگویی به پرسش‌های زیر است:

۱ـ عوامل مؤثر بر رفتار اعتباری متقاضیان تسهیلات خرد چیست؟

۲ـ سهم هر یک از عوامل مؤثر در شکل‌دهی رفتار اعتباری متقاضیان تسهیلات خرد چگونه است؟

۳ـ قواعد حاکم بر رفتار اعتباری[1] مشتریان چیست؟

پاسخ به پرسش‌های بالا این امکان را به بانک می‌دهد که ضمن حفظ حجم تسهیلات اعطایی در آینده، نرخ رشد تسهیلات غیرجاری را کاهش دهد.

۳. پیشینه پژوهش

تاکنون مقالات زیادی در خصوص شناسایی، پیش‌بینی و دسته‌بندی عوامل مؤثر بر احتمال نکول متقاضیان تسهیلات، چاپ و منتشر شده است. در ادامه به تعدادی از آنها اشاره می‌کنیم:

کشاورز و آیتی (۱۳۸۶) در پژوهشی که در بانک مسکن انجام داده‌اند، اقدام به تعریف شاخص‌های تعداد فرزندان، درجه تحصیلی و وضعیت شغلی همسر و شخص درخواست‌کننده اعتبار، صاحبخانه‌بودن یا نبودن گیرنده اعتبار، سن و جنسیت گیرنده اعتبار برای ارزیابی رفتار متقاضیان تسهیلات کرده‌اند. آنها با استفاده از مدل پارامتریک رگرسیون لجستیک و مدل ناپارامتریک درخت تصمیم کارت اقدام به پیش‌بینی و دسته‌بندی ۲۴۰ نفر از مشتریان حقیقی یکی از شعب بانک مسکن به دو گروه خوش‌حساب و بدحساب کردند. نتیجه این پژوهش نشانگر برتری دقت پیش‌بینی با استفاده از درخت تصمیم کارت بر رگرسیون لجستیک در تمامی نمونه‌ها بوده است.

با هدف اعتبارسنجی مشتریان حقیقی، جلیلی، خدایی و کنشلو (۱۳۸۷) با استفاده از رگرسیون لجستیک، اقدام به جمع‌آوری اطلاعات فردی و اعتباری یک نمونه تصادفی ۱۰۰۰نفری از اعتبارگیرندگان سیستم بانکی از سال ۱۳۸۰ تا ۱۳۸۶ کرده‌اند. شاخص‌های مورد بررسی شامل جنسیت، سن، وضعیت تاهل، تحصیلات، شغل، مدت وام، مبلغ تسهیلات، وثیقه، ارزش وثیقه و هدف متقاضی از دریافت وام است. نتیجه پژوهش نشان می‌دهد اثر دو شاخص

۱ـ دریافت تسهیلات و بازپرداخت آن

سن و تحصیلات بر وضعیت اعتباری تأیید نشده ولی اثر سایر شاخص‌ها بر وضعیت اعتباری تأیید شده است. همچنین مدل مورد استفاده از قدرت پیش‌بینی خوبی برخوردار است.

اسماعیلی و رحمانی (۱۳۸۹) طی پژوهشی اقدام به مقایسه توانایی مدل‌های شبکه عصبی، رگرسیون لجستیک و تحلیل تشخیص برای پیش‌بینی ریسک نکول کرده‌اند. این پژوهشگران با استفاده از اطلاعات ۲۳۸۰۱ قرارداد لیزینگ از سه شرکت مختلف با انتخاب متغیرهای مدت قرارداد، مبلغ قرارداد، نوع صنعت، نوع قرارداد، نوع تضمین و سیاست‌های اعتباری به عنوان متغیر، پیش‌بینی و ریسک نکول دو وجهی به عنوان متغیر پیش‌بینی‌شونده اقدام به برازش مدل‌های یادشده کردند. نتایج کلی پژوهش حاکی از معنادار بودن متغیرهای پیش‌بین در برآورد مدل‌های فوق است. مقایسه قدرت پیش‌بینی، نشان‌دهنده برتری شبکه عصبی بر سایر مدل‌هاست.

با استفاده از اطلاعات مالی و اعتباری ۹۰ مشتری حقوقی و بهره‌گیری از روش رگرسیون لجستیک دو وجهی پرویزیان، ذکاوت و محمدیان اقدام به دسته‌بندی مشتریان خوش‌حساب و بدحساب و با استفاده از رگرسیون لجستیک سه وجهی اقدام به دسته‌بندی مشتریان به سه دسته مشتریانی که بدون تأخیر، با تأخیر سه‌ماهه و با تأخیر بیش از سه ماه اقساط خود را بازپرداخت نموده‌اند، کرده‌اند. نتایج حاصل نشان‌دهنده قدرت تفکیک‌کنندگی بیشتر مدل دووجهی نسبت به مدل سه وجهی است به گونه‌ای که در مدل دو وجهی و سه وجهی به ترتیب ۷۵ و ۷۰ درصد مشتریان به درستی دسته‌بندی شده‌اند.

در پژوهشی دیگر محمدی صداقت و سپه‌وند (۱۳۸۸)، اقدام به رتبه‌بندی ۲۰۰ نفر از مشتریان حقیقی بانک تجارت در حوزه کسب‌وکار با استفاده از مدل‌های لاجیت و شبکه عصبی کردند. نتایج پژوهش نشان داد از بین ۱۸ شاخص مورد بررسی، متغیرهای نوع مجوز کسب، میزان تجربه در حوزه فعالیت، چک برگشتی و وثیقه بهترین پیش‌بینی‌کننده‌های مشتریان از نظر خوش‌حسابی و بدحسابی هستند. همچنین مقایسه میانگین نرخ رده‌بندی صحیح کلی دو مدل نشان داد شبکه عصبی با نرخ ۸۴/۵ بر مدل لاجیت با نرخ ۸۱/۵ درصد برتری دارد.

دهقانی و سوری (۱۳۸۸)، با استفاده از ۴۰ متغیر مستقل اقدام به طراحی مدلی برای رتبه‌بندی اعتباری مشتریان حقیقی بانک کارآفرین با استفاده از دو مدل Logit و Cart کردند. سپس نتایج حاصل از دو مدل را با روش فعلی ارزیابی مدیران بانک مقایسه کردند. یافته‌های پژوهش نشان داد. متغیرهای میانگین موجودی حساب، درآمد و تحصیلات مشتریان، نوع وثیقه و مبلغ وام در دو روش Logit و Cart دارای بیشترین تأثیر در احتمال قصور مشتری بودند. همچنین دقت دو مدل بر اساس نسبت Roc محاسبه شده، برای دو مدل Logit و Cart به ترتیب

برابر ۷۰ و ۷۲ درصد است. در ضمن مقایسه دو مدل برآورد شده و روش مورد استفاده مدیران بانک نشان از بازدهی حداقل ۱۰ درصدی بیشتر این مدل‌ها بر روش ارزیابی مدیران دارد.

٤. چارچوب نظری

در این بخش الگوریتم‌های سنجش درجه اهمیت عوامل مؤثر بر رفتار اعتباری متقاضیان تسهیلات و روش‌های ارزیابی الگوریتم‌های به کارگرفته شده، معرفی می‌شود.

٤ـ۱. الگوریتم‌های سنجش درجه اهمیت عوامل مؤثر بر رفتار اعتباری متقاضیان تسهیلات

٤ـ۱ـ۱. درخت تصمیم

در سال ۱۹۶۳ میلادی درخت تصمیم برای اولین بار توسط مورگان[۱] و سانکویست[۲] بر مبنای ایده بلسون[۳] و دیگران (۱۹۵۹) معرفی شد. از معروف‌ترین روش‌های درخت تصمیم می‌توان به CHAID[۴]، ART[۵]، ID3[۶]، C4.5، C5، QUEST[۷] و AID[۸] اشاره کرد.

به طور کلی درخت‌های تصمیم بر دو گونه‌اند:

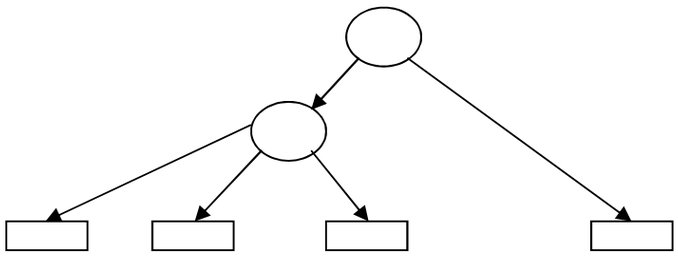
- درخت‌های تصمیم دوـ دویی، که هر گره، فقط دو شاخه انشعابی دارد.
- درخت‌های تصمیم خطی[۹]، که هر گره، می‌تواند به چند شاخه منشعب شود.

درخت تصمیم یک روش کارآمد طبقه‌بندی داده‌هاست که به دلیل برخورداری از ویژگی‌هایی مانند سادگی، قابل فهم بودن، دقت و سرعت الگوریتم، بسیار پرکاربرد است. یک درخت تصمیم از مجموعه نمونه‌های ورودی و خروجی ایجادشده، از روش یادگیری نظارتی استفاده کرده و با بیان یکسری قوانین ضمنی تعمیم‌پذیر، اقدام به پیش‌بینی رفتار متغیر وابسته (برچسب دسته) می‌کند.

ساختار درخت تصمیم، شبیه فلوچارت است و بالاترین گره در درخت، گره ریشه بوده و

1- Morgan
2- Sonquist
3- Belson
4- Chi-Squared Automatic Interaction Detection
5- Classification and Regression Trees
6- Iterative Dichotomizererd
7- Quick, Unbiased, Efficient & Statistical Tree
8- Automatic Interaction Detection
9- Classification & Regression Tree

گره‌های برگ، دسته‌ها یا توزیع دسته‌ها را نشان می‌دهند. در این فلوچارت گره با نماد دایره و برگ با نماد مستطیل نشان داده شده است.



٤-١-٢. روش کار درخت تصمیم

طراحی درخت تصمیم با طرح یکسری پرسش آغاز می‌شود، در ادامه با مشخص‌شدن پاسخ هر پرسش، سؤالی دیگر پرسیده می‌شود. در صورتی که پرسش‌ها متناسب با اهداف مورد نظر پرسیده شوند، مجموعه‌ای کوتاه از پرسش‌ها برای پیش‌بینی‌کردن دسته مربوط به شیء جدید کافی بوده و این پرسش‌ها تا آنجا ادامه می‌یابد که مجموعه‌ای از مشاهدات باقی بمانند که آنها را به هیچ طریقی نتوان در دسته یا گروه جدیدی قرار داد.

