فصلنامه پژوهشهای اقتصادی ـ سال هفتم ـ شماره چهارم ـ زمستان ۱۳۸۶ ـ صفحات ۷۱-۹۷

مقایسهٔ کارکرد مدل لاجیت و روش درختهای طبقه‌بندی و رگرسیونی در فرایند اعتبارسنجی متقاضیان حقیقی برای استفاده از تسهیلات بانکی

دکتر غلامرضا کشاورزحداد[۱]
حسین آیتی گازار[۲]

تاریخ دریافت: ۱۳۸۵/۶/۱۵ تاریخ پذیرش: ۱۳۸۶/۸/۷

چکیده

در صنعت بانکداری امروز وامها نقشی اساسی دارند، به‌طوری که بخش زیادی از دارایی‌های یک بانک از وامهای پرداخت‌شده به افراد و شرکتها تشکیل می‌شود و در نتیجه با توجه به افزایش تعداد درخواستهای وام از سوی افراد و با توجه به ریسک موجود در این رشته از فعالیتها، ارائهٔ روشی برای مدیریت این وامها ضروری به نظر می‌رسد. در بین ریسک‌هایی که بانک با آن مواجه است، ریسک اعتباری از اهمیت ویژه‌ای برخوردار است. یکی از راه‌های کمی‌سازی و اندازه‌گیری ریسک اعتباری و در نتیجه مدیریت مناسب آن، استفاده از مدلهای امتیازدهی اعتباری (CS) است. مدل CS بر اساس معیارهای کمّی(مانند اطلاعات مالی افراد) و نیز معیارهای غیرکمّی(مثل مشخصات مربوط به شخصیت اجتماعی افراد)، ویژگیها و عملکرد وامهای قبلی را مدل‌سازی می‌نماید تا عملکرد آتی وامهایی با مشخصات مشابه را پیش‌بینی می‌کند. در CS یک نمره به هر مشتری اختصاص داده می‌شود که این نمره به‌عنوان شاخصی از ریسک مشتریان شناخته می‌شود. با مقایسه این نمره مقدار آستانه، مشتریان پرریسک و کم‌ریسک از یکدیگر تفکیک می‌گردند. به‌رغم عمومیت استفاده از روش لاجیت در فرایند اعتبارسنجی افراد، در این تحقیق سعی شده است روشی جایگزین که با توجه به وضعیت اطلاعات موجود در خصوص مشتریان حقیقی بانکها در ایران که نسبت به مدل لاجیت برتری داشته باشد، ارائه و با استفاده از یک مجموعه داده، کارآیی و دقت آن در مقابل مدل لاجیت بررسی شود. به‌منظور ارزیابی مشتریان حقیقی بانک، از مدل امتیازدهی لاجیت و روش غیرپارامتریک $CART$ استفاده شده و نتایج نشان می‌دهد که روش دوم از دقت بالاتری در پیش‌بینی مشتریان خوب و بد با یکدیگر برخوردار است. علاوه بر این، ضمن این‌که مدلهای ساخته‌شده برای نمونه‌های تصادفی با حجم کوچک به روش نمونه‌گیری بازگردان نیز مورد ارزیابی قرار گرفتند که بار دیگر در نتایج حاصل از مشاهده نمونه مورد مطالعه، دقت در تشخیص نوع مشتری از نظر خوش‌حسابی تأیید گردید.

---

۱. استادیار گروه اقتصاد دانشگاه صنعتی شریف

email: g.k.haddad@sharif.edu

۲. فارغ‌التحصیل کارشناسی ارشد علوم اقتصادی دانشگاه صنعتی شریف

طبقه‌بندی JEL: G21 ، G32 ، G33

واژگان کلیدی: امتیازدهی اعتباری[1]، ریسک اعتباری[2]، مدیریت ریسک، مهندسی مالی[3] و نمونه‌گیری بازگردان[4].

1. Credit Scoring 2. Credit Risk 3. Financial Engineering
4. Bootstrapping

۱- مقدمه

برای سالیان متمادی بانکها به وامهای شخصی با شک و تردید می‌نگریستند و در صنعت بانکداری، وزن ویژه‌ای برای آن قائل نبودند. وام‌دهی بانکها به وامهای تجاری محدود بود و این وامها در مبالغ قابل توجه به شرکتها و به سرمایه‌گذاران حقوقی پرداخت می‌شد. بالابودن سود حاصل از این نوع وامها و شفاف‌بودن فرصتهای سرمایه‌گذاری از لحاظ ریسک و بازده، از ویژگیهای این نوع وامها به‌شمار می‌رفت. به‌تدریج با ایجاد رقابت در بازارها و به‌وجودآمدن شرکتهای جدید و افزایش ریسک سرمایه‌گذاری در صنعت، سود وامهای تجاری کاهش یافت و به این ترتیب نگرشها به وامهای شخصی دچار تغییر گردید.

در سالهای اخیر رقابت بین بانکها و کاهش نرخ تسهیلات، سوددهی وامهای تجاری را در مقایسه با ریسک بالای آنها کاهش داده و این امر باعث کاهش تعداد وامهای تجاری و روی‌آوردن بانکها به سمت وامهای شخصی شده است. سود حاصل از وامهای شخصی را می‌توان از دو طریق افزایش داد. اول اینکه اگر طی مکانیزمی بتوان تعداد وامها را افزایش و از سوی دیگر هزینه‌های بررسی درخواستها را کاهش داد و با توجه به حجم بالای این نوع وامها می‌توان سود عظیمی را از این مجرا نصیب بانکها کرد. دوم اینکه با توجه به مقدار کم اعتبار تخصیصی، در صورت عدم توانایی مشتری برای بازپرداخت دیون، زیان زیادی متوجه بانک نخواهد شد و از طرف دیگر وثیقه‌های قابل نقدشدن فراوانی را می‌توان پشتوانه این وامها قرار داد.

آنچه ممکن است بانک را در مواجهه با این نوع وامها دچار مشکل کند، تعداد زیاد وامهای(اقساط) پرداخت‌نشده یا پرداخت با تأخیر است که به علت حجم زیاد آنها، ممکن است حتی به ورشکستگی یک بانک منجر شود. به این ترتیب در صورت وجود یک سیستم کارا و هوشمند مدیریت این وامها، می‌توان تعداد وامهای سوخت‌شده را کاهش داد. یکی از مهمترین ابزارهایی که بانکها، برای مدیریت و کنترل ریسک اعتباری بدان نیازمندند، سیستم امتیازدهی اعتباری مشتریان(CS)[1] است. از طرف دیگر ورود مؤسسات اعتباری به عرصهٔ رقابت داخلی و جهانی و رویارویی با حجم گستردهٔ تقاضا برای اعتبار، فرصتها و تهدیدات جدیدی را برای آنها، ایجاد نموده است. لذا شاهد توسعه رو به تزاید نقش تکنولوژی در فرایند مدیریت اعتبار مؤسسات بانکی و نهادهای مالی هستیم(Liu, 2001).

امتیازدهی اعتباری به‌عنوان یک روش ارزیابی اعتبار، بیش از ۵۰ سال است که مورد استفاده قرار گرفته است. اولین دفتر مشاوره که از روشهای امتیازدهی استفاده کرده است، در سانفرانسیسکو

1. credit scoring

توسط بیل‌فر و ارل ایساک[1] در اوایل دهه ۱۹۵۰ شکل گرفت. در حقیقت اولین موفقیت در حوزه امتیازدهی اعتباری، در حوزه کارتهای اعتباری رخ داده است(Fujita & Tamai, 1987). پس از آن بانکها استفاده از امتیازدهی را برای سایر محصولات نظیر وام شخصی، وام خودرو و وام مسکن آغاز کردند و در حال حاضر این سیستمها برای ارزیابی وامهای تجاری کوچک هم مورد استفاده قرار گرفته است. درخواست برای ایجاد سیستمهای امتیازدهی، محدود به بانکها و شرکتهای بیمه نمی‌شود، بلکه در مسائل مشابه برای تصمیم‌گیری در سایر بخشها نظیر مخابرات و شرکتهای پست هم می‌توانند مورد استفاده قرار گیرند.

پیش‌بینی عدم توانایی یک مشتری برای بازپرداخت وام، به‌عنوان یکی از مسائل مهم و مورد بحث در حسابداری از زمانی که فیتز پاتریک(Fitspatrik, 1930) مطالعاتی را در مورد آن انجام داد مطرح بوده و طی ۶۰ سال اخیر این موضوع به زمینهٔ مهمی در تحقیقات نظری و تجربی در حوزه اقتصاد مالی تبدیل شده است[2]. یکی از اولین محققانی که در این موضوع مطالعات زیادی انجام داده و می‌توان او را در شمار پیشگامان این مبحث برشمرد، ادوارد آلتمن (Edward Altman, 1968) است. او تلاش بسیاری برای یافتن یک رابطه معنی‌دار بین متغیرهای حسابداری یک شرکت و احتمال عدم توانایی در پرداخت دیون این شرکت در آینده انجام و رابطه‌ای معروف به *Z-Score* را ارائه داد. این روش مبتنی بر تحلیل تفکیکی خطی[3] بین شرکتهای خوب و بد بود. با این روش فقط شرکتهای خوب از بد جدا می‌شدند. با اینکه این روش بسیار ابتدایی بود اما تا حدودی وضعیت شرکتهای بد را می‌توانست پیش‌بینی نماید.

با گسترش تکنیکها و استفاده از روشهای پیچیده و دقیقتر مانند رگرسیون‌های خطی، مدلهای لاجیت و پروبیت[4] (کرامر، ۲۰۰۱) در دهه‌های اخیر و استفاده از مدلهای مبتنی بر شبکه‌های عصبی[5]، دقت این پیش‌بینی‌ها تا حد قابل قبولی افزایش یافته است. ابتدا روش تحلیل تفکیکی خطی(بیور، ۱۹۶۶) روش مورد استفاده بود. در سال ۱۹۶۸ این روش توسط آلتمن به یک روش چندمتغیره[6] توسعه یافت و تا سال ۱۹۸۰ چارچوب تحلیلی اکثر مطالعات بود. در خلال دهه ۱۹۸۰ روش تحلیلی لاجستیک[7] جایگزین این روش شد که تاکنون هم در شمار پرکاربردترین روشهای آماری مورد استفاده برای پیش‌بینی احتمال نکول است.

---

1. Bill Fair and Earl Isacc

۲. برای مطالعه بیشتر به Altman & etal(1994) و Lo(1986) مراجعه شود.

3. Discreminant Analysis 4. Probit 5. Neural Networks

6. Multivariate Discriminant Analysis 7. Logistic

هر دو این روشها سعی بر جداسازی وامهای خوب و بد از بین درخواستهای پذیرفته‌شده دارند؛ اما آنچه باعث ایجاد مشکل می‌شود این است که این مدلها مقید به چولگی موجود در نمونه انتخاب‌شده هستند. به این ترتیب که چون مدل نهایی بر اساس داده‌های انتخاب‌شده ساخته می‌شود، چولگی موجود در نمونه هنگامی که مدل به درخواستهای جدید اعمال می‌گردد، باعث کاهش دقت در کار می‌شود. به این دلیل، اغلب تحقیقات اخیر روشهای ناپارامتریک نظیر مدلهای نزدیکترین k همسایه[1](Hand & Henley, 1997)، داده شماری[2](Dionne & etal., 1996)، شبکه‌های عصبی(Leea & etal., 2002) و درختهای طبقه‌بندی و رگرسیونی (Arminger & etal., 1997) را مورد استفاده قرار داده‌اند.

هر چند در این تحقیق برای نخستین بار در ایران به امتیازدهی مشتریان حقیقی یک بانک پرداخته شده است، اما رویکرد اغلب محققان که علاقه‌مند به مطالعه در خصوص سیستمهای امتیازدهی هستند به سوی استفاده از مدل لاجیت است. این روش دارای برخی مزایاست. به‌عنوان مثال ضرورتاً به فروض مدلهای تحلیل تفکیک خطی و تحلیل تفکیک چندمتغیره[3] نیاز ندارد و بر خلاف این دو مدل، فرض نرمال‌بودن چندمتغیره و یکسان‌بودن ماتریس کوواریانس را در نظر نمی‌گیرد(سبزواری، ۱۳۸۴).

آنچه استفاده از روش رگرسیون لاجستیک را دچار مشکل می‌سازد، نوع متغیرهای موجود در ارزیابی اعتباری افراد است. به این صورت که اکثر متغیرهای(توضیحی) موجود گسسته(کیفی) و به شکل انتخاب دوگانه ظاهر می‌شوند. در این حالت به‌نظر می‌رسد استفاده از یک روش ناپارامتریک می‌تواند راه حل کاراتری برای طبقه‌بندی افراد نسبت به روشهای پارامتریک نظیر رگرسیون لاجستیک باشد. روش مورد نظر که به اختصار $CART$[4] نامیده می‌شود، در دهه ۸۰ توسط بریمن و دیگران(Breiman & etal., 1984) در مقاله‌ای تحت عنوان "درختهای طبقه‌بندی و رگرسیونی" منتشر شد. استفاده از درختهای طبقه‌بندی برای ارزیابی و طبقه‌بندی افراد در فرایند وام‌دهی بر اساس کار کوئین لن قرار گرفته است که به‌عنوان یکی از رهیافتهای یادگیری ماشینی شناخته می‌شود(Quinlan, 1983). کارتر و کاتلت الگوریتم مشابهی برای استفاده در فرایند ارزشیابی کارتهای اعتباری ایجاد کردند. آنها سیستم نمونه‌ای را بر اساس ۶۰۰ مشاهده ارائه نمودند و نشان دادند این روش به‌صورت رضایت‌بخشی مشکل دسته‌بندی مشتریان را حل می‌کند (Carter & cutlett, 1987). اخیراً هم، لی و دیگران نتایج مقایسه‌ای بین شبکه‌های عصبی و

1. k-nearest neighborhood 2. Count Data
3. Linear Discriminant Analysis and Multivariate Discriminant Analysis
4. Classification And Regression Trees

CART را ارائه کرده‌اند. نتایج مطالعه آنها نشان می‌دهد، هنگامی که حجم داده‌ها و تعداد متغیرها زیاد است استفاده از روش *CART* راه حل ساده‌تر و کاراتری است(Lee & etal., 2006).

آنچه در این تحقیق در پی آن خواهیم بود، نشان‌دادن این امر است که با توجه به ساختار اطلاعات موجود در مورد مشتریان بانکها در ایران، استفاده از روش ناپارامتریک "درختهای طبقه‌بندی" در مقایسه با مدل لاجیت برای طبقه‌بندی مشتریان دقت و کارآیی بیشتری دارد. ابتدا این دو روش به اختصار از لحاظ نظری بررسی شده و سپس نتایج به‌دست‌آمده از یک نمونه برای مقایسهٔ دو روش مورد ارزیابی قرار می‌گیرد. هدف اصلی انجام این تحقیق مقایسهٔ نتایج به‌دست‌آمده از این دو روش است و در نتیجه، استفاده از یک مجموعه داده مربوط به یک شعبهٔ خاص از یک بانک برای هر دو روش، می‌تواند فرایند مقایسه را امکان‌پذیر نماید. هدف از تحقیق تخمین پارامترهای جامعهٔ مورد نظر نیست و در نتیجه چولگی احتمالی مجموعه دادهٔ مورد استفاده، تأثیری در صحت نتایج نخواهد داشت.

۲- داده‌های پژوهش

جهت برآورد یک مدل امتیازدهی اعتباری برای مشتریان حقیقی از مجموعه اطلاعات یکی از شعب بانک مسکن استفاده می‌شود. یک مجموعه مشاهده از مشتریان در فاصله زمانی ۱۳۸۳-۱۳۸۰ در اختیار بود که به‌طور تصادفی تعداد ۲۴۰ مشاهده از این مجموعه انتخاب شده است. از این بین تعدادی از مشاهدات موجود، به دلیل نقص اطلاعات اساسی در پروندهٔ مشتریان کنار گذاشته شد. بنابراین تعداد ۲۰۰ مشاهده برای استفاده در مدل مورد ارزیابی قرار گرفت. از مجموع مشتریان فوق تعداد ۱۳۶ مشتری به موقع اقساط خود را پرداخت کرده و خوش‌حساب[1] تلقی شدند و عدد صفر(۰) در تناظر با آنها قرار داده شده است و این در حالی است که بقیه آنان، یعنی ۶۴ مشتری، بدحساب بوده و برای آنها عدد یک(۱) اختصاص می‌یابد.

متغیر وابسته در مدل لاجیت، نوع رفتار پرداختی مشتری یعنی خوش‌حساب‌بودن یا بدحساب‌بودن آن است که متناظر عدد صفر و یک خواهد بود. این متغیر با سرواژهٔ *banr* نشان داده می‌شود. متغیرهایی که به‌صورت بالقوه امکان قرارگرفتن در طرف راست مدل به‌عنوان متغیرهای توضیحی را خواهند داشت عبارتند از[2]:

---

۱. منظور از مشتریان خوش‌حساب، مشتریانی است که (طبق تعریف بال) یا هیچ‌گونه تأخیری در پرداخت اقساط خود نداشته‌اند و یا حداکثر ۳ ماه تأخیر دارند. درحالی‌که مشتری بدحساب دارای حداقل ۳ ماه تأخیر است.

۲. روزباخ به حدود ۴۸ متغیر در خصوص ارزیابی مشتریان حقیقی اشاره می‌کند(Roszbach, 1998) که با توجه به وضعیت کسب اطلاعات از مشتریان در بانکهای ایران، اغلب این متغیرها قابل دستیابی نیستند و بنابراین در این تحقیق از اطلاعات و متغیرهای موجود استفاده شده است.