به عبارت دیگر، ساختار کلی درخت تصمیم به این صورت است که یک گره ریشه در بالای آن و برگ‌ها در پایین آن هستند، هر گاه یک رکورد یا ویژگی جدید در گره ریشه وارد شود در این گره یک آزمون صورت می‌گیرد تا معلوم شود این رکورد یا ویژگی متغیر به کدام یک از گره‌های فرزند تعلق دارد. روش‌های مختلفی برای انتخاب این آزمون وجود دارد که هدف همه آنها، انتخاب روشی است که بهترین جداسازی را در دسته‌های هدف انجام می‌دهد. این فرآیند تا آنجا ادامه می‌یابد که رکورد جدید به گره برگ برسد تا در یک دسته قرار گیرند. شایان ذکر است که برای رسیدن از ریشه به یک برگ، تنها یک راه وجود دارد و آن بیان قاعده‌ای است که برای دسته‌بندی رکوردها استفاده شده است.

متغیرهایی که برای پیش‌بینی متغیر هدف یا وابسته استفاده می‌شود، متغیرهای مستقل، توضیح‌دهنده یا پیش‌بینی‌کننده نامیده می‌شوند و به متغیرهای وابسته، برچسب دسته[1] می‌گویند.

1- Class Label

٤ـ١ـ٣. خصوصیات و نقاط قوت درخت تصمیم

- پیش‌بینی‌های حاصل از درخت تصمیم، در قالب یکسری قواعد توضیح داده می‌شود.
- از همه داده‌ها استفاده می‌کند.
- درک نتایج حاصل از درخت تصمیم آسان است.
- دسته‌بندی‌هایی که توسط درخت تصمیم ایجاد می‌شوند، از روی شباهت داده‌های ذخیره شده در پارامترهای پیش‌بینی‌کننده، انجام می‌شود.
- محدودیتی در استفاده از انواع داده‌ها ندارد.

این روش می‌تواند درجه تأثیر متغیرها در پیش‌بینی و دسته‌بندی را نشان دهد. به گونه‌ای که هر چه متغیر به ریشه نزدیک‌تر باشد از اهمیت بالاتر و بیشتری برخوردار است.

٤ـ١ـ٤. استخراج قواعد رده‌بندی حاصل از درخت‌های تصمیم (قواعد ضمنی)

می‌توان برای هر مسیر از ریشه تا هر برگ یک قانون ضمنی به صورت IF … THEN … (اگر ... آنگاه ...) تعریف کرد که به آسانی برای مدیران و کارمندان قابل توصیف است. در صورتی که قسمت اگر شامل تمام آزمون‌های یک مسیر باشد، قسمت آنگاه، بیانگر طبقه‌بندی نهایی است. قوانین در این فرم، قوانین تصمیم نامیده می‌شوند.

٤ـ١ـ٢. رگرسیون لجستیک[1]

رگرسیون لجستیک در اواخر دهه ۶۰ میلادی و اوایل دهه ۷۰ میلادی به عنوان به دیلی برای رگرسیون خطی و تحلیل تشخیصی در شرایطی که متغیر وابسته اسمی و متغیرهای مستقل از نوع متغیرهای ترتیبی و فاصله‌ای هستند، مطرح شد. این مدل تعمیم‌یافته رگرسیون خطی است و احتمال وقایع اتفاق افتاده (متغیر وابسته) را که تابع خطی از مجموع متغیرهای مستقل هستند، پیش‌بینی می‌کند. در این مدل با یک مجموعه داده (مجموعه آموزشی) می‌توان یک مدل لجستیک را برازش و با گروه دیگر داده‌ها می‌توان کیفیت مدل را در پیش‌بینی احتمال وقوع وجوه متغیر وابسته تحلیل کرد.

پس از پیش‌بینی احتمال موفقیت، باید انطباق مدل و معنادار بودن سهم هر کدام از متغیرهای مورد استفاده در مدل را به ترتیب با استفاده از آماره انحراف و شاخص والد که از توزیع نرمال پیروی می‌کند، بررسی کرد. آماره انحراف مبتنی بر توزیع نرمال بوده و با مقایسه مقدار آن با ناحیه بحرانی متناظر، در مورد انطباق مدل باید اظهار نظر شود.

1- Logistic Regression

٤ـ١ـ٣. تحلیل تشخیصی[1]

نظریه آغازین تحلیل تشخیصی به دهه ۳۰ میلادی و پژوهش‌های کارل پیرسون[2] و همکارانش در زمینه فواصل گروه‌ها و ضرایب تشابه نژادی باز می‌گردد. این روش در سال ۱۹۳۶ میلادی توسط فیشر[3] ابداع و برپایه روش‌شناسی مورد استفاده در رگرسیون خطی چندمتغیره و جبر ماتریسی، توسعه یافت.

در مواردی که متغیر وابسته اسمی، دو یا چند وجهی بوده و متغیرهای مستقل کمی باشند برای پیش‌بینی تغییرات متغیر وابسته از تحلیل تشخیصی استفاده می‌شود. دراین روش متغیر وابسته یا ملاک، با استفاده از تابع تشخیصی به شکل دسته‌ای تفکیک می‌شود. هدف اصلی این روش، تشخیص تفاوت بین گروه‌ها و پیش‌بینی احتمال تعلق یک عضو به گروه خاص است.

٤ـ١ـ٣ـ١. تابع تشخیصی[4] (DF)

ترکیب خطی متغیرهای مستقل استاندارد شده است که بزرگ‌ترین تفاوت‌های میانگین بین گروه‌ها را نشان می‌دهد. این تابع بر اساس مرکز ثقل گروه‌ها به گونه‌ای که این گروه‌ها کمترین همپوشی را داشته باشند، ایجاد می‌شود. به عبارت ساده‌تر متغیر جدیدی که از ترکیب متغیرهای مستقل به دست می‌آید و متغیر ملاک را به طبقات مختلف تقسیم می‌کند، تابع تشخیصی نامیده می‌شود که آن را با نماد DF نشان می‌دهند. این تابع از طریق حداقل مجذورات معمولی[5] و حداکثر درست‌نمایی محاسبه می‌شود که شکل جبری آن به صورت زیر است:

$$DF= b_1x_1 + b_2x_2 + \ldots + b_kx_k + b_0$$

که در آن:

DF: تابع تشخیصی

b_i: مقدار ضریب تشخیصی یا وزن هر متغیر که معادل ضریب β در رگرسیون خطی است.

x_i: متغیرهای مستقل و پیش‌بینی‌کننده است

٤ـ١ـ٣ـ٢. روش اجرای تحلیل تشخیصی

در این روش برای دسته‌بندی داده‌ها، مرزهایی بین داده‌ها به وسیله یکسری توابع ایجاد می‌کنند که این توابع مشخص‌کننده و جداکننده دسته‌های مختلف هستند. برای اجرای این

1- Discriminant Analysis
2- Carl Pearson
3- Fishe
4- Discriminant Function
5- Ordinary Least Squares

روش، ابتدا یک تابع که قابلیت تفکیک گروه‌ها از همدیگر را دارد، انتخاب می‌شود. سپس، دومین تابع با ویژگی دسته‌بندی‌کنندگی با رعایت دو شرط: ۱ـ با تابع قبلی ارتباط نداشته باشد؛ ۲ـ تا حد ممکن گروه‌ها را به تعداد بیشتری از همدیگر تفکیک کند، انتخاب می‌شود. این روند همچنان ادامه می‌یابد تا زمانی که بیشترین تعداد توابع جداکننده به دست آید.

٤ـ١ـ٤. شبکه‌های عصبی مصنوعی[1]

پیشینه شبکه عصبی به اوایل قرن بیستم و اواخر قرن نوزدهم بر می‌گردد، در این دوره کارهای بنیادی در فیزیک، روانشناسی و نورولوژی انجام شد. در آغاز دهه ۱۹۴۰ وارن مک کلوث[2] و والتر پیتز[3] نشان دادند که شبکه‌های عصبی می‌توانند هر تابع حسابی منطقی را محاسبه کنند.

نخستین کاربرد شبکه عصبی هنگامی که فرانک روزنبلات[4] و همکاران در سال ۱۹۵۸ شبکه پرسپترون با توانایی شناخت الگو را معرفی کردند، مطرح شد. تا سال‌های دهه ۷۰ میلادی شبکه عصبی به دلیل ضعف فناوری‌های نوین نتوانست نقش قابل توجهی در تحقیقات کاربردی ایفا کند. در سال‌های دهه ۸۰ میلادی با رشد تکنولوژی ریزپردازنده‌ها[5]، شبکه‌های عصبی با رشد فزاینده‌ای توسعه یافتند.

به طور کلی می‌توان گفت، شبکه عصبی روشی است که قصد دارد با استفاده از مدل‌های ریاضی و توان کامپیوتری، به شبیه‌سازی فرآیند یادگیری انسان بپردازد. به عبارت دیگر، شبکه عصبی مصنوعی یک «پردازنده توزیع‌شده موازی»[6] است که میل طبیعی برای ذخیره دانش تجربی و کاربردی‌کردن آن را دارد. شبکه عصبی مصنوعی از دو جهت به مغز انسان شباهت دارد.

- دانش از طریق یک فرآیند یادگیری توسط شبکه کسب می‌شود.
- قدرت ارتباطی که به عنوان وزن‌های سیناپسی شناخته می‌شود، برای ذخیره دانش مورد استفاده قرار می‌گیرد.

٤ـ١ـ٤ـ١. مراحل طراحی یک شبکه عصبی

- معماری: تعیین تعداد لایه‌های شبکه، نرون‌های هر لایه، پیش‌خور یا پس‌خوربودن شبکه و ... متناسب با نوع مسئله انتخاب نوع تابع فعال‌سازی
- آموزش شبکه: شبکه، قانون کار یا الگوی مناسب را یاد می‌گیرد یعنی پس از آموزش

1- Artificial Neural Networks
2- Warren McCulloch
3- Walter Pitts
4- Frank Rosenblatt
5- Microprocessors
6- Parallel Distributed Processor

به ازای هر ورودی، خروجی مناسب ارائه دهد.

٤ـ١ـ٤ـ٢. مزیت‌های شبکه عصبی

- شبکه عصبی، به دلیل پردازش‌های موازی، از سرعت پردازش بالایی برخوردار است.
- شبکه عصبی همانند مغز انسان همواره در حال یادگیری است.
- در شبکه عصبی، عدم عملکرد صحیح قسمتی از نرون‌ها، موجب از کار افتادگی شبکه نمی‌شود.
- این روش می‌تواند، برای داده‌ها در شرایط عدم اطمینان جواب منطقی ارائه دهد.