Child : نشانگر تعداد فرزندان مشتری است. لذا شامل اعداد ۰، ۱، ۲ و ... خواهد بود.
Wedu : سطح تحصیلات همسر فرد درخواست‌کننده اعتبار. بدین‌ترتیب که برای سطح تحصیلات دیپلم و پایین‌تر از آن عدد صفر، برای سطح لیسانس عدد ۱ و برای فوق لیسانس و بالاتر عدد ۲ در نظر گرفته شده است.
Wjob : شاغل‌بودن یا نبودن همسر مشتری. لذا اعداد صفر و یک متناظر آن خواهد بود. اگر همسر مشتری شاغل باشد عدد ۱ و در غیر این صورت به آن عدد صفر اختصاص می‌یابد.
Hom : اگر مشتری دارای خانه باشد به این متغیر عدد ۱ و در غیر این صورت عدد صفر داده می‌شود.
Edu : سطح تحصیلات متقاضی اعتبار است، لذا به مانند *Wedu* برای سطح تحصیلات دیپلم و پایین‌تر از آن عدد صفر، برای سطح لیسانس عدد ۱ و برای فوق لیسانس و بالاتر عدد ۲ در نظر گرفته شده است.
Work : نوع کار مشتری. اگر مشتری دارای شغل دولتی باشد، مقدار این متغیر برای آن ۱ و اگر آزاد باشد مقدار صفر خواهد بود.
Age : سن فرد متقاضی اعتبار را نشان می‌دهد.
Gend : جنسیت درخواست‌کننده اعتبار را نشان می‌دهد. اگر مرد باشد عدد ۱ و اگر زن باشد عدد صفر در نظر گرفته می‌شود.

توجه داریم که علاوه بر متغیرهای فوق که به‌صورت بالقوه می‌توانند وارد مشخص نمایی مدل نهایی شوند، متغیرهای متنوع‌تر دیگری نیز می‌توانستند در مدل در نظر گرفته شوند، نظیر وضعیت بدهی متقاضی اعتبار به سایر بانکها، درآمد ماهانه مشتری، وضعیت پرداختی مشتری نزد بانک مسکن و نزد سایر بانکها، میزان و ارزش دارایی‌ها و از این قبیل که این دسته اطلاعات در مجموعه اسناد بانکی، آماری وجود نداشت و در نتیجه امکان بررسی آنها در مطالعه تجربی وجود ندارد. هر چند نباید از اهمیت فوق‌العاده برخی از این متغیرها مثل درآمد ماهیانه غافل بود و حتی می‌توان بنا به تئوری و نیز مطالعات تجربی اذعان کرد این متغیر از توان توضیح‌دهندگی قابل توجهی برخوردار است(Roszbach, 1998).

۳- چارچوب نظری

در این نوشتار دو مدل متفاوت مورد بررسی قرار می‌گیرند که یک مدل پارامتریک و دیگری ناپارامتریک است. این روشها عبارتند از:

- مدل با انتخاب دوگانه در متغیر وابسته[1]
- روش درختهای طبقه‌بندی[2]

۱-۳- مدل با انتخاب دوگانه در متغیر وابسته

مدلهای لاجیت و پروبیت در مواردی استفاده می‌شوند که متغیر وابسته قابل مشاهده نباشد. متغیر وابسته در این موارد به‌صورت انتخاب دوگانه[3] ظاهر می‌شود. مدل استفاده شده در این قسمت مدل لاجیت است که از رگرسیون لاجستیک پیروی می‌کند. در رگرسیون لاجستیک مانند رگرسیون چندمتغیره، ضرایب متغیرهای مستقل برآورد می‌شود، لیکن چگونگی عملکرد آن کاملاً متفاوت است. در رگرسیون چندمتغیره از روش کوچکترین مجموع مجذورات استفاده می‌شود. در این روش، مجموع مجذورات اختلاف بین مقادیر واقعی و مقادیر پیش‌بینی‌شده متغیر وابسته حداقل می‌گردد. به‌دلیل تابع غیرخطی تبدیل لاجستیک، روش حداکثر درستنمایی[4] مورد استفاده قرار می‌گیرد. با این حال روش برآورد ضرایب هنوز از بسیاری جهات شبیه رگرسیون معمولی است.

مدل لاجستیک از منحنی لاجستیک پیروی می‌کند، بدین ترتیب این منحنی بر اساس داده‌های واقعی برازش می‌شود. داده‌های واقعی مربوط به متغیر وابسته، بر اساس اینکه پدیده مورد نظر اتفاق افتاده یا اتفاق نیفتاده، دو مقدار صفر و یک اختصاص داده می‌شود، لذا در بالا و پایین نمودار مزبور قرار می‌گیرند. وقوع یا عدم‌وقوع پدیده مزبور با توجه به سطوح مختلف از ترکیبات خطی متغیرهای مستقل تعیین می‌شود. برتری رگرسیون لاجستیک در این است که برای تعیین مقادیر صفر و ۱ تنها اطلاع از وقوع پدیده مورد نظر(به‌طور مثال خرید یک کالا، ریسک اعتباری یا موفقیت یا عدم‌موفقیت یک شرکت) کافی است. بدین ترتیب از این متغیر وابسته می‌توان به‌منظور تخمین وقوع یا عدم‌وقوع اتفاق مورد نظر سود جست. اگر احتمال وقوع بیش از ۰/۵ پیش‌بینی شود، در این صورت وقوع پدیده مورد نظر، حتمی تلقی می‌شود و در غیر این صورت وقوع پدیده، غیرحتمی خواهد شد.

آنچه در استفاده از این مدل مدنظر است، در درجه اول به‌دست‌آوردن احتمال کل ناتوانی در بازپرداخت وام دریافتی توسط افراد و در مرحله بعد، استخراج اثرات نهایی[5] هر یک از متغیرهای توضیحی است. اثر نهایی، تغییر در میزان احتمال رخدادن متغیر وابسته، در ازای یک واحد افزایش در متغیرهای توضیحی است. در این تحقیق متغیر وابسته، قصور مشتری در پرداخت دیون است. با استفاده از مدل لاجیت داریم:

---

1. Binary Choice Model
2. Regression Trees
3. Binary Choice
4. Maximum Likelihood
5. Marginal Effects

$$P(y=1|\beta'x)=\frac{1}{1+\exp(\beta'x)} \quad (1)$$

که اثر نهایی متغیر فرضی x_r از رابطهٔ زیر به‌دست می‌آید.

$$\frac{\partial p}{\partial x_r}=f(\beta'x)\beta_r \quad (2)$$

بنابراین اثر نهایی متغیر x_r به‌طور ضمنی وابسته به خود متغیر، توسط مقدار عددی چگالی توزیع لاجستیک است(Maddala, 1983).

در تحلیل نتایج به‌دست‌آمده، اول احتمال کل به‌دست‌آمده به شرط قصور مشتری که از آن در محاسبه احتمال عدم توانایی مشتری در بازپرداخت وام استفاده می‌گردد، محاسبه می‌شود و سپس اثرات نهایی متغیرها که برای به‌دست‌آوردن نتیجهٔ افزایش یک واحد در هر یک از متغیرها بر احتمال قصور فرد لازم است استخراج می‌شوند.

۲-۳- روش درختهای طبقه‌بندی

درختهای طبقه‌بندی یک روش داده‌کاوی[1] است که ساختاری به نام درختهای تصمیم را به‌وجود می‌آورند. این درختهای تصمیم برای طبقه‌بندی داده‌های جدید مورد استفاده قرار می‌گیرند. در ابتدای به‌وجودآمدن، این نوع دسته‌بندی بیشتر در کارخانه‌ها و ماشینی‌کردن سیستمها مورد استفاده قرار می‌گرفتند(Quinlan, 1983). اکنون حوزه‌های مختلفی در اقتصاد مالی وجود دارند که می‌توان از روشهای مهندسی در آنها استفاده کرد؛ مانند ارزیابی وام، بیمه‌کردن، انتخاب فرصتهای سرمایه‌گذاری، مدیریت بدهی‌ها و دارایی‌ها، خرید و فروش سهام، اوراق قرضه، ارز و پیش‌بینی در بازارهای پولی(Lee & etal. 2006).

درختهای تصمیم از یک دسته سؤال تشکیل می‌شوند که نمونه آموزشی را به بخشهای کوچک و کوچکتری تقسیم می‌کنند. این امر تا جایی ادامه پیدا می‌کند که مجموعه‌ای از مشاهدات باقی بمانند که آنها را به هیچ طریقی نتوان در گروه‌های جدا از هم قرار داد. در روش *CART* سؤالها فقط به‌صورت بله یا خیر هستند. به‌عنوان مثال یک سؤال ممکن می‌تواند به‌صورت «آیا سن فرد بزرگتر از ۵۰ است یا خیر؟» و یا «جنسیت فرد مذکر است یا مؤنث؟» طرح شود. الگوریتم استفاده از *CART* برای تعیین خصوصیات و سؤالهای تفکیک‌کننده به‌صورت بازگشتی است، به‌طوری که این الگوریتم تمام حالتهای جداسازی و تمام متغیرهایی که می‌تواند عامل تفکیک دو گروه باشد را مشخص کرده و سپس بر اساس آن شروع به تقسیم و دسته‌بندی داده‌ها می‌نماید. هر سؤال،

1. Data mining

داده‌های باقیمانده را به دو گروه تقسیم می‌کند که بیشترین تفاوت را با هم داشته باشند و از طرفی داده‌های قرار گرفته در هر گروه، بیشترین همگنی را با سایر داده‌های هم‌گروه خواهند داشت. این فرایند برای هر قسمت باقیمانده از داده‌ها تکرار خواهد شد.

روش $CART$ دارای سه بخش عمده است:

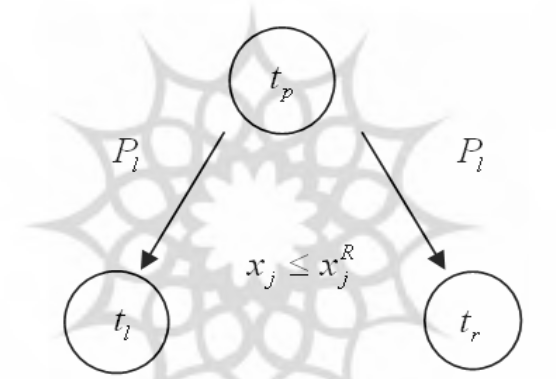
۱. ساختن بزرگترین درخت ممکن

۲. انتخاب اندازه صحیح درخت

۳. طبقه‌بندی داده‌های جدید با استفاده از ساختار درختی ساخته شده.

در مرحلهٔ اول با استفاده از یک الگوی تفکیک، اقدام به ساختن درخت اولیه می‌نماییم. هر الگوی تفکیک سعی بر کمینه‌کردن تابع ناخالصی بین گره اصلی و گره‌های فرعی دارد. شکل (۱) این موضوع را نشان می‌دهد.

شکل ۱. الگوریتم تفکیک در روش CART



در این شکل t_p ، t_r و t_l به ترتیب گره اصلی، گره فرعی راست و گره فرعی چپ هستند. x_j متغیر jام و x_j^R بهترین مقدار x_j برای تفکیک می‌باشد. بیشترین همگنی یک گره فرعی با تابعی به نام تابع ناخالصی $i(t)$ تعیین می‌شود. برای هر کدام از حالتهای تفکیک چپ و راست، میزان ناخالصی گره اصلی ثابت باقی می‌ماند، یعنی برای تمام حالتهای $x_j \le x_j^R$ و $j = ۱,...,M$ مقدار همگنی و ناخالصی گره اصلی ثابت است. بنابراین بیشترین مقدار همگنی گره‌های فرعی چپ و راست برابر با بیشترین مقدار تغییر در تابع ناخالصی خواهد بود. به عبارت دیگر گره‌های سمت چپ و راست گره اصلی، زمانی بهترین حالت تفکیک را خواهند داشت که ناخالصی بین آنها که همان تغییر در ناخالصی کل است، بیشینه باشد که آن را با $\Delta i(t)$ نشان می‌دهیم.

فرض کنید p_r و p_l به ترتیب احتمالهای گره‌های سمت چپ و راست باشند. به این معنی که یک مشاهده با چه احتمالی در گره سمت راست و با چه احتمالی در گره سمت چپ قرار خواهد گرفت. بنابراین می‌توانیم بنویسیم:

$$\Delta i(t) = i(t_p) - p_l i(t_l) - p_r i(t_r) \qquad (۴)$$

در نتیجه در هر گره در روش $CART$ به حل یک مسأله ماکزیمم‌سازی به‌صورت زیر خواهیم پرداخت:

$$\underset{x_j \le x_j^R,\ j=۱....M}{\arg\max} \quad [i(t_p) - p_l i(t_l) - p_r i(t_r)] \qquad (۵)$$

برای استفاده در الگوریتم تفکیک، از توابع ناخالصی متعددی می‌توان استفاده نمود. از بین این توابع، الگوی تفکیک جینی یا شاخص جینی بیشترین موارد استفاده را داراست. این الگو از تابع ناخالصی زیر برای جداسازی استفاده می‌کند.

$$i(t) = \sum_{k,l=1}^{K} p(k \mid t)\, p(l \mid t) \qquad (۶)$$

که در آن l و k ، $۱...K$، شاخصی برای کلاسها هستند و $P(k \mid t)$ احتمال شرطی کلاس k است، هنگامی که در گره t قرار داریم. با استفاده از تابع ناخالصی جینی(۶) در مسأله حداکثرسازی (۵) می‌توانیم مقدار تغییر در تابع ناخالصی را اندازه‌گیری نماییم. بنابراین تغییرات در تابع ناخالصی را به‌صورت زیر بازنویسی می‌کنیم:

$$\Delta i(t) = -\sum_{k=۱}^{K} p^{۲}(k \mid t_p) + P_l \sum_{k=۱}^{K} p^{۲}(k \mid t_l) + P_r \sum_{k=۱}^{K} p^{۲}(k \mid t_r) \qquad (۷)$$

سرانجام الگوریتم جینی مسأله زیر را حل خواهد کرد:

$$\underset{x_j \le x_j^R,\ j=۱....M}{\arg\max} \left[-\sum_{k=۱}^{K} p^{۲}(k \mid t_p) + P_l \sum_{k=۱}^{K} p^{۲}(k \mid t_l) + P_r \sum_{k=۱}^{K} p^{۲}(k \mid t_r) \right] \qquad (۸)$$

طبق این الگوریتم تمام نمونه آموزشی، مورد ارزیابی و جستجو قرار می‌گیرد تا بزرگترین کلاسهای موجود از داده‌ها تشکیل شده و سرانجام جداسازی شوند. این الگو به خوبی با داده‌هایی هم که دچار پراکندگی هستند کار می‌کند.

در مرحلهٔ دوم باید اندازهٔ بهینهٔ درخت تعیین شود. درختی که به‌صورت اولیه و به‌عنوان درخت بیشینه ایجاد می‌شود، ممکن است دارای پیچیدگی بسیار و همچنین دارای صدها سطح و گره باشد که این امر، تحلیل این درختها را مشکل می‌کند. همچنین با افزایش تعداد مشاهدات، این پیچیدگی افزایش می‌یابد. بنابراین درخت به‌وجودآمده قبل از به‌کارگیری توسط روشهایی باید

بهینه شده و شاخه‌های اضافی و حتی برخی زیردرختها باید از ساختار این درخت حذف شوند که به این کار هرس درخت می‌گویند. بهینه‌سازی درخت به معنی انتخاب بهترین اندازه درخت است. در عمل از دو الگوریتم برای هرس شاخه‌های اضافی استفاده می‌شود: اول بهینه‌سازی با استفاده از تعداد نقاط موجود در هر گره و دوم استفاده از وارسی اعتباری[1] است.

در این تحقیق با استفاده از روش اول، اقدام به تعیین اندازهٔ صحیح درخت خواهیم نمود. در این روش تعیین می‌کنیم که اگر تعداد مشاهدات موجود در یک گره از مقدار $N_{\min}$ معین کمتر بود روند تفکیک متوقف شده و این گره به‌عنوان یک گره نهایی محسوب می‌شود. کاملاً واضح است که با افزایش $N_{\min}$ اندازه درخت کاهش می‌یابد. دو مقدار نهایی برای $N_{\min}$ می‌توان در نظر گرفت، ۱ و N. اولی در صورتی است که در هر یک از گره‌های نهایی فقط یک مشاهده باقی بماند، این حالت همان درخت بیشینه است و حالت دوم یعنی تمام مشاهدات، که در این صورت هیچ تفکیکی صورت نخواهد گرفت. تعیین مقدار $N_{\min}$ بستگی زیادی به هدف دسته‌بندی یا رگرسیون دارد. این روش، یعنی استفاده از $N_{\min}$ بسیار سریع بوده و به سادگی قابل استفاده است. نتایج به‌دست‌آمده از آن هم در حد قابل قبولی قرار دارند. در عمل، مقدار $N_{\min}$ به اندازه ۱۰٪ نمونه مورد آزمایش قرار داده می‌شود که در اغلب موارد هم جواب صحیحی به‌دست می‌آید. به‌عنوان مثال برای داده‌های مورد آزمایش با احتساب $N_{\min}=۱۵$ مقدار تابع ناخالصی به‌دست‌آمده برابر است با $i=۰/۴۱۴$ و تعداد گره‌های انتهایی $\tilde{T}=۷$ است. اگر $N_{\min}=۱$ باشد، آنگاه $i=۰/۴۱۴$ و $\tilde{T}=۱۲$ خواهد بود. به‌راحتی می‌توانیم مشاهده کنیم که با افزایش مقدار $N_{\min}$ ناخالصی در درخت افزایش می‌یابد و به عبارت دیگر، پیچیدگی درخت که تعداد گره‌های انتهایی نشان‌دهنده آن است کاهش می‌یابد(Timofeev, 2004).

هنگامی که یک درخت طبقه‌بندی و رگرسیونی ساخته شد، می‌تواند برای طبقه‌بندی داده‌های جدید، مورد استفاده قرار گیرد. آنچه از اعمال هر کدام از این روشها بر داده‌های جدید به‌دست می‌آید، نسبت یک کلاس یا یک مقدار پاسخ به هر کدام از مشاهدات جدید است. با توجه به یک سری پرسشهایی که در بدنه درخت قرار گرفته‌اند، هر کدام از مشاهدات جدید در یک گره انتهایی جای خواهند گرفت و به این ترتیب هر مشاهده در دسته مناسب قرار می‌گیرد. بنابراین هر مشاهده بر اساس نزدیکترین کلاس و یا مقدار پاسخ داده شده به آن، طبقه‌بندی شده و در یک گره نهایی قرار داده می‌شود. کلاس هر گره بر اساس مشاهدات مربوط به بیشترین کلاس تعیین می‌شود، به‌عنوان مثال اگر در یک گره ۵ مشاهده مربوط به کلاس ۱ و ۲ مشاهده مربوط به کلاس ۲ و صفر

1. Cross Validation

مشاهده مربوط به کلاس ۳ قرار گرفته باشد، کلاس ۱ غالب خواهد بود و گره متعلق به کلاس ۱ خواهد بود.