٤ـ١ـ٤ـ٣. محدودیت‌های شبکه عصبی

- شبکه عصبی توانایی توضیح منطق کار را ندارد. بنابراین برای حل مسائلی که کشف نوع و اندازه رابطه و قواعد حاکم مهم است، فاقد کارایی است.
- محاسبات شبکه عصبی نیازمند تعداد زیادی داده به ویژه در مرحله آموزش مدل است.

٤ـ١ـ٥. ماشین بردار پشتیبان(SVM)[1]

استفاده از بردارهای پشتیبان خطی در مسائل دسته‌بندی، رویکرد جدیدی است که در چند سال اخیر مورد توجه بسیاری قرار گرفته است. ماشین بردار پشتیبان در ابتدا توسط واپنیک[2] در سال ۱۹۹۰ طراحی شد و بر پایه نظریه آماری یادگیری بنا نهاده شده است. ماشین‌های بردار پشتیبان دارای یکسری خواص هستند که در زیر به تعدادی از آنها اشاره می‌شود:

- طراحی دسته‌بندی‌کننده با حداکثر تعمیم
- رسیدن به بهینه سراسری تابع هزینه
- تعیین خودکار ساختار و توپولوژی بهینه برای طبقه‌بندی‌کننده.

٤ـ١ـ٥ـ١. روش عملکرد ماشین‌های بردار پشتیبان

الگوریتم‌های مبتنی بر ماشین‌های بردار پشتیبان، الگوریتم‌هایی هستند که سعی می‌کنند برای دسته‌بندی داده‌های برچسب یک حاشیه[3] را بیشینه کنند. این الگوریتم‌ها برای پیداکردن خط جداکننده دسته‌ها در مجموعه‌ای از داده‌ها، از دو خط موازی استفاده می‌کنند، ابتدا این خطوط را در خلاف جهت یکدیگر حرکت می‌دهند تا هر کدام از خطوط به یک نمونه از یک دسته خاص در هر سمت خود برسد. پس از انجام این مرحله، میان دو خط موازی یک نوار یا حاشیه

1- Support Vector Machine
2- Vapink
3- Margin

شکل می‌گیرد. هر چه پهنای این نوار بیشتر باشد، به این معناست که الگوریتم توانسته، حاشیه را بیشینه کند و هدف، بیشینه‌کردن این حاشیه است. به عبارت دیگر هر چه پهنای این نوار بیشتر باشد، فاصله بین دسته‌ها بیشتر می‌شود و گروه‌ها به راحتی تفکیک‌پذیر می‌شوند.

٤ـ١ـ٥ـ٢. مقایسه برخی جنبه‌های علمی SVM و شبکه عصبی

- در مرحله آموزش SVM همیشه یک کمینه سراسری یافت می‌شود. ویژگی‌های هندسی مربوط به ماشین‌های بردار پشتیبانی، امکان بررسی بهتر فضای جواب را فراهم می‌سازند.
- ماشین‌های بردار پشتیبان برخلاف شبکه عصبی، به شکل خودکار اندازه مدل را انتخاب می‌کنند.
- برخلاف روش‌های آماری و شبکه عصبی در SVM تلاش نمی‌شود تا پیچیدگی مدل با توجه به تعداد خصیصه‌ها کنترل می‌شود.
- SVM برخلاف شبکه عصبی دارای شالوده نظری منسجمی است.

٤ـ٢. معرفی برخی از روش‌های ارزیابی الگوریتم‌های دسته‌بندی‌ها

میزان صحت یک روش دسته‌بندی بر روی یک مجموعه داده‌های آموزشی (Train) و آزمون، درصد مشاهداتی است که مدل به درستی آنها را دسته‌بندی کرده است که آن را «نرخ تشخیص»[1] می‌نامند. نرخ تشخیص به طور معمول با استفاده از داده‌های آزمون محاسبه می‌شود که آن را با ACC[2] نشان می‌دهند. مهم‌ترین معیار برای تعیین کارایی یک الگوریتم یا مدل دسته‌بندی است.

ماتریس اغتشاشی[3] ابزاری مفید برای تحلیل چگونگی عملکرد روش دسته‌بندی در صحت تشخیص داده‌ها یا مشاهدات دسته‌های مختلف است. این ماتریس چگونگی عملکرد الگوریتم دسته‌بندی را با توجه به مجموعه داده ورودی به تفکیک انواع دسته‌های مسئله دسته‌بندی نشان می‌دهد. برای درک بهتر به توصیف یکسری مفاهیم در قالب یک ماتریس اغتشاشی فرضی می‌پردازیم، فرض می‌کنیم یک متغیر برچسب با دو وجه داریم که به دو کلاس C_2 و C_1 قابل تقسیم هستند که در آن:

TN: عنصر «منفی درست»[4] به مشاهداتی از دسته C_2 دلالت دارد که توسط روش دسته‌بندی به درستی تشخیص داده شده است.

1- Diagnosis Rate
2- Accuracy
3- Confusion Matrix
4- True Negative

FP: «مثبت غلط»[1] مشاهداتی از دسته C_2 است که به نادرستی در دسته C_1 قرار گرفته‌اند.
FN: مقدار «منفی غلط»[2] مشاهداتی از دسته C_1 است که توسط روش دسته‌بندی به نادرستی در دسته C_2 قرار گرفته است.
TP: عنصر «مثبت درست»[3] به مشاهداتی از دسته C_1 که توسط روش دسته‌بندی به درستی تشخیص داده شده است، اطلاق می‌شود.

جدول (۸) ماتریس اغتشاشی فرضی یک دسته‌بندی

دسته	C1	C2
C1	TP	FN
C2	FP	TN

از آنجا که مهم‌ترین معیار برای تعیین کارایی یک الگوریتم یا مدل دسته‌بندی، شاخص صحت یا ACC است ابتدا به روش محاسبه این شاخص اشاره می‌شود. در این مقاله از شاخص صحت یا همان نرخ تشخیص برای مقایسه و ارزیابی مدل‌ها استفاده شده است.

حساسیت[4] $= \frac{TP}{POS[5]}$

شفافیت[6] $= \frac{TN}{Neg[7]}$

دقت[8] $= \frac{TP}{TP + FP}$

حساسیت[9] $= \frac{TP}{POS[10]}$

صحت[11] $= \text{حساسیت} \times \frac{Pos}{POS+Neg} + \text{شفافیت} \times \frac{Neg}{Pos+Neg}$

1- True Negative
2- False Negative
3- True Positive
4- Sensitivity
۵ـ کل تعداد داده‌های مثبت
6- Specificity
۷ـ کل تعداد داده‌های منفی
8- Precision
9- Sensitivity
۱۰ـ کل تعداد داده‌های مثبت
11- Accuracy

در ادامه به معرفی دو شاخص دیگر برای ارزیابی و گزینش دقیق‌تر مدل یا مدل‌های نهایی از میان مدل‌های برتر می‌پردازیم. در مسائل دو دسته‌ای (با متغیرهای خروجی دو وجهی)، شاخص‌های ارزیابی DR و FAR اهمیت ویژه‌ای دارند. این شاخص‌ها توانایی الگوریتم یا مدل را در تشخیص دسته مثبت و نیز تاوان این توانایی تشخیص را تبیین می‌کنند. شاخص DR، نشان می‌دهد که دقت تشخیص دسته مثبت چه اندازه است و شاخص FAR بیانگر نرخ هشدار غلط با توجه به دسته منفی است. به بیان دیگر معیار FAR نشان می‌دهد چه نسبتی از رکوردهای منفی به اشتباه مثبت تشخیص داده شدند. روش محاسبه دو شاخص فوق به شرح زیر است:

$$FAR = \frac{FP}{TN+FP} \qquad DR = \frac{TP}{FN+TP}$$

۵. روش پژوهش

پژوهش حاضر طی مراحل زیر انجام شده است:

۱ـ انتخاب و بررسی داده‌های پژوهش، ۲ـ آماده‌سازی داده‌ها، ۳ـ برازش مدل‌ها

۵ـ۱. انتخاب و بررسی داده‌های پژوهش

در این پژوهش ابتدا با تکیه بر پژوهش‌های انجام شده در سیستم بانکی (داخلی و خارجی) و با استفاده ازسیستم‌های انتخاب و گزینش شاخص مؤثر در رفتار اعتباری متقاضیان تسهیلات اعتباری شامل 5C[1]، LAPP[2] و 5P[3] با نظرکارشناسان مجرب بانکی، ۳۵ متغیر مؤثر بر رفتار اعتباری گیرندگان حقیقی تسهیلات خرد تا سطح یک میلیارد ریال در سطح استان تهران شناسایی شد، سپس با استفاده از روش نمونه‌گیری احتمالی‌ـ تصادفی طبقه‌بندی‌شده از ۱۳،۰۲۶ پرونده تسهیلاتی موجود تا پایان سال ۱۳۸۹ با استفاده از روش کوکران ۴۲۹ نمونه با فاصله اطمینان ۹۵ درصد انتخاب شد. شاخص‌های مورد استفاده برای نمونه‌گیری طبقه‌بندی‌شده بر حسب اهداف پژوهش به ترتیب جاری یا غیرجاری‌بودن پرونده تسهیلاتی، شعب، مبلغ مصوب و نوع عقد است که در قالب جدول زیر آورده شده است. مرجع هر یک از متغیرها نیز در این جدول ذکر شده است.[4]

1- Five C_s of Credit
2- Liquidity, Activity, Profitability, Potential
3- Five P_s of Credit
۴ـ شایان توجه است مرجع متغیرهای مورد استفاده به نام ویگانو، اسکرینر و کروک از پژوهش دهقانی و سوری اخذ شده است.