روش درختهای طبقه‌بندی دارای مزایایی است که تعدادی از آنها را ذکر می‌کنیم.

- یک روش ناپارامتریک است، در نتیجه این روش احتیاج به تعریف و یا تعیین شکل تابعی ندارد.
- احتیاجی به اینکه متغیرها از پیش انتخاب شده باشند، ندارد.
- نتایج آن در یک انتقال یکنواخت در متغیرهای مستقل یکسان باقی می‌مانند.
- به‌راحتی می‌تواند داده‌های پرت[1] را کنترل کند.
- هیچ فرض اولیه وجود ندارد، که این امر از لحاظ محاسباتی باعث کاهش در مدت زمان انجام محاسبات می‌شود.

۴-تحلیل نتایج تجربی

در این قسمت نتایج تجربی به‌دست‌آمده از دو روش را بیان می‌کنیم. سپس مقایسه‌ای بین نتایج به‌دست‌آمده از دو روش صورت می‌گیرد.

۴-۱- برآورد یک مدل رگرسیون لاجستیک

جهت معرفی یک مدل مناسب نمره‌دهی، ابتدا متغیر وابسته بر روی تمام متغیرهای توضیحی که در بالا بدانها اشاره شد برازانیده می‌شود. خروجی نهایی این مدل در جدول شمارهٔ ۱ مشاهده می‌گردد. این مدل شامل ۸ متغیر توضیحی می‌باشد که ضرایب دو تا از این متغیرها اختلاف معنی‌داری از صفر نداشته و در نتیجه، از نظر آماری، این متغیرها دارای ضریب صفر می‌باشند. این دو متغیر، سطح تحصیلات همسر($Wedu$) و سن مشتری(Age) می‌باشند و بنابراین از مدل نهایی حذف گردیده‌اند.

1. Outliers

جدول ۱. خروجی نهایی مدل رگرسیونی لاجیت

متغیر وابسته	ضرائب	انحراف استاندارد خطا	مقدار z استاندارد	مقدار p - value
تعداد فرزندان	۱/۶	۰/۴۴	۳/۶۷	۰/۰۰
شغل همسر	-۱/۷	۰/۵۵	-۳/۱۸	۰/۰۰
مالکیت مسکن	-۳/۱	۰/۶۷	-۴/۶۱	۰/۰۰
سطح تحصیلات	-۱/.	۰/۳۹	-۲/۶۲	۰/۰۱
نوع کار(دولتی و آزاد)	-۲/۰	۰/۵۵	-۳/۷۱	۰/۰۰
جنسیت	۲/۴	۰/۶۷	۳/۶۳	۰/۰۰

در مورد متغیر سطح تحصیلات همسر، بهتر است به این نکته اشاره شود که وجود این متغیر در مدل باعث از دست دادن تعدادی از مشاهدات و کاهش تعداد آنها به ۱۴۳ مشاهده می‌گردد، اما به‌علت اهمیت آن در مدل و معنی‌داربودن ضریب آن، در مدل نهایی گنجانیده شده است.

جدول ۲. چگونگی پیش‌بینی در مدل رگرسیونی لاجیت

کلاس	موارد تشخیص به عنوان بدحساب	موارد تشخیص به عنوان خوش حساب	کل
۱	۴۷	۱۱	۵۸
۰	۱۰	۷۵	۸۵
مجموع	۵۷	۸۶	۱۴۳
حساسیت(*Sensitivity*)	۸۲/۴۶ درصد		
ویژگی(*Specificity*)	۸۶/۲۱ درصد		
میزان تفکیک در کلاس ۱	۸۱/۰۳ درصد		
میزان تفکیک در کلاس ۰	۸۸/۲۴ درصد		
خطای نوع اول	۱۸/۹۷ درصد		
خطای نوع دوم	۱۱/۷۶ درصد		
میزان تفکیک کلی مدل	۸۵/۳۱ درصد		

در جدول شمارهٔ ۲ نتایج مربوط به پیش‌بینی و طبقه‌بندی مدل نهایی مشاهده می‌گردد. از مجموع ۱۴۳ مشاهده، تعداد ۵۸ مشاهده بدحساب و ۸۵ مشاهده خوش‌حساب هستند. بر اساس جدول، تعداد ۴۷ مشاهده به‌درستی بدحساب تشخیص داده شده‌اند و همین‌طور ۷۵ مشتری نیز به‌درستی در طبقه خوش‌حساب قرار داده شده‌اند. لذا دقت مدل در مورد مشتریان بدحساب برابر

۸۱/۰۳ درصد، در حالی که برای مشتریان خوش‌حساب برابر ۸۸/۲۴ درصد می‌باشد. بنابراین میزان طبقه‌بندی نهایی مدل، برابر با ۸۵/۳۱ درصد است. **حساسیت**[1] نشان‌دهنده نسبتی از مشتریان بدحسابی است که به‌درستی توسط مدل در طبقه مشتریان بدحساب قرار گرفته‌اند، در مقایسه با کل مشتریانی که به‌عنوان بدحساب تشخیص داده شده‌اند. به عبارت دیگر، احتمال قصور این مشتریان بیش از نمره برش(۰/۵) پیش‌بینی شده است. **ویژگی**[2] نشان‌دهنده نسبتی از مشتریان خوش‌حساب است که به‌درستی در طبقه مشتریان خوش‌حساب قرار گرفته‌اند، در مقایسه با کل مشتریانی که خوش‌حساب تشخیص داده شده‌اند. در این خصوص در مورد مشتریانی که به اشتباه بدحساب تشخیص داده شده‌اند، خطای نوع اول و در مورد مشتریانی که به اشتباه خوش‌حساب تشخیص داده شده‌اند، خطای نوع دوم صورت پذیرفته است.

۱-۱-۴- آزمون نیکویی برازش

نرم‌افزارهای آماری و اقتصادسنجی دو آزمون نیکویی برازش هاسمر- لمشوف[3] و اندروز[4] را انجام می‌دهد. ایده‌ی اصلی این دو آزمون، آن است که مقادیر برازش‌شده مورد انتظار را با مقادیر واقعی هر گروه مقایسه می‌کند و اگر اختلافات بزرگ باشد، مدل را رد می‌کنیم؛ چرا که برازش نامناسبی برای داده‌ها فراهم می‌کند. به‌طور خلاصه می‌توان گفت این دو آزمون در گروه‌بندی مشاهدات و در توزیع مجانبی آماره آزمون متفاوتند. آزمون هاسمر- لمشوف، مشاهدات را بر پایه پیش‌بینی احتمال اینکه $Y = ۱$ باشد، گروه‌بندی می‌کند. آماره $\chi^۲$ در پایین جدول ۳ گزارش شده است. مقدار آماره هاسمر- لمشوف(۳/۹۱) از ۱۶ کمتر است، لذا این آزمون نیز نیکویی مدل برازش‌شده را تأیید می‌نماید.

جدول ۳. خروجی آزمون نیکویی برازش

آزمون نیکویی برازش	
تعداد مشاهدات	۱۴۳
مقدار آماره هاسمر- لمشوف	۳/۹۱
$Prob > chi2$	۰/۱

1. Sensitivity 2. Specificity 3. Hosmer-Lemeshow (1989)
4. Anrdews(1988a,1988b)

۲-۱-۴- تحلیل اثرات نهایی

تفسیر مقادیر ضرایب مدل لاجیت پیچیده است؛ چرا که ضرایب برآورد شده حاصل یک مدل دوگزینه‌ای است که نمی‌تواند به عنوان اثر نهایی روی متغیر وابسته تفسیر شود. اثر نهایی x_j روی احتمال شرطی به‌وسیله رابطه‌ی زیر تعیین می‌شود.

$$\frac{\partial E(y/x,\beta)}{\partial x_j} = f(-x'\beta).\beta_j$$

در این رابطه $f(x)=\frac{dF(x)}{dx}$ تابع چگالی $F(x)$ است. توجه داشته باشیم که β_j به‌وسیله عامل f که خود بستگی به مقادیر همه توضیح‌دهنده‌ها، در بردار x دارد، وزن‌دار می‌شود. از آنجایی‌که تابع f همیشه دارای مقدار مثبت است، جهت اثر نهایی به علامت β_j بستگی دارد. اگر β_j عددی مثبت باشد، افزایش x_j باعث افزایش احتمال وقوع متغیر وابسته می‌شود.

به منظور بررسی میزان تأثیرگذاری متغیرهای گوناگون در احتمال بدحساب بودن، اثرات نهایی و علامت مربوط به ضرایب آنان بررسی می‌شود. نتایج مربوط به محاسبه اثرات در جدول(۴) آمده است.

جدول ۴. محاسبه اثرات نهایی بعد از تخمین مدل لاجیت

متغیر	مقدار dy/dx	مقدار $prob$
تعداد فرزند	۰/۲۶۱	۰
وضعیت مسکن	-۰/۴۸۳	۰
سطح تحصیلات	-۰/۲۵۵	۰
نوع کار	-۰/۳۵۵	۰
جنسیت	۰/۱۹۴	۰

برای متغیر تعداد فرزندان اثر نهایی ۰/۲۶۱ می‌باشد و این بدان معنی است که با افزایش ۱ فرزند احتمال قصور به میزان ۲۶/۱ درصد افزایش پیدا می‌کند. اثر نهایی و یا کشش در مطالعه تجربی فوق نشان می‌دهد که متغیر وابسته یا احتمال قصور به ترتیب به متغیرهای مسکن، نوع کار و تعداد فرزندان، حساسیت بیشتری دارد. به عبارت دیگر احتمال قصور به وضعیت خانه‌داربودن یا نبودن مشتری، نوع کار وی و نیز تعداد فرزندان خانواده بستگی شدیدتری نسبت به سایر متغیرها دارد. در نتیجه اگر بانک بخواهد بر اساس این مدل در مورد مشتریان خود تصمیم‌گیری کند، باید در مورد مالکیت خانه، نوع کار و تعداد فرزندان فرد، دقت بیشتری به خرج دهد.

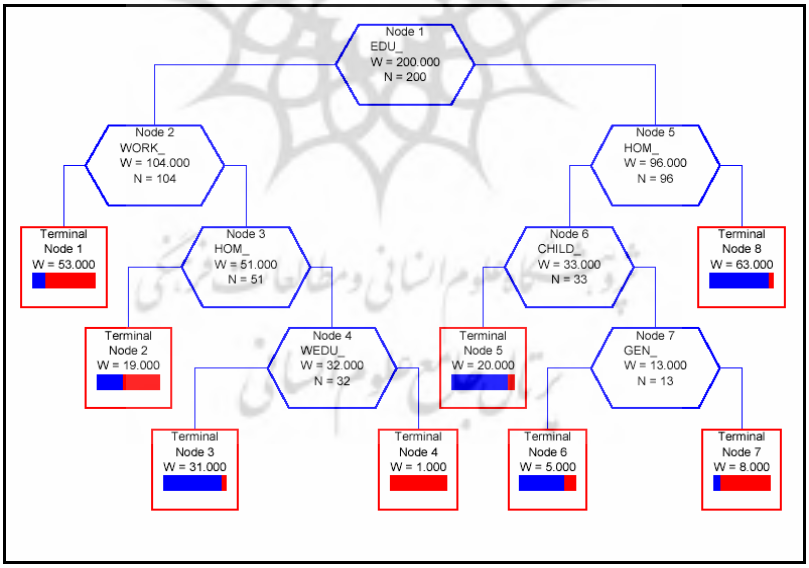
البته این نتایج را نمی‌توان برای همه بانکها و در همه کشورها تعمیم داد، چرا که نوع و چگونگی متغیرها می‌تواند جامع‌تر و متنوع‌تر از مجموعه متغیرهای مدل ما باشد. برای مثال ما به دلیل در اختیار نداشتن درآمد مشتریان بانک از یک متغیر بسیار اساسی صرف‌نظر کرده‌ایم و این در حالی است که برای بانکهای موفق خارجی، این متغیر از اهمیت بسیاری برخوردار است.

۲-۴- تفکیک و طبقه‌بندی داده‌ها با استفاده از روش درختهای طبقه‌بندی

در این قسمت با استفاده از روش طبقه‌بندی درختی، مشتریان بانک را رتبه‌بندی خواهیم کرد و به هر گروه، یک کلاس نسبت خواهیم داد. آنچه در این بخش خواهیم دید این است که بر خلاف مدل رگرسیونی لاجیت که برای هر مشاهده یک احتمال به‌دست می‌آید، در این روش مشاهدات بر اساس نزدیکترین کلاس مرتبط دسته‌بندی شده، و سعی می‌شود مشاهدات موجود در هر کلاس بیشترین شباهت را به یکدیگر داشته باشند. در این روش، دیگر از متغیرها به‌عنوان یک رابطه استفاده نمی‌شود؛ بلکه متغیرها فقط به‌عنوان تفکیک‌کننده مورد استفاده قرار می‌گیرند.

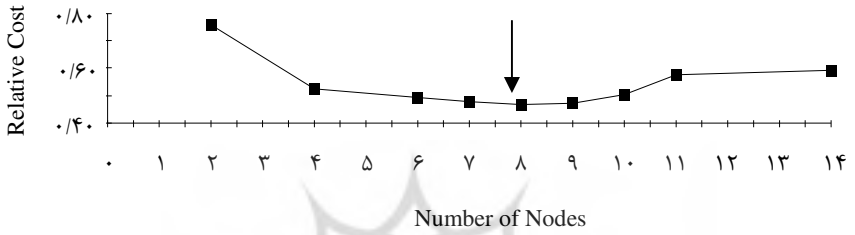
اکنون دسته‌بندی مشتریان را با استفاده از روش ناپارامتریک درختهای طبقه‌بندی انجام می‌دهیم. برای این منظور از نرم‌افزار *CART* طراحی‌شده توسط شرکت *Salford System* استفاده خواهیم کرد. نمودار (۱) درخت ایجادشده با این روش را نشان می‌دهد.

نمودار ۱. درخت ایجادشده بر اساس نمونه



در ابتدا بزرگترین درخت ممکن بر اساس نمونه ساخته می‌شود. سپس بر اساس پارامترهای تنظیم‌شده، بهترین اندازهٔ درخت با کمترین مقدار خطا تعیین می‌شود. همان‌طور که می‌دانیم اندازهٔ درخت بر اساس گره‌های نهایی تعیین می‌شود که در این مورد، تعداد گره‌های نهایی یا به عبارتی طبقات $T = ۸$ است. نمودار (۲) منحنی خطا را برای این درخت نشان می‌دهد که در مقدار $T = ۸$ به حداقل می‌رسد. به‌راحتی قابل مشاهده است که با افزایش یا کاهش تعداد گره‌های نهایی، میزان خطا افزایش می‌یابد. برای درخت مورد نظر این مقدار برابر با $i = ۰/۴۷$ است.

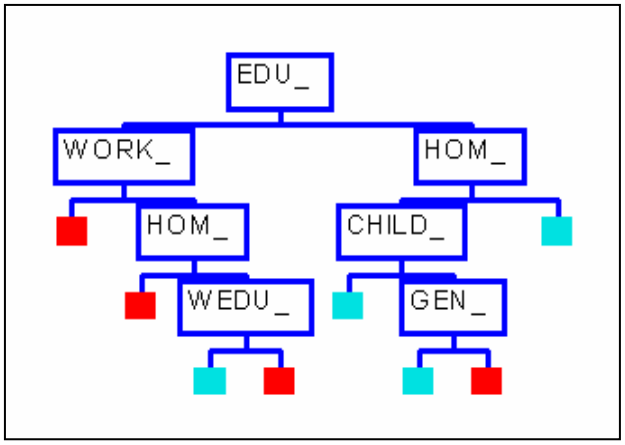
نمودار ۲. منحنی خطا به صورت تابعی از گره‌های نهایی درخت



البته با تغییر در پارامترهای مدل نظیر تغییر در الگوی تفکیک و مقدار $N_{\min}$ درختهایی با ساختار متفاوت به‌وجود می‌آید که در این مورد، بهترین حالت استفاده از الگوی جینی و مقدار $N_{\min} = ۱۰$ تشخیص داده شد. در الگوریتم تکراری ساخت درخت بارها ساختارهایی با متغیرهای تفکیک‌کننده ایجاد می‌شود و در نهایت بهترین درخت با توجه به گره‌های نهایی و کمترین میزان خطا انتخاب می‌شود. از قسمت (۳-۲) به خاطر داریم که با افزایش گره‌های نهایی، میزان خطا کاهش پیدا می‌کند و در حالت درخت بیشینه این مقدار صفر خواهد بود اما در این حالت پیچیدگی درخت که تعداد گره‌های نهایی بیانگر آن است، بیشترین مقدار خواهد بود. بنابراین بین کاهش میزان خطا و افزایش تعداد گره‌های نهایی باید یک مقدار بهینه انتخاب شود که در این مورد همان مقادیر ذکرشده در بالاست.

در درخت تصمیم به‌وجودآمده تفکیک بر اساس اهمیت متغیرها در به‌وجودآوردن گروه‌های تفکیکی انجام می‌گیرد. بر این اساس صورت کلی الگوی تفکیکی و متغیرهای تفکیک‌کننده در نمودار (۳) و اهمیت نسبی این متغیرها در جدول (۴) ارائه شده است که از این جدول می‌توانیم در امتیازدهی به مشتریان استفاده کنیم.

نمودار ۳. ساختار متغیرهای تفکیک‌کننده



جدول ۵. اهمیت نسبی متغیرها

متغیر	امتیاز
سطح تحصیلات	۱۰۰
شغل همسر	۹۶/۴۴
نوع کار	۷۷/۹۷
تعداد فرزندان	۶۶/۸۰
وضعیت مسکن	۶۲/۱۴
تحصیلات همسر	۵۳/۵۱
سن	۲۳/۷۵
جنسیت	۱۹/۳۸

در آخرین جدولی که در این قسمت ارائه می‌شود، دقت نهایی روش را نشان می‌دهیم. مقادیر به‌دست‌آمده برای درصد طبقه‌بندی صحیح، دقت نهایی مدل را نشان می‌دهد. این مقادیر برای مشتریان خوش‌حساب تقریباً هشتاد و پنج درصد و برای مشتریان بدحساب هشتاد و هفت درصد است.