جدول (۱) شاخص‌ها و متغیرهای پژوهش

شاخص	نوع متغیر	منبع
فردی و خانوادگی	سن گیرنده تسهیلات	اکرو گورس (۱۹۸۹)، بویز،هافمن و لو(۱۹۸۹)، لی و چیو (۲۰۰۲)، ویگانو (۱۹۹۳)، تهرانی، محمدی و رحیمی (۱۳۸۸)، جلیلی، خدایی و کنشلو (۱۳۸۹)
	جنسیت گیرنده تسهیلات	لی و چیو (۲۰۰۲)، اسکرینر (۱۹۹۹)، ویگانو (۱۹۹۳)
	سطح تحصیلات گیرنده تسهیلات	بویز، هافمن و لو (۱۹۸۹)، ویگانو (۱۹۹۳)، تهرانی، محمدی و رحیمی (۱۳۸۸)، جلیلی، خدایی و کنشلو (۱۳۸۹)
	وضعیت تاهل	بویز، هافمن و لو (۱۹۸۹)، داین، آرتیس و گویلن (۱۹۹۶)، ویگانو (۱۹۹۳)، تهرانی، محمدی و رحیمی (۱۳۸۸)، جلیلی، خدایی و کنشلو (۱۳۸۹)
	تعداد فرزندان	کروک (۱۹۹۲)، تهرانی، محمدی و رحیمی (۱۳۸۸)
	وضعیت مالکیت خانه توسط گیرنده تسهیلات یا همسر ایشان	اکرو گورس (۱۹۸۹)، وست (۲۰۰۰)، داین، آرتیس و گویلن (۱۹۹۶)، بویز، هافمن و لو (۱۹۸۹)، لی و چیو (۲۰۰۲)، تهرانی، محمدی و رحیمی (۱۳۸۸)
سایر	شغل ضامن	نظر کارشناسان خبره بانکی
	کد شعبه	اسکرینر (۱۹۹۹)
	کد بخش اقتصادی	اسکرینر (۱۹۹۹)، ویگانو (۱۹۹۳)، موسوی، قلی‌پور (۱۳۸۸)
	آیا اسم فرد متقاضی در لیست سیاه مشتریان ثبت شده است؟	نظر کارشناسان خبره بانکی
اطلاعات حساب‌های سپرده‌ای	تعداد حساب‌های سپرده‌ای	ویگانو (۱۹۹۳)
	مدت سپرده‌گذاری (سپرده کوتاه‌مدت)	دیوید وست (۲۰۰۰)
	میانگین گردش حساب اصلی گیرنده تسهیلات شش ماه قبل از دریافت تسهیلات	اکرامی و رهنما (۱۳۸۸)
	نرخ سود تسهیلات دریافتی	ویگانو (۱۹۹۳)
	مبلغ تسهیلات دریافتی	جلیلی، خدایی و کنشلو (۱۳۸۷)، دیوید وست (۲۰۰۰)، ویگانو (۱۹۹۳)، اسکرینر (۱۹۹۹)، عرب‌مازار، رویین‌تن (۱۳۸۵)
	مبلغ قسط	نظر کارشناسان خبره بانکی
	زمان پرداخت تسهیلات توسط بانک	اسکرینر (۱۹۹۹)
	تعداد پرداخت	نظر کارشناسان خبره بانکی
	نوع عقد	البرزی، پورزرندی و خان‌بابایی (۱۳۸۹)
	طول دوره زمانی بازپرداخت	اکرو گورس (۱۹۸۹)، داین، آرتیس و گویلن (۱۹۹۶)، تهرانی، محمدی و رحیمی (۱۳۸۸)، جلیلی، خدایی و کنشلو (۱۳۸۹)
سابقه اعتباری در نظام بانکی	تعداد قرارداد تسهیلات دریافتی	ابدو و پوینتون (۲۰۰۸)، دیوید وست (۲۰۰۰)
	مبلغ تسهیلات دریافتی	
	تعداد قرارداد تسهیلات دریافتی که وارد سرفصل غیرجاری شده‌اند.	
	مبلغ مانده تسهیلات غیرجاری	
	تعداد چک‌های برگشتی	
	مبلغ چک‌های برگشتی	

ادامه جدول (۱) شاخص‌ها و متغیرهای پژوهش

شاخص	نوع متغیر	منبع
امهال	آیا این قرارداد مشمول امهال شده است؟	مقدم آرانی، امین ناصری و قدسی‌پور (۱۳۸۳)
	آیا پس از امهال دوباره وارد سرفصل غیرجاری شده است؟	
	تعداد قراردادهای تسهیلاتی که شامل امهال شده‌اند	
وثیقه یا تضمینات	نوع وثیقه و تضمینات	جلیلی، خدایی و کنشلو (۱۳۸۷)، ویگانو (۱۹۹۳)، تهرانی، محمدی و رحیمی (۱۳۸۸)، موسوی، قلی‌پور (۱۳۸۸)
	ارزش وثیقه و تضمینات	
چک برگشتی	تعداد چک برگشتی قبل از دریافت تسهیلات	اکرامی و رهنما (۱۳۸۸)
	مبلغ چک برگشتی قبل از دریافت تسهیلات	
شغلی	شغل گیرنده تسهیلات	اکرو گورس (۱۹۸۹)، کروک (۱۹۹۲)، تهرانی، محمدی و رحیمی (۱۳۸۸)، جلیلی، خدایی و کنشلو (۱۳۸۹)
	وضعیت اشتغال همسر گیرنده تسهیلات	کشاورز و آیتی (۱۳۸۶)

در ادامه اطلاعات مورد نیاز از طریق مکاتبه با شعب دریافت شد. لازم به یادآوری است که برخی شاخص‌های مؤثر نظیر سابقه کاری، تجربه و غیره به دلیل شرایط ویژه و فقدان اطلاعات لازم در شعب در مجموع شاخص‌های لیست ارسالی وارد نشدند. در این پژوهش متغیر مستغل، شاخص‌های ارسالی مندرج در جدول شماره (۱) و متغیر وابسته دو وجهی، مشتریان خوش‌حساب و بدحساب بانک هستند.

۵ـ۲. آماده‌سازی داده‌ها

حدود۶۰ تا ۹۰ درصد زمان پژوهش، صرف آماده‌سازی داده‌ها می‌شود و ۷۵ تا ۹۰ درصد موفقیت پژوهش به انجام درست این بخش وابسته است. پردازش اولیه بر روی داده‌های پژوهشی به منظور بهبود کیفیت داده‌ها، آماده‌سازی داده‌ها نامیده می‌شود که به واسطه آن تلاش می‌شود تا داده‌ها عاری از داده‌های ناقص[۱]، مغشوش[۲] و ناسازگار[۳] باشد، این فرآیند طی مراحل زیر انجام شده است:

۱ـ پرونده‌ها و متغیرهایی که با فقدان یا کمبود داده‌ها مواجه‌اند.

۲ـ شاخص‌ها و متغیرهایی که با داده‌های ناقص یا مغشوش مواجه‌اند.

۳ـ تکمیل داده‌ها با استفاده از مدل‌های پیش‌بینی، برخورد با داده‌های پرت.

۱ـ Incomplete: تعداد نمونه‌های ناکافی، کمبود برخی مقادیر شاخص‌ها

۲ـ Noisy: داده‌های دارای مقادیرخطا و نامفهوم که مانع از دسترسی به اصل داده‌ها می‌شود.

۳ـ Inconsistent: داده‌های دارای تناقض

۵ـ۲ـ۱. پرونده‌ها و متغیرهایی که با فقدان یا کمبود داده‌ها مواجه‌اند

در مرحله اول برای بهبود بخشیدن به وضعیت داده‌ها، نگارندگان به شناسایی و حذف متغیرهایی که با فقدان شدید و کمبود داده مواجه هستند، به شرح جدول زیر اقدام کرده‌اند. داده‌های موجود برای متغیرهای زیر حدود ۲۰ درصد بوده است. در این مرحله اطلاعات مربوط به ۲۰ پرونده اعتباری به دلیل نقص شدید و فقدان داده های مربوط به بیش از ۲۴ متغیر از جدول داده‌ها حذف شدند.

جدول (۲) شاخص‌هایی که با فقدان یا کمبود داده مواجه‌اند

شاخص	نوع متغیر
شغلی	وضعیت اشتغال همسر گیرنده تسهیلات
سایر	شغل ضامن
سابقه اعتباری در نظام بانکی	تعداد قرارداد تسهیلات دریافتی
	مبلغ تسهیلات دریافتی
	تعداد قرارداد تسهیلات دریافتی که وارد سرفصل غیرجاری شده‌اند
	مبلغ مانده تسهیلات غیرجاری
	تعداد چک‌های برگشتی
	مبلغ چک‌های برگشتی

۵ـ۲ـ۲. شاخص‌ها و متغیرهایی که با داده‌های ناقص یا مغشوش مواجه‌اند

در این بخش شاخص‌ها و متغیرهای زیر به دلیل ناقص یا مغشوش‌بودن از فرآیند محاسبه کنار گذاشته شدند.

جدول (۳) شاخص‌های دارای داده ناقص یا مغشوش

شاخص	نوع متغیر
امهال	آیا این قرارداد مشمول امهال شده است؟
	آیا پس از امهال دوباره وارد سرفصل غیر جاری شده است؟
	تعداد قراردادهای تسهیلاتی که شامل امهال شده‌اند.
سایر	آیا تاکنون اسم فرد متقاضی درلیست سیاه مشتریان ثبت شده است؟
	مبلغ چک برگشتی

۵ـ۲ـ۳. تکمیل داده‌ها با استفاده از مدل‌های پیش‌بینی و برخورد با داده‌های پرت

۵ـ۲ـ۳ـ۱. پس از پاکسازی داده‌ها[1] از داده‌های مغشوش و بسیار ناقص، وضعیت شاخص‌های باقیمانده به شرح زیر مورد بررسی نهایی قرار گرفت و کمبود داده‌ها با استفاده از الگوریتم تکمیل شد.

جدول (٤) بررسی وضعیت داده‌ها

ردیف	نام فیلد	داده پرت[3]	درصد تکمیل	تعداد رکوردهای تکمیل شده[2]
۱	میانگین گردش، شش ماه قبل از دریافت تسهیلات		۴۱/۵۶	۱۷۰
۲	تعداد چک برگشتی	۲	۵۳/۷۹	۲۲۰
۳	تعداد فرزندان	۵	۷۵/۰۶	۳۰۷
۴	سطح تحصیلات		۷۵/۳	۳۰۸
۵	وضعیت مالکیت خانه		۷۷/۵	۳۱۷
۶	مدت سپرده‌گذاری (سپرده کوتاه‌مدت)		۸۰/۹۲	۳۳۱
۷	جنسیت		۹۱/۱۹	۳۷۳
۸	وضعیت تاهل		۸۴/۱	۳۴۴
۹	سن	۱	۸۹/۴۸	۳۶۶
۱۰	شغل		۹۲/۴۲	۳۷۸
۱۱	ارزش وثیقه	۱	۹۱/۱۹	۳۷۳
۱۲	نوع وثیقه		۹۵/۵۹	۳۹۱
۱۳	تعداد حساب سپرده	۴	۹۴/۸۶۶	۳۸۸
۱۴	کد شعبه		۱۰۰	۴۰۹
۱۵	نوع عقد		۱۰۰	۴۰۹
۱۶	مبلغ تسهیلات		۱۰۰	۴۰۹
۱۷	بخش اقتصادی		۱۰۰	۴۰۹
۱۸	زمان پرداخت تسهیلات توسط بانک		۱۰۰	۴۰۹
۱۹	مبلغ قسط		۱۰۰	۴۰۹
۲۰	نرخ سود		۱۰۰	۴۰۹
۲۱	تعداد پرداخت		۱۰۰	۴۰۹
۲۲	مدت (دوره زمانی بازپرداخت)		۱۰۰	۴۰۹

داده‌های ناقص به شرح جدول زیر تکمیل شدند.