جدول ۶. دقت نهایی روش درختهای طبقه‌بندی

کلاس	موارد تشخیص به‌عنوان بدحساب	موارد تشخیص به‌عنوان خوش‌حساب	کل
۱	۶۱	۹	۷۰
۰	۲۰	۱۱۰	۱۳۰
مجموع	۸۱	۱۱۹	۲۰۰
حساسیت(*Sensitivity*)		۷۵/۳۱ درصد	
ویژگی(*Specificity*)		۹۲/۴۴ درصد	
میزان تفکیک در کلاس ۱		۸۷/۱۴ درصد	
میزان تفکیک در کلاس ۰		۸۴/۶۱ درصد	
خطای نوع اول		۱۲/۸۶ درصد	
خطای نوع دوم		۱۵/۳۸ درصد	
میزان تفکیک کلی مدل		۸۵/۵۰ درصد	

در روش "درختهای طبقه‌بندی" یک مشاهدهٔ جدید بر اساس درخت شکل‌یافته از نمونهٔ مورد آزمایش طبقه‌بندی می‌گردد. به این طریق که برای هر نمونه بر اساس مهمترین متغیر(گره اول درخت) تفکیک صورت می‌گیرد و به همین ترتیب با توجه به گره‌های بعدی، مشاهدهٔ مزبور در یکی از هشت گره یا طبقهٔ نهایی دسته‌بندی می‌شود.

۱-۲-۴- استفاده از روش نمونه‌گیری بازگردان تکراری[1] برای ارزیابی کارآیی روش CART در مقابل مدل لاجیت

در قسمتهای گذشته نشان دادیم که استفاده از روش ناپارامتریک درختهای طبقه‌بندی و رگرسیونی در زمانی که داده‌های مورد استفاده اغلب ماهیت کیفی داشته، یعنی به‌صورت گسسته باشند، کارآیی بهتری نسبت به مدل لاجیت دارد. این امر برای یک نمونه از مشتریان یک بانک با تعداد ۲۰۰ مشاهده به اثبات رسید. اما سؤالی که ممکن است در ذهن خواننده ایجاد شود، این است که آیا این برتری برای نمونه‌های دیگر هم استوار است یا فقط در این نمونهٔ مورد مطالعه برتری وجود دارد. به‌علت محدودیت دسترسی به مجموعه داده‌های جدید، امکان بررسی این امر روی نمونه‌های دیگر از بانک امکان‌پذیر نیست. برای رفع این مشکل و نشان‌دادن کارآیی و استواری

1. Bootstrap

نتایج به‌دست‌آمده برای نمونه‌های مختلف از روش نمونه‌گیری بازگردان تکراری استفاده می‌شود. در این روش با ایجاد چندین نمونه تصادفی از بین مشاهدات اولیه، برای هر کدام از این نمونه‌ها، مدل اصلی به‌طور تکراری با استفاده از مدل لاجیت و درختهای طبقه‌بندی تخمین زده می‌شود. در این نمونه‌ها امکان تکرار برای مشاهدات وجود دارد و بنابراین صحت نتایج در خصوص نمونه‌های چوله نیز بررسی می‌شود. در جدول(۷) تعدادی از نتایج به‌دست‌آمده از دو روش در نمونه‌های تصادفی حاصل از نمونه‌گیری بازگردان تکراری آورده می‌شود.

جدول ۷. نتایج به‌دست‌آمده از روش نمونه‌گیری بازگردان تکراری برای نمونه‌های تصادفی با انتخاب جایگزین

دقت پیش‌بینی(درصد)		پیش‌بینی صحیح		کلاس	حجم نمونه	
روش CART	مدل لاجیت	روش CART	مدل لاجیت		روش CART	مدل لاجیت
۹۱/۹	۸۶/۶	۱۱۳	۷۴	۰	۲۰۰	۱۵۰
۹۳/۵	۸۹/۲	۷۲	۵۸	۱		
۸۴/۷	۸۵/۹	۱۲۴	۷۳	۰	۲۰۰	۱۴۰
۹۴/۷	۸۹/۱	۷۶	۴۹	۱		
۹۱/۸	۸۸/۰	۱۲۲	۶۶	۰	۲۰۰	۱۴۵
۱۰۰/۰	۸۱/۴	۷۸	۵۷	۱		
۹۲/۲	۹۲/۴	۱۲۸	۷۳	۰	۲۰۰	۱۴۶
۹۳/۱	۸۵/۱	۷۲	۵۷	۱		
۸۹/۱	۸۶/۲	۱۲۸	۷۵	۰	۲۰۰	۱۳۹
۸۶/۱	۸۴/۶	۷۲	۴۴	۱		
۸۹/۸	۸۲/۸	۱۱۸	۷۲	۰	۲۰۰	۱۵۰
۹۶/۳	۸۲/۵	۸۲	۵۲	۱		

همان‌طور که مشاهده می‌شود در تمام نمونه‌ها، روش درختهای طبقه‌بندی، برتری محسوسی از لحاظ دقت طبقه‌بندی دارد. یکی از دلایل آن وجود تعداد زیاد مشاهداتی است که برای آنها مقداری برای متغیر "شغل همسر" وجود ندارد. این امر را در تعداد مشاهدات برای هر روش می‌توان ملاحظه نمود. دلیل دیگر، دقت بیشتر روش درختهای طبقه‌بندی، حذف مشاهدات پراکنده است که در روش لاجیت امکان آن وجود ندارد.

۵- نتیجه‌گیری

آنچه دلیل جستجو برای یافتن روشی جایگزین برای مدل لاجیت است، ماهیت اطلاعات موجود در مورد مشتریان است. در طی تحقیق نشان داده شد که در حال حاضر اطلاعات موجود در بانکها به صورت صفتهای کیفی است که اغلب حالتی دوگانه دارند. بنابراین روشی که از ابتدا و به‌صورت ذاتی بتواند مشتریان را طبقه‌بندی نماید، برترخواهد بود.

روشهای پارامتریک(همچون لاجیت) در برابر روشهای غیرپارامتریک دارای برخی از ایرادات است. یکی از مهمترین ایرادات روشهای آماری و رگرسیونی، این است که آنها دارای فروض قوی و محدودکننده هستند. برای نمونه در روش لاجیت، رابطه بین ترکیب خطی متغیرهای مستقل و متغیر وابسته از یک تابع سیگموئید پیروی می‌کند. روشهای غیرپارامتریک مثل *CART* از این ایراد مستثنا هستند.

در این تحقیق ۲۰۰ مشاهده از مشتریان حقیقی بانک، به‌طور تصادفی به‌منظور مدل‌سازی مورد استفاده قرار گرفت. تعداد ۱۳۶ مشاهده، در دسته مشتریان خوش‌حساب و ۶۴ مشاهده، در گروه مشتریان بدحساب قرار گرفت. برای ساختن مدل لاجیت از روش انتخاب رو به جلو و همچنین حذف رو به عقب و برای ایجاد درخت تصمیم از روش جینی و نرم‌افزار *CART* استفاده نمودیم. مدل لاجیت نهایی شامل ۶ متغیر توضیحی است. اثرات نهایی متغیرهای توضیحی نشان می‌دهد که اگر بانک بخواهد بر اساس این مدل به مشتریان خود وام دهد، ابتدا می‌باید به متغیرهای با اثرات نهایی بالا مانند سطح تحصیلات، تعداد فرزندان و وضعیت مسکن توجه بیشتری داشته باشد، چرا که تأثیر بیشتری در قصور یا عدم‌قصور آنان دارد.

روش دومی که در این تحقیق استفاده گردید، روش *CART* است. در این روش، درخت نهایی بر اساس دقّت آن به‌ویژه در نمونه آزمون و همچنین رعایت میزان بهینه خطا و تعداد گره‌های نهایی انتخاب گردید. در این درخت با توجه به اهمیت متغیرها، در تفکیک و کلاس‌بندی، ۸ متغیر نهایی تفکیکی قرار دارد. ضمن اینکه اهمیت عواملی چون سطح تحصیلات، شغل همسر و نوع کار از سایر عوامل بیشتر است.

مقایسه مدل لاجیت با روش *CART* نشان می‌دهد که برای تمام نمونه‌ها دقّت پیش‌بینی روش *CART* بیشتر است. که این امر بیانگر مزیّت مدلهای غیرپارامتریک(همچون *CART*) نسبت به مدلهای آماری و پارامتریک(همچون لاجیت) خواهد بود.

نگاهی کلی به ساختار *CART* روشن‌کنندهٔ برخی برتریهای این روش بر مدل لاجیت است. اول اینکه با این روش از ابتدا عمل دسته‌بندی انجام می‌شود و نیازی نیست که مانند مدل لاجیت ضرایب را به‌دست آورده و سپس کار طبقه‌بندی را انجام دهیم. دوم اینکه در مدل لاجیت فقط دو

گروه خوب و بد را از هم تمیز می‌دهیم و تعیین گروه‌هایی که مشاهدات مربوط به هر گروه بیشترین شباهت را به هم داشته باشند بسیار مشکل است، زیرا مشتریان را فقط بر اساس احتمال نکول به‌دست‌آمده می‌توان تقسیم‌بندی نمود؛ در حالی که در اینجا هشت گروه تفکیک‌شده وجود دارد که در هر کدام مشاهدات مربوط به کلاسی قرار گرفته‌اند که بیشترین شباهت را به یکدیگر دارند. مورد سومی که می‌توان ذکر کرد استفاده از ۸ متغیر تفکیک‌کننده و در کل استفاده از تمام متغیرهاست در حالی که در مدل لاجیت دو متغیر سن و سطح تحصیلات همسر بی‌معنی تشخیص داده شده و از مدل نهایی حذف شدند. مورد چهارم مربوط به استفاده از تمام مشاهدات است. همان‌طور که در فصل پیش دیدیم متغیر شغل همسر به‌صورتی بود که برای برخی مشاهدات موجود بود و برای سایر مشاهدات وجود نداشت که این امر باعث ایجاد مشکل در استفاده از مدل می‌شد. اما در روش درختهای طبقه‌بندی این متغیر هم با اهمیت تشخیص داده شده و هم اینکه در مدل مورد استفاده قرار می‌گیرد.