۱ـ Data Cleansing: عملیاتی که به برطرف‌شدن مشکل پایین‌بودن کیفیت داده‌ها می‌انجامد.

۲ـ عبارت است از تعداد فیلد یا شاخص تکمیل شده از ۴۰۹ پرونده‌ای که اطلاعات آن از شعب درخواست شده است.

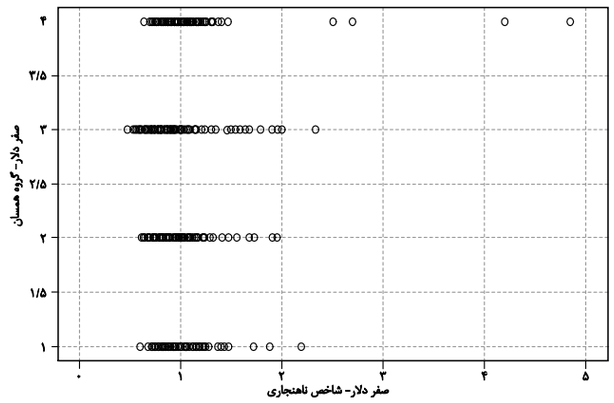
3- Outlier

جدول (۵) الگوریتم‌های استفاده شده برای تکمیل داده‌ها

ردیف	نام متغیر	نحوه تکمیل
۱	جنسیت، وضعیت تاهل، شغل، سطح تحصیلات، تعداد فرزندان، سن و مدت سپرده‌گذاری (سپرده کوتاه‌مدت)	رندم[1]
۲	وضعیت مالکیت	کارت
۳	میانگین گردش، شش ماه قبل از دریافت تسهیلات، تعداد چک برگشتی، تعداد حساب سپرده، ارزش و نوع وثیقه	شبکه عصبی

۵-۲-۳-۲. داده پرت

همان‌گونه که در جدول شماره (۴) نشان داده شد، تعدادی از پرونده‌ها با داده‌های پرت مواجه هستند. برای حل این مشکل از یک مدل با قابلیت کشف مغایرت در اطلاعات گروه‌های غالب[2] به شکل سیستمی استفاده می‌شود. این مدل ابتدا پرونده‌ها را به چند گروه دسته‌بندی کرده، سپس پرونده‌هایی را که در هیچ یک از گروه‌ها جای نمی‌گیرند به عنوان داده پرت معرفی می‌کند.



نمودار (۱) بررسی وضعیت رکوردها (پرونده‌ها)

در نمودار (۱) شاخص نابهنجاری در سطر و گروه‌های چهارگانه (نقاط به هم پیوسته) در ستون نمایش داده شده است. برحسب شاخص نابهنجاری آن گروه از نقاطی را که بیشترین فاصله (شاخص نابهنجاری بزرگ‌تر) را از یکی ازگروه‌ها دارند، از مجموع نقاط حذف می‌کنیم.

1- Random
2- Peer Group

در این میان چهار نقطه فاصله محسوسی با گروه‌های چهارگانه دارند و در نتیجه آن از مدل حذف شده‌اند. شاخص نابهنجاری در اینجا بزرگ‌تر از ۲/۳۸ در نظر گرفته شده است.

۵ـ۳. برازش مدل

۵ـ۳ـ۱. پالایش متغیرهای کم‌اثر

در این بخش داده‌های تکمیل‌شده در مرحله قبل با استفاده از الگوریتم‌های Feature Selection مورد ارزیابی قرار گرفته‌اند. این الگوریتم‌ها برای غربال داده‌ها در دو سطح الف- بر حسب درصد مقادیر گمشده، پراکندگی پایین متغیرها و غیره، ب- آزمون‌های همبستگی با ضریب همبستگی کرامر[1]، پیرسون[2] و غیره اقدام به ارزش‌گذاری داده‌ها می‌کنند.

چکیده نتایج حاصل، در قالب جدول زیر نشان داده شده است. در این جدول شاخص‌ها بر حسب ارزش[3] به سه گروه مهم، حاشیه‌ای و غیرمهم تقسیم شده‌اند. به شکل معمول متغیرهای مهم و حاشیه‌ای حفظ و متغیرهای غیرمهم از فرآیند پژوهش حذف می‌شوند. ولی از آنجا که آزمون‌های همبستگی معیار کاملی برای ارزیابی متغیرها تلقی نمی‌شوند، بنابراین متغیرهای با ارزش بیش از ۰/۸۵۰ در فرآیند پژوهش باقی ماندند. در اینجا بنا بر نظر مدیریتی متغیر سن گیرنده تسهیلات درفرآیند آماده‌سازی ماندگار شد.

جدول (٦) درجه اهمیت متغیرها

رتبه	نام متغیر	درجه اهمیت	ارزش
۱	کد بخش اقتصادی	مهم	۱
۲	نوع عقد	مهم	۱
۳	تعداد پرداخت‌ها	مهم	۱
۴	مدت (دوره زمانی بازپرداخت)	مهم	۱
۵	میانگین گردش، شش ماه قبل از دریافت تسهیلات	مهم	۱
۶	کد شعبه	مهم	۱
۷	نوع وثیقه	مهم	۱
۸	مبلغ قسط	مهم	۱
۹	نرخ سود تسهیلات	مهم	۱
۱۰	تعداد چک برگشتی	مهم	۱

1- Cramer
2- Pearson
3- Value

ادامه جدول (٦) درجه اهمیت متغیرها

رتبه	نام متغیر	درجه اهمیت	ارزش
۱۱	مدت سپرده‌گذاری (سپرده کوتاه‌مدت)	مهم	۰/۹۹۸
۱۲	سطح تحصیلات	مهم	۰/۹۸۷
۱۳	تعداد فرزندان	حاشیه‌ای	۰/۹۴۶
۱۴	شغل	حاشیه‌ای	۰/۹۴۱
۱۵	جنسیت	غیرمهم	۰/۸۸۷
۱۶	وضعیت مالکیت خانه	غیرمهم	۰/۸۷۲
۱۷	ارزش وثیقه	غیرمهم	۰/۸۳۶
۱۸	تعداد حساب سپرده	غیرمهم	۰/۷۹۵
۱۹	سن	غیرمهم	۰/۲۸۳
۲۰	مبلغ تسهیلات	غیرمهم	۰/۲۳

در این میان متغیرها یا فیلدهای مطروحه به دلایل زیر حذف شدند.

جدول (۷) متغیرهای حذف‌شده

ردیف	نام متغیر	دلیل حذف
۱	وضعیت تاهل	یکی از طبقات وجه غالب[1] است
۲	زمان پرداخت تسهیلات توسط بانک	ضریب تغییرات کم
۳	ارزش وثیقه	معنادار نبودن ارزش
۴	تعداد حساب سپرده	معنادارنبودن ارزش
۵	مبلغ تسهیلات	معنادارنبودن ارزش

۵ـ۳ـ۲. شناسایی شاخص‌های مهم و مدل‌های برتر

در این مرحله، به برازش داده‌ها با هدف شناسایی مدل‌های برتر و شناسایی شاخص‌های غیرمؤثر اقدام شد. همچنین به منظور داوری دقیق‌تر درباره داده‌ها، آنها را به دو گروه Train[2] و Test[3] به نسبت ۶۵ و ۳۵ تقسیم نموده سپس با پنج Random Seed اقدام به حل هشت مدل CHAID، CART، C5، QUEST، رگرسیون لجستیک، شبکه عصبی، SVM و تحلیل تشخیص کرده، حاصل برآورد که نشانگر صحت پیش‌بینی رفتار اعتباری مشتریان است، در قالب جدول زیر نشان داده شده است.

۱ـ تعداد متأهلان بسیار بیش از مجردها می‌باشد.

۲ـ آموزشی

۳ـ آزمون

جدول (۹) ارزیابی صحت روش‌های دسته‌بندی

نوع درخت	صحت[1]		اندازه تغییر[2]	نوع درخت	صحت		اندازه تغییر
	Train	Test			Train	Test	
C51	۹۰/۴۹	۹۴/۳۷	۳/۸۸	CART1	۹۰/۴۹	۹۵/۰۷	۴/۵۸
C52	۸۸/۹۷	۸۸/۰۳	۰/۹۴	CART2	۹۱/۲۵	۸۴/۵۱	۶/۷۴
C53	۹۱/۶۳	۹۲/۲۵	۰/۶۲	CART3	۸۷/۴۵	۹۱/۵۵	۴/۱
C54	۸۹/۷۳	۹۳/۶۶	۳/۹۳	CART4	۸۷/۴۵	۸۹/۴۴	۱/۹۹
C55	۸۸/۲۱	۹۰/۱۴	۱/۹۳	CART5	۸۵/۹۳	۸۷/۳۲	۱/۳۹
میانگین	۸۹/۸۰۶	۹۱/۶۹	[3]۲/۲۶	**میانگین**	۸۸/۵۱۴	۸۹/۵۷۸	۳/۷۶
نوع درخت	**صحت**		**اندازه تغییر**	**نوع درخت**	**صحت**		**اندازه تغییر**
	Train	**Test**			**Train**	**Test**	
CHAID1	۸۸/۵۹	۸۸/۷۳	۰/۱۴	QUEST1	۸۸/۵۹	۸۸/۷۳	۰/۱۴
CHAID2	۸۸/۲۱	۸۶/۶۲	۱/۵۹	QUEST2	۸۵/۹۳	۸۵/۲۱	۰/۷۲
CHAID3	۸۸/۹۷	۹۱/۵۵	۲/۵۸	QUEST3	۸۸/۹۷	۹۱/۵۵	۲/۵۸
CHAID4	۸۹/۳۵	۸۹/۳۵	۸۸/۰۳	QUEST4	۸۶/۳۱	۸۹/۴۴	۳/۱۳
CHAID5	۸۶/۶۹	۸۶/۶۹	۸۸/۷۳	QUEST5	۸۸/۵۹	۸۹/۴۴	۰/۸۵
میانگین	۸۸/۳۶۲	۸۸/۳۶۲	۸۸/۷۳۲	**میانگین**	۸۷/۶۷۸	۸۸/۸۷۴	۱/۴۸۴
نوع درخت	**صحت**		**اندازه تغییر**	**نوع درخت**	**صحت**		**اندازه تغییر**
	Train	**Test**			**Train**	**Test**	
D1	۸۵/۲۷	۸۲/۳۱	۲/۹۶	S1	۹۴/۰۹	۹۱/۳۹	۲/۷
D2	۸۲/۵۶	۸۳/۶۷	۱/۱۱	S2	۹۲/۶۴	۹۳/۲	۰/۵۶
D3	۸۲/۹۵	۸۳/۶۷	۰/۷۲	S3	۹۳/۸	۹۴/۵۶	۰/۷۶
D4	۸۱/۴	۸۳/۶۷	۲/۲۷	S4	۹۳/۸	۹۲/۵۲	۱/۲۸
D5	۸۸/۵۸	۸۰/۱۳	۸/۴۵	S5	۹۳/۰۲	۹۳/۲	۰/۱۸
میانگین	۸۴/۱۵۲	۸۲/۶۹	۳/۱	**میانگین**	۹۳/۴۷	۹۲/۹۷۴	۱/۱
شبکه عصبی	**صحت**		**اندازه تغییر**	**لجستیک**	**صحت**		**اندازه تغییر**
	Train	**Test**			**Train**	**Test**	
N1	۹۰/۳۱	۹۲/۵۲	۲/۲۱	L1	۹۰/۷	۸۴/۳۵	۶/۳۵
N2	۸۷/۶	۸۹/۸	۲/۲	L2	۸۸/۳۷	۸۷/۷۶	۰/۶۱
N3	۹۱/۸۶	۸۹/۸	۲/۰۶	L3	۸۸/۳۷	۸۷/۷۶	۰/۶۱
N4	۹۱/۸۶	۹۱/۱۶	۰/۷	L4	۸۸/۷۶	۸۸/۴۴	۰/۳۲
N5	۸۷/۲۱	۹۱/۱۶	۳/۹۵	L5	۸۸/۷۶	۸۷/۰۷	۱/۶۹
میانگین	۸۹/۷۶	۹۰/۸۸	۲/۲۲	**میانگین**	۸۸/۹۹	۸۷/۰۸	۱/۹۲

۱ـ درصد تحقق پیش‌بینی در مقایسه با آنچه در دنیای واقعی رخ داده است.

۲ـ اندازه تکرارپذیری صحت در Test و Train را نشان می‌دهد.

۳ـ میانگین اندازه تغییر، از طریق سنجش درجه تمرکز اندازه تغییر با استفاده از فرمول میانگین هندسی محاسبه شد.

در ادامه بر اساس بالاترین میانگین صحت مدل‌های برتر تعیین و رتبه‌بندی شده است.

جدول (۱۰) رتبه بندی مدل‌ها برحسب شاخص صحت

رتبه	نام مدل	میانگین صحت		اندازه تغییر
		Train	Test	
۱	SVM	۹۳/۴۷	۹۲/۹۷۴	۱/۱
۲	درخت تصمیم C5	۸۹/۸۰۶	۹۱/۶۹	۲/۲۶
۳	شبکه عصبی	۸۹/۷۶	۹۰/۸۸	۰/۰۲
۴	درخت CHAID	۸۸/۳۶۲	۸۸/۷۳	۱/۵۳۴

نکته: به دلیل اندازه تغییر کمتر، درخت CHAID بر درخت تصمیم CART ترجیح داده شد. ضمن تأکید بر ویژگی درخت تصمیم CHAID که قابلیت تقسیم درخت به بیش از دو شاخه را دارد. از دیگر نتایج برآورد مدل‌های فوق؛ تعیین درجه اهمیت شاخص‌های مورد بررسی است که در قالب جدول زیر برحسب میانگین درجه اهمیت، رتبه‌بندی شده‌اند.

جدول (۱۱) رتبه‌بندی میانگین درجه اهمیت شاخص‌ها

رتبه	شاخص	میانگین اهمیت
۱	نوع عقد	۰/۲۰۴
۲	تعداد پرداخت‌ها	۰/۱۸۹
۳	میانگین گردش شش ماه قبل از دریافت تسهیلات	۰/۱۱۲
۴	مدت (دوره زمانی بازپرداخت)	۰/۱۰۳
۵	کد بخش اقتصادی	۰/۰۹۸
۶	نرخ سود	۰/۰۸۹
۷	تعداد چک برگشتی قبل از دریافت تسهیلات	۰/۰۸۱
۸	کد شعبه	۰/۰۷۶
۹	جنسیت گیرنده تسهیلات	۰/۰۶۶
۱۰	مدت سپرده‌گذاری (سپرده کوتاه‌مدت)	۰/۰۶۵
۱۱	نوع وثیقه	۰/۰۵۳
۱۲	وضعیت مالکیت	۰/۰۴۸
۱۳	شغل گیرنده تسهیلات	۰/۰۳۳
۱۴	تعداد فرزندان	۰/۰۳۲
۱۵	سطح تحصیلات گیرنده تسهیلات	۰/۰۳۲
۱۶	مبلغ قسط	۰/۰۳
۱۷	سن گیرنده تسهیلات	۰/۰۲۵

مطابق جدول شماره (۱۱)، شاخص‌هایی از قبیل سن، سطح تحصیلات و تعداد فرزندان گیرنده تسهیلات به دلیل میانگین اهمیت پایین از مدل حذف شده‌اند. در این میان شاخص‌های کد شعبه و نرخ سود به ترتیب به دلیل:

- عدم قابلیت تسری (تعداد پرونده به شعبه)
- ناتوانی در کاهش نرخ سود دست‌کم در کوتاه‌مدت

از مدل حذف و شاخص مبلغ قسط و شغل گیرنده تسهیلات با نظر کارشناسی و مدیریتی در مدل باقی ماند و در نهایت مدل با ۱۳ شاخص مورد بررسی قرار گرفت.

۵ـ۳ـ۳. محاسبه و برآورد مدل نهایی

در این بخش با استفاده از ۱۳ شاخص نهایی شده، مدل‌های CHAID، C5، شبکه عصبی و SVM به دو گروه Train و Test به نسبت ۶۵ و ۳۵ تقسیم شده و با استفاده از پنج Random Seed برآورد شد. در جدول زیر نتایج برآورد، به همراه شاخص‌ها و معیارهای ارزیابی نمایش داده شده است.

جدول (۱۲) ارزیابی و رتبه‌بندی الگوریتم‌های دسته‌بندی

نوع درخت	اندازه تغییر	Test			Train		
		Far	DR	صحت	Far[۲]	DR[۱]	صحت
C51	۳/۱۱	۲۲/۴۵	۹۳/۶۷	۸۷/۵	۲۰	۹۴/۹۲	۹۰/۶۱
C52	۲/۹۶	۲۴/۳۹	۹۲/۷۸	۸۷/۶۸	۱۹/۳۲	۹۵/۵۳	۹۰/۶۴
C53	۵/۵۱	۱۰/۸۷	۹۵	۹۳/۱۵	۲۶/۵۱	۹۴/۳۲	۸۷/۶۴
C54	۴/۷۴	۲۰	۹۷/۹۶	۹۲/۷۵	۲۱/۳۵	۹۲/۷	۸۸/۰۱
C55	۲/۶	۲۲/۷۳	۹۳/۱۸	۸۷/۸۸	۲۰	۹۵/۲۱	۹۰/۴۸
میانگین	۳/۷۸	۲۰/۰۹	۹۴/۵۲	۸۹/۷۹	۲۱/۴۳	۹۴/۵۴	۸۹/۴۸
نوع درخت	**اندازه تغییر**	**Test**			**Train**		
		Far	**DR**	**صحت**	**Far**	**DR**	**صحت**
Chi1	۶/۸۴	۲۴/۴۹	۸۴/۸۱	۸۱/۲۵	۱۳/۷۵	۸۸/۸۳	۸۸/۰۹
Chi2	۲/۸۳	۴۱/۴۶	۹۴/۸۵	۸۴/۰۶	۳۱/۸۲	۹۶/۰۹	۸۶/۸۹
Chi3	۱/۱۵	۱۰/۸۷	۸۳	۸۴/۹۳	۱۳/۲۵	۸۲/۳۹	۸۳/۷۸
Chi4	۲/۳۲	۲۰	۸۹/۸	۸۶/۹۶	۱۷/۹۸	۸۵/۹۶	۸۴/۶۴
Chi5	۸/۷۳	۳۶/۳۶	۸۷/۵	۷۹/۵۵	۲۱/۱۸	۹۲/۵۵	۸۸/۲۸
میانگین	۴/۳۷	۲۶/۶۴	۸۷/۹۹	۸۳/۳۵	۱۹/۶	۸۹/۱۶	۸۶/۳۴

۱ـ نسبت رکوردهای دسته خوش‌حساب‌ها که به درستی در دسته خوش‌حساب‌ها قرار گرفته‌اند به همه رکوردهایی که خوش‌حساب هستند.
۲ـ نسبت رکوردهای دسته بدحساب‌ها که نادرست در دسته خوش‌حساب‌ها قرار گرفته‌اند به همه رکوردهایی که بدحساب هستند.