فهرست منابع

سبزواری، حسن(۱۳۸۴) مقایسه‌ای بین روش پارامتریک لاجیت و مدل ناپارامتریک AHP برای ارزیابی اعتباری مشتریان حقوقی بانک پارسیان؛ پایان‌نامهٔ کارشناسی ارشد، دانشگاه صنعتی شریف.

جمشیدی، سعید(۱۳۸۳) شیوه‌های اعتبارسنجی مشتریان؛ پژوهشکدهٔ پولی و بانکی.

کشاورزحداد، غلامرضا(۵-۲۰۰۴) ویژگی‌های کوچک نمونه‌ای آماره والد-کاربردی از تکنیک نمونه‌گیری بازگردان تکراری و شبیه‌سازی مونت کارلو؛ مجموعه مقالات پژوهشی دانشگاه صنعتی شریف، معاونت پژوهش و فناوری.

Altman, E. I. (1993) Corporate Financial Distress and Bankruptcy: A Complete Guide to Predicting & Avoiding Distress and Profiting from Bankruptcy; *2nd Edition, New York.*

Altman, E., (1968) Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy; *Journal of Finance*, September, pp. 189-209.

Altman, E., A. Agarwal, and F. Varetto (1994) Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks, *Journal of Banking and Finance* 18.

Arminger, G.D., Enach and T. Bonne (1997) Analyzing Credit Risk Data: a Comparison of Logistic Discrimination, Classification Tree Analysis and Feed Forward Network; *Computational Statistics 12, 293-310.*

Barbro Back, Teija Laitinen, Kaisa Sere & Michiel Van Wezel (1996) Choosing Bankruptcy Predictors Using Discriminant Analysis, Logit Analysis, and Genetic Algorithms; *Turku Centre for Computer Science, Technical Report No. 40, September.*

Breiman L., Frydman H., Olshen R.A., and Stone C.J., (1984) *Classification and Regression Trees*, Chapman and Hall, New York, London.

Carter, C. and Catlett, J. (1987) Assessing Credit Card Application Using Machine Learning; *IEEE Expert, Vol. 2, No. 3, pp. 71–79.*

Cramer, JS (2001) An Introduction to the Logit Model for Economists, 2nd ed., Timberlake Consultants Ltd.

Dean Caire and Robert Kossmann(2003) Credit Scoring: Is It Right for Your Bank?; *February, Bannock Consulting.*

Dionne, G., M. Artis and M. Gulien(1996) Count Data Models for a Credit Scoring Model; *Journal of Empirical Finance 3, 303-325.*

Eisenbeis, Robert, A. (1996) Recent Developments in the Application of Credit Scoring Techniques to the Evaluation of Commercial Loans; *In: IMA Journal of Mathematics Applied in Business and Industry, 7, P. 271-290.*

Eivind Bernhardsen (2001) A Model of Bankruptcy Prediction; *Norges Bank, December 5.*

Elaine Fortowsky & Michael LaCour-Little (2001) Credit Scoring and Disparate Impact; *Wells Fargo Home Mortgage Dec 13.*

Fair, Isaac and Company (1990) Incorporated Annual Report; *San Rafael, CA*.

Fair, Isaacs and Company. http://www.fairisaac.com, 06.12.2001.

Feelders, A. J.,le Loux, A. J. F. & Zand, J. W. (1995) Data Mining for Loan Evaluation at ABN AMRO: A Case Study; *Fayyad, U. M. & Uthurusamy, R., (eds.), Proceedings of the First International Conference on Knowledge Discovery & Data Mining, P. 106-111.*

Fujita, M. and Tamai, T. (1987) Credit Card Application Assessment with a Profiling System, Proc. *5th Knowledge Engineering Symposium, The Society of Instrument and Control Engineers, pp. 135–140, in Japanese.*

Hand D.J. and Henley W.E. (1997) Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society, Series A*, 160, 523-541.

Henley, W.E. and D.J. Hand (1996) A K-Nearest-Neighbor Classifier for Assessing Consumer Credit Risk; *The Statistician, 45(1), 77-95.*

Johnson, R. W. (1992) Legal, Social, and Economic Issues in Implementing Scoring in the US; *In: Thomas, L. C., Crook, J. N. & Edelman, D. B. (eds.), Proceedings of the IMA Conference on Credit Scoring and Credit Control, Clarendon Press, Oxford, P.19-32.*

Jost, A. (1998) Data Mining; In: Mays, E. (eds.), Credit Risk Modeling: Design and Application, *Glenlake Publishers, Chicago, P. 128-147.*

Kevin, J. Leonard (1995) The Development of Credit Scoring Quality Measures for Consumer Credit Applications; *International Journal of Quality & Reliability Management, Vol. 12, No. 4, pp. 79-85.*

Koch Timothy, W. & Macdonald Scott, S. (2002) Bank Management; *Fifth Edition.*

Lee, Tian-Shyug, Chih-Chou Chiu, Yu-Chao Chou and Chi-Jie Lu (2006) Mining the Customer Credit Using Classification and Regression Tree and Multivariate Adaptive Regression Splines;*Computational Statistics & Data Analysis, Vol. 50, No. 4.*

Leea, T.Sh. , Ch.Ch. Chiub, Ch. J. Luc and I.F. Chend (2002) Credit Scoring Using the Hybrid Neural Discriminant Technique; *Expert Systems with Applications, 245–254.*

Liu, Yang (2001) New Issues in Credit Scoring Application; *Arbeitsbericht No. 16, Hrsg.: Matthias Schumann.*

Lo, A. W., (1986) Logit Versus Discriminant Analysis: A Specification Test and Application to Corporate Bankruptcies, *Journal of Econometrics* 31, 151–178.

Loretta, J. Mester (1997) What's the Point of Credit Scoring?; *Business Review September/October.*

Lundy, M. (1992) Cluster Analysis in Credit Scoring; *Thomas, L. C.,Crook, J. N. & Edelman, D. B. (eds.), Proceedings of the IMA Conference on Credit Scoring and Credit Control. Clarendon Press, Oxford, P. 92-107.*

Maddala, G.S. (1983) Limited Dependent and Qualitative Variables in Econometrics, Cambridge University Press, Cambridge, England.

Quinlan, R. (1983) Learning Efficient Classification Procedures and Their Application to the Chess End Games, in Machine Learning: An Artificial Intelligence Approach; *Tioga Publishing Co., pp. 436–482.*

Reichert, A.K., Cho, C.C. and Wagner,G.M. (1983) An Examination of the Conceptual Issues Involved in Developing Credit Scoring Models; *Journal of Business Economic Statistics, Vol. 1, No.2, pp. 101-104.*

Rose Peter, S. (2002) Commercial Bank Management; *Mc Graw-Hill Higher Education, International Edition.*

Rosenberg, E. & Gleit, A. (1994) Quantitative Methods in Credit Management: A Survey; *Operations Research, Vol. 42, No.4, 589-613.*

Roszbach, Kasper (1998) Bank Lending Policy, Credit Scoring and the Survival of the Loan; *28 September, Doctoral Thesis School of Business and Economics, Stokholm Uni.*

Stephen, A. Hillegeist, Elizabeth K. Keating, Donald P. Cram and Kyle G. Lundstedt (2003) Assessing the Probability of Bankruptcy; *September.*

Stephen, Kealhofer (2003) Quantifying Credit Risk I: Default Prediction; *January/February, Financial Analysts Journal.*

Stephen, Kealhofer (2003) Quantifying Credit Risk II: Debt Valuation; *May/June, Financial Analysts Journal.*

Thomas, H. Stanton Fellow (1999) Credit Scoring and Loan Scoring: Tools for Improved Management; *Grant Repor, July.*

Thomas, L. C. (2000) A Survey of Credit and Behavioural Scoring: Forcasting Financial Risk of Lending to Consumers; *International Journal of Forecasting 16, 149-172.*

Timofeev R. (2004) Classification and Regression Trees (CART) Theory and Applications; Master Thesis, Center of Applied Statistics and Economics Humboldt University, December 20.

Vladimir Bugera, Hiroshi Konno and Stanislav Uryasev (2002) Credit Cards Scoring With Quadratic Utility Function; *University of Florida, January, 15.*

Yang Liu (2002) A Framework of Data Mining Application Process for Credit Scoring; *Arbeitsbericht, No. 01.*

Yang Liu (2002) The Evaluation of Classification Models for Credit Scoring; *Arbeitsbericht, No. 02.*