ادامه جدول (۱۲) ارزیابی و رتبه‌بندی الگوریتم‌های دسته‌بندی

شبکه عصبی	اندازه تغییر	Test			Train		
		Far	DR	صحت	Far	DR	صحت
N1	۰/۵۹	۶/۱۲	۹۱/۱۴	۹۲/۱۹	۳/۷۵	۹۱/۳۷	۹۲/۷۸
N2	۱/۳۵	۴/۸۸	۹۲/۷۸	۹۳/۴۸	۴/۵۵	۹۰/۵	۹۲/۱۳
N3	۳/۰۱	۲/۱۷	۹۳	۹۴/۵۲	۶/۰۲	۹۰/۳۴	۹۱/۵۱
N4	۲/۴۴	۵	۹۳/۸۸	۹۴/۲	۴/۴۹	۸۹/۸۹	۹۱/۷۶
N5	۲	۲/۲۷	۹۲/۰۵	۹۳/۹۴	۵/۸۸	۹۰/۹۶	۹۱/۹۴
میانگین	۱/۸۸	۴/۰۹	۹۲/۵۷	۹۳/۶۷	۴/۹۴	۹۰/۶۱	۹۲/۰۲
SVM	**اندازه تغییر**	**Test**			**Train**		
		Far	**DR**	**صحت**	**Far**	**DR**	**صحت**
SVM1	۷/۲۶	۲۸/۵۷	۸۶/۰۸	۸۰/۴۷	۲۱/۲۵	۹۱/۳۷	۸۷/۷۳
SVM2	۳/۲۹	۱۶/۳۳	۸۶/۰۸	۸۵/۱۶	۱۶/۲۵	۹۰/۳۶	۸۸/۴۵
SVM3	۳/۱۰	۱۹/۵۷	۸۷	۸۴/۹۳	۱۴/۴۶	۸۹/۲	۸۸/۰۳
SVM4	۰/۴۴	۲۷/۵	۹۲/۸۶	۸۶/۹۶	۱۶/۸۵	۸۸/۲	۸۶/۵۲
SVM5	۷/۲۲	۲۹/۵۵	۸۶/۳۶	۸۱/۰۶	۱۶/۴۷	۹۰/۴۳	۲۲/۲۸
میانگین	۴/۲۶	۲۴/۳	۸۷/۶۷	۸۳/۷۲	۱۷/۰۶	۸۹/۹۱	۸۷/۸

همان‌گونه که مشاهده می‌شود، بر اساس شاخص DR، الگوریتم‌های C5 و شبکه عصبی بیش از ۹۰ درصد داده‌های مربوط به خوش‌حساب‌ها را به درستی در دسته مربوطه قرار داده‌اند، همچنین الگوریتم‌های شبکه عصبی و C5 بر مبنای شاخص Far در داده‌های آزمون[۱] (Test)، به ترتیب ۴/۰۹ و ۲۰/۰۹ درصد از بدحساب‌ها را در دسته خوش‌حساب‌ها قرار داده‌اند. مقدار این شاخص به دلیل پیامدهای ناگوار آن[۲] برای بانک اهمیت ویژه‌ای دارد. در ادامه با در نظرگرفتن دو شاخص Far و DR، همچنین معیار صحت که مهم‌ترین شاخص ارزیابی الگوریتم‌ها است، الگوریتم‌های شبکه عصبی و درخت تصمیم C5 به ترتیب به عنوان الگوریتم‌های کارا در مقایسه با سایر الگوریتم‌ها شناخته می‌شوند. بنابراین پیشنهاد می‌شود، برای کشف قواعد از مدلC5 و برای پیش‌بینی نرم‌افزاری از شبکه عصبی استفاده شود.

۱ـ داده‌های Test از آنجا که در فرآیند یادگیری نقشی ندارند، بنابراین قدرت تطابق بیشتری با داده‌های خارج از نمونه در مقایسه با داده‌های Train دارند.

۲ـ افزایش هزینه پیگیری‌های حقوقی و عدم بازپرداخت اصل و سود تسهیلات دریافتی

۵ـ۳ـ۵. ترکیب مدل

گاهی پژوهشگر در پی استفاده از قابلیت‌های مدل‌های برآورد شده به طور همزمان می‌باشد. در این شرایط می‌توان ترکیب مدل‌های مختلف با روش‌های رأی‌گیری[1]، رأی‌گیری وزنی براساس فاصله اطمینان[2] و غیره با random Seedهای مشخص، برای تولید یک مدل ترکیبی استفاده نمود. در این پژوهش ترکیبات مختلفی از چهار مدل CHAID، C5، شبکه عصبی و SVM با Random Seedهای مشخص مورد آزمون قرارگرفت که در همه ترکیبات شبکه عصبی به شکل محسوسی نسبت به ترکیب به دست‌آمده برتری دارد.

٦. یافته‌های پژوهش

٦ـ۱. شاخص‌های مهم

با استفاده از نتایج حاصل از حل دو مدل درخت تصمیم (C5 وCHAID)، شبکه عصبی و SVM، دوازده متغیر مهم به همراه میانگین اندازه اهمیت در قالب جدول شماره (۱۳) نشان داده شده است.

جدول (۱۳) میانگین اهمیت متغیرهای مهم مؤثر در رفتار اعتباری متقاضیان تسهیلات

ردیف	متغیرها	میانگین اهمیت متغیر بر حسب هر الگوریتم[3]				میانگین اهمیت کل	میانگین اهمیت کل نرمال‌شده[4] (درصد)	تعداد تکرار در هر الگوریتم
		درخت تصمیم		SVM	شبکه عصبی			
		C5	CHILD					
۱	مدت بازپرداخت	۰/۴۱	۰/۴۹	۰/۱	۰/۱۳	۰/۲۸	۲۲/۵۸	۴
۲	نوع عقد	۰/۳۹	۰/۳۳	۰/۲۱	۰/۰۸	۰/۲۵	۲۰/۱۶	۴
۳	تعداد پرداخت‌ها	۰/۰۶	۰/۳۹	۰/۱۱	۰/۱۳	۰/۱۷	۱۳/۷۱	۴
۴	نوع وثیقه		۰/۱۲	۰/۱۹	۰/۰۸	۰/۱۳	۱۰/۴۸	۳
۵	تعداد چک برگشتی قبل از دریافت تسهیلات	۰/۰۸	۰/۱۵	۰/۰۳	۰/۱۳	۰/۱۰	۸/۰۶	۴
۶	مبلغ قسط		۰/۱۶	۰/۰۳	۰/۰۵	۰/۰۸	۶/۴۵	۳

1- Voting
2- Confidence-Weighted Voting
۳ـ از میانگین‌گیری اهمیت هر الگوریتم برحسب Random Seedهای پنج‌گانه به دست می‌آید.
۴ـ از تقسیم میانگین اهمیت کل هر یک از متغیرها بر مجموع این میانگین به دست می‌آید.

ادامه جدول (۱۳) میانگین اهمیت متغیرهای مهم مؤثر در رفتار اعتباری متقاضیان تسهیلات

ردیف	متغیرها	میانگین اهمیت متغیر بر حسب هر الگوریتم				میانگین اهمیت کل	میانگین اهمیت کل نرمال‌شده (درصد)	تعداد تکرار در هر الگوریتم
		درخت تصمیم		SVM	شبکه عصبی			
		C5	CHILD					
۷	مدت سپرده‌گذاری	۰/۰۶	۰/۰۲	۰/۰۱	۰/۱۳	۰/۰۶	۴/۸۴	۴
۸	کد بخش اقتصادی	صفر	صفر	۰/۱۳	۰/۰۷	۰/۰۵	۴/۰۳	۳
۹	جنسیت گیرنده تسهیلات		۰/۰۵	۰/۰۳	۰/۰۳	۰/۰۴	۳/۲۳	۳
۱۰	شغل گیرنده تسهیلات	صفر	صفر	۰/۰۴	۰/۱	۰/۰۴	۳/۲۳	۳
۱۱	میانگین گردش شش ماه قبل از اعطای تسهیلات	صفر	صفر	۰/۰۷	۰/۰۵	۰/۰۳	۲/۴۲	۳
۱۲	وضعیت مالکیت خانه	صفر	صفر	۰/۰۳	۰/۰۲	۰/۰۱	۰/۸۱	۴

همان‌گونه که مشاهده می‌شود شاخص‌های مربوط به شخصیت گیرنده تسهیلات به استثنای جنسیت، شغل و وضعیت مالکیت هیچ‌گونه اثری بر خوش‌حسابی یا بدحسابی مشتریان ندارد.

نکته: مطابق جدول بالا عوامل مؤثر بر رفتار اعتباری به طور عموم تحت تأثیر عوامل قابل کنترل به وسیله بانک است.

٦ـ۲. در ذیل به تعدادی از قواعد حاصل اشاره می‌شود.

٦ـ۲ـ۱. قواعد رفتاری مشتریان بدحساب

- تعداد نمونه ۸۳

اگر طول دوره بازپرداخت کوچک‌تر از ۸۴ ماه و نوع عقد؛ اجاره به شرط تملیک، مشارکت مدنی یا مضاربه باشد و در صورتی که مشتری چک برگشتی داشته باشد، با احتمال ۸۸ درصد مشتری بدحساب است.

- تعداد نمونه ۵۲

اگر طول دوره بازپرداخت کوچک‌تر از ۸۴ ماه، نوع عقد؛ اجاره به شرط تملیک، عمر حساب سپرده کوتاه‌مدت مشتری کمتر از چهار سال و مشتری دارای چک برگشتی، تضمین دریافتی، سفته و مبلغ قسط مشتری بیش از ۱،۸۲۰،۰۰۰ ریال باشد با احتمال ۹۴ درصد مشتری بدحساب است.

٦ـ۲ـ۲. قواعد رفتاری مشتریان خوش‌حساب

- تعداد نمونه ۸۰

اگر طول دوره بازپرداخت کوچک‌تر از ۸۴ ماه، نوع عقد جعاله، فروش اقساطی یا قرض‌الحسنه و عمر حساب سپرده کوتاه‌مدت مشتری کمتر از شش سال باشد با احتمال ۹۵ درصد مشتری خوش‌حساب است.

- تعداد نمونه ۱۶۶

اگر طول دوره زمانی بازپرداخت بزرگ‌تر از ۸۴ ماه باشد مشتری ۱۰۰ درصد خوش‌حساب است.

- تعداد نمونه ۶۸

اگر طول مدت بازپرداخت کمتر از ۸۴ ماه، عمر حساب سپرده کوتاه‌مدت مشتری کمتر از چهار سال و پرداخت تسهیلات در چند مرحله انجام شده باشد با احتمال ۹۷ درصد مشتری خوش‌حساب است.

- تعداد نمونه ۵۰

اگر کد بخش اقتصادی ۱۰یا ۱۲ و عمر حساب سپرده کوتاه‌مدت مشتری بیش از شش سال باشد به احتمال ۸۲ درصد مشتری خوش‌حساب است.

- تعداد نمونه ۹۸

اگر کد بخش اقتصادی ۱۰ یا ۱۲ و عمر حساب سپرده کوتاه‌مدت مشتری کمتر یا مساوی شش سال باشد به احتمال ۹۸ درصد مشتری خوش‌حساب است.

وفق قواعد حاصل، فاکتور اساسی در دسته‌بندی مشتریان اعتباری، دوره زمانی بازپرداخت است که این امر می‌تواند به دلیل عدم لحاظ دوره زمانی لازم برای درآمدزایی طرح‌های تسهیلاتی باشد. در ضمن از قواعد به دست آمده، می‌توان به نکات زیر پی برد.

۱ـ به دلیل عدم رعایت تناسب بین مبلغ اقساط و درآمد مشتری در اعطای تسهیلات خرد، در این پژوهش (نمونه مورد مطالعه) در صورتی که مبلغ اقساط بیش از ۱۸۲ هزار تومان باشد، به احتمال زیاد مشتری بدحساب خواهد بود.

۲ـ طول دوره سپرده‌گذاری مشتریان با خوش‌حسابی یا بدحسابی مشتریان ارتباط مشخصی ندارد.

۳ـ بین نوع عقد تسهیلاتی و وضعیت اعتباری مشتریان ارتباط مشخص و مستقیمی مشاهده نشده است.

۴- از قواعد فوق می‌توان برای طراحی یک مدل امتیازدهی اعتباری استفاده کرد.
۵- مدل مناسب برای پیش‌بینی رفتار اعتباری متقاضیان تسهیلات شبکه عصبی با میانگین درصد صحت پیش‌بینی ۹۳ درصد است.

۷. نتیجه‌گیری و پیشنهادات

به دلیل اهمیت زیاد کاهش رشد تسهیلات غیرجاری، در مقاله حاضر تلاش شد به پرسش‌های زیر پاسخ داده شود:

بهترین مدل برای ارزیابی احتمال نکول مشتریان چه مدلی است؟

عوامل مؤثر بر احتمال نکول تعهدات مشتریان حقیقی بانک تا چه اندازه‌ای قابل کنترل هستند؟

برای پاسخ به پرسش شماره (۱)، در پژوهش حاضر، پس از نمونه‌گیری طبقه‌بندی‌شده ۴۲۹ پرونده از بیش از ۱۳۰۰۰ پرونده اعتباری انتخاب و درخواست اطلاعات ۳۶ شاخص مورد بررسی به شعب منتخب ارسال شد. پس از آماده‌سازی و پالایش داده‌های دریافتی از شعب با بهره‌گیری از چهار نوع درخت تصمیم (CHAID، CART، QUEST، C5)، شبکه عصبی، ماشین بردار پشتیبان، رگرسیون لجستیک و تحلیل تمایزی اقدام به محاسبه احتمال نکول تعهدات در دو مرحله شد. در مرحله اول چهار مدل تحلیل تمایزی، رگرسیون لجستیک، درخت تصمیم (QUEST و CART) از فرآیند محاسبه حذف شدند. در این مرحله، مدل SVM بهترین صحت را داشته است. در ادامه چهار مدل باقیمانده دو مدل درخت تصمیم (C5 و CHAID)، شبکه عصبی و SVM دوباره حل شد. نتایج این مرحله نشان داد بیشترین صحت مربوط به مدل شبکه عصبی بوده است و برتری بی‌چون و چرایی بر سایر مدل‌ها دارد. در ضمن در فرآیند پژوهش روشن شد، از ۳۵ شاخص مورد بررسی ۱۲ شاخص مدت (دوره زمانی بازپرداخت)، نوع عقد، تعداد اقساط بازپرداخت، نوع وثیقه، تعداد چک برگشتی قبل از دریافت تسهیلات، مبلغ قسط، مدت سپرده‌گذاری (سپرده کوتاه‌مدت)، بخش اقتصادی، جنسیت، شغل، میانگین گردش شش ماه قبل از دریافت تسهیلات و وضعیت مالکیت به عنوان شاخص‌های نهایی به ترتیب بیشترین اهمیت را در توضیح رفتار اعتباری وام‌گیرندگان دارند که تنها سه شاخص شغل، جنسیت و مالکیت توسط بانک قابل کنترل نیستند. این مطلب بیانگر قابل کنترل‌بودن ۹ شاخص نهایی دیگر با مجموع اهمیت بیش از ۹۰ درصد است. در ادامه با توجه به نتایج پژوهش حاضر اقدامات زیر پیشنهاد می‌شود:

- بهره‌گیری از فرآیند اعتبارسنجی در اعطای تسهیلات

- طراحی و ساخت سیستم نرم‌افزاری اعتبارسنجی مبتنی بر مدل‌های شبکه عصبی، C5، CHAILD و SVM
- بهره‌گیری از دست‌کم ۵ تا ۱۵ درصد داده‌ها برای رعایت اصل «تعمیم‌پذیری»
- شناسایی عوامل مؤثر بر رفتار اعتباری متقاضیان حقوقی تسهیلات با درنظرگرفتن سطوح مختلف تسهیلاتی

منابع و مآخذ

شهرابی، جمال و ذوالقدر، علی. (۱۳۹۰). *داده‌کاوی پیشرفته*. تهران: انتشارات جهاد دانشگاهی دانشگاه امیرکبیر.

طالبی، محمد و شیرزادی، نازنین. (۱۳۹۰). *ریسک اعتباری: اندازه‌گیری و مدیریت*. تهران: انتشارات سمت.

قراچورلو، نجف و انجمن آذری، ارسلان. (۱۳۸۷). *مدیریت ریسک*. تبریز: جهاد دانشگاهی واحد استان آذربایجان شرقی.

فلاح شمس، میرفیض و رشنو، مهدی. (۱۳۸۹). *مدیریت ریسک اعتباری در بانک‌ها و مؤسسات مالی و اعتباری، مفاهیم و مدل‌ها*. تهران: دانشکده علوم اقتصادی.

سروش، علیرضا و بحرینی‌نژاد، اردشیر. (۱۳۸۸). *هوشمندی کسب‌وکار و داده‌کاوی*. تهران: انتشارات ناقوس.

علیزاده، سمیه و ملک‌محمدی، سمیرا. (۱۳۹۰). *داده‌کاوی و کشف دانش: گام به گام با نرم‌افزار*. تهران: دانشگاه خواجه‌نصیرالدین طوسی.

شهرابی، جمال و هداوندی، اسماعیل. (۱۳۹۰). *داده‌کاوی در صنعت بانکداری*. تهران: انتشارات جهاد دانشگاهی دانشگاه امیرکبیر.

صفری شال، رضا و پورگتابی، کرم. (۱۳۸۸). *راهنمای جامع کاربرد SPSS در تحقیقات پیمایشی*. تهران: نشر لویه.

کانتاردزیک، مهمد. (۱۳۸۹). *داده‌کاوی*. امیرعلیخانزاده. تهران: نشر علوم رایانه.

علیزاده، سمیه؛ تیمورپور، بابک و غضنفری، مهدی. (۱۳۸۷). *داده‌کاوی و کشف دانش*. تهران: دانشگاه علم و صنعت ایران.

راسل، بیل و جکسون، تام. (۱۳۸۶). *آشنایی با شبکه‌های عصبی*. محمود البرزی. تهران: دانشگاه صنعتی شریف.

صنیعی آباده، محمد؛ محمودی، سینا و طاهرپرور، محدثه. (۱۳۹۱). *داده‌کاوی کاربردی*. تهران: نیاز دانش.

انواری رستمی، علی‌اصغر و فتحی، سعید. (۱۳۸۲)، بررسی تحلیلی- تطبیقی الگوها و مدل‌های سنجش و اندازه‌گیری اعتباری مشتریان، *بررسی‌های حسابداری و حسابرسی*، ش ۳۳.

تهرانی، رضا، محمدی، محسن و رحیمی، امیرمحمد. (۱۳۸۸). *نظام سنجش اعتبار و جایگاه آن در بهبود نظام تأمین مالی*، اولین کنفرانس بین‌المللی توسعه نظام تأمین مالی در ایران

موسوی، رضا و قلی‌پور، الناز. (۱۳۸۸). *رتبه‌بندی معیارهای اعتبارسنجی مشتریان بانکی با رویکرد دلفی*، اولین کنفرانس بین‌المللی بازاریابی خدمات بانکی.

محمدی صداقت، سارا و سپه‌وند، مهرداد. (۱۳۸۸). *رتبه‌بندی اعتباری مشتریان حقیقی حوزه کسب‌وکار با استفاده از مدل شبکه‌های عصبی و مدل لاجیت*، پایان‌نامه کارشناسی ارشد، مؤسسه عالی بانکداری ایران

دهقانی، محمدعلی و سوری، داود. (۱۳۸۸). *مدلی برای رتبه‌بندی اعتباری مشتریان حقیقی در بانک کارآفرین*، پایان‌نامه کارشناسی ارشد، مؤسسه عالی بانکداری ایران

کشاورز حداد، غلامرضا و آیتی گلزار، حسین. (۱۳۸۶). مقایسه کارکرد مدل لاجیت و روش درخت‌های طبقه‌بندی و رگرسیونی در فرآیند اعتبارسنجی متقاضیان حقیقی برای استفاده از تسهیلات بانکی، *فصلنامه پژوهش‌های اقتصادی*، (۷)۴.

جلیلی، محمد، خدایی، محمد و مهدیه کنشلو. (۱۳۸۹). *اعتبارسنجی مشتریان حقیقی در سیستم بانکی کشور، مطالعات کمی در مدیریت*، ش ۳.

عرب مازار، عباس و رویین‌تن، پونه. (۱۳۸۵). عوامل مؤثر بر ریسک اعتباری مشتریان بانکی؛ مطالعه موردی بانک کشاورزی، *فصلنامه جستارهای اقتصادی*، ش ۶.

البرزی، محمود؛ محمدپور زرندی، محمد ابراهیم و خان‌بابایی، محمد. (۱۳۸۹). به کارگیری الگوریتم ژنتیک دربهینه‌سازی درختان تصمیم‌گیری برای اعتبارسنجی مشتریان بانک‌ها، *نشریه مدیریت فناوری اطلاعات*، دوره دوم، ش ۴.

رحمانی، علی و اسماعیلی، غریبه. (۱۳۸۹).کارایی شبکه‌های عصبی، رگرسیون لجستیک و تحلیل تمایزی در پیش‌بینی نکول، *فصلنامه اقتصاد مقداری*، ش ۷.

پرویزیان، کوروش؛ ذکاوت، سیدمرتضی و محمدیان، مهدی. (۱۳۸۸). رتبه‌بندی داخلی مشتریان بانک‌ها با استفاده از مدل‌های رگرسیونی لاجیت، *پژوهشنامه اقتصادی*، ش ۶. (ویژه‌نامه بانک)

مقدم آرانی، عباس؛ امین‌ناصری، محمدرضا و قدسی‌پور، حسن. (۱۳۸۳). *مدل ارزیابی وام‌های بانکی با استفاده از تکنیک AHP*، کنفرانس بین‌المللی مهندسی صنایع