چکیده

در بین ریسک‌هایی که بانک با آن مواجه است، ریسک اعتباری از اهمیت ویژه‌ای برخوردار است. یکی از راه‌های کمی‌کردن و اندازه‌گیری ریسک اعتباری و در نتیجه مدیریت مناسب آن، استفاده از مدل‌های امتیازدهی اعتباری (CS) است. مدل CS بر اساس معیارهای کمی (مانند اطلاعات مالی شرکت) و نیز معیارهای غیرکمی (مثل بخش اقتصادی آن)، ویژگی‌ها و عمل‌کرد وام‌های قبلی را مدل-سازی می‌نماید تا عمل‌کرد آتی وام‌هایی با مشخصات مشابه را پیش‌بینی کند. در CS یک نمره به هر مشتری اختصاص داده می‌شود. این نمره به‌عنوان شاخصی از ریسک مشتریان شناخته می‌شود. با مقایسه این نمره با نمره مرزی، مشتریان پرریسک و کم‌ریسک از هم‌دیگر مجزا می‌شوند. در این تحقیق به‌منظور ارزیابی مشتریان حقوقی بانک کارآفرین از مدل امتیازدهی لاجیت و روش غیرپارامتریک CART استفاده می‌گردد، سپس این دو مدل از نظر میزان دقت در پیش‌بینی مشتریان خوب و بد با یک‌دیگر مقایسه می‌شوند. مدل‌های ساخته‌شده برای نمونه‌های تصادفی با حجم اندک مشاهدات مورد آزمایش قرار گرفته و دقت آنها در تفکیک مشتریان به صورت خوش‌حساب یا بدحساب مورد ارزیابی قرار می‌گیرد.

کلمات کلیدی: امتیازدهی اعتباری[1]، ریسک اعتباری[2]، قصور[3]، لاجیت[4] و طبقه‌بندی و رگرسیون درختی[5] (CART).

[1] credit scoring
[2] credit risk
[3] default
[4] logit
[5] classification and regression tree (CART)

مقدمه

ارایه‌ی تسهیلات مالی یکی از فعالیت‌های مهم نظام بانکی تلقی می‌شود. برای اعطای تسهیلات، باید درجه اعتبار و قدرت گیرنده‌ی تسهیلات را در بازپرداخت اصل و سود تسهیلات اعطایی تعیین نمود. احتمال عدم بازگشت اصل و سود تسهیلات اعطایی را ریسک اعتباری گویند. بحران‌های مشاهده شده در نظام بانکی کشورها عمدتاً ناشی از عدم کارآیی در مدیریت ریسک اعتباری بوده است. یکی از مهم‌ترین ابزارهایی که بانک‌ها، برای مدیریت و کنترل ریسک اعتباری بدان نیازمندند، سیستم امتیازدهی اعتباری مشتریان[1] (CS) است. از طرف دیگر ورود مؤسسات اعتباری به عرصه‌ی رقابت داخلی و جهانی و رویارویی با حجم گسترده‌ی تقاضا برای اعتبار، فرصت‌ها و تهدیدات جدیدی را برای آنها ایجاد نموده است. لذا شاهد توسعه رو به تزاید نقش تکنولوژی در فراگرد مدیریت اعتبار مؤسسات بانکی و نهادهای مالی هستیم[2].

مدل‌های امتیازدهی اعتباری (CS) اثر بخشی تصمیمات اعتباری را در تولید خدمت و رفع نیازهای مشتریان افزایش داده و موجبات کاهش محسوس هزینه‌ها و قصور وام‌گیرندگان را فراهم ساخته است. برای بانک که اعطاکننده اعتبار است مساله‌ی اساسی، تعیین سطح ریسک اعتبارگیرندگان و مدیریت اعتبارات است. لذا مدیران بانکی بایستی شناخت مناسبی نسبت به تعیین، اندازه‌گیری، نظارت و کنترل ریسک اعتباری و نیز تعیین ذخیره‌ی سرمایه[3] مناسب با توجه به ریسک آنها داشته باشند. بنابراین بانک‌ها بایستی ابزار مناسبی جهت اندازه‌گیری ریسک اعتباری مشتریان خود، طراحی کنند که این ابزار چیزی جز کمی‌کردن ریسک اعتباری از طریق روش امتیازدهی اعتباری نیست.

هدف از این تحقیق آن است که با بهره‌گیری از روش‌های پذیرفته شده‌ی مالی[4] همچون مدل لاجیت و روش طبقه‌بندی و رگرسیون درختی یا CART، مدل‌هایی ارایه کند که با دریافت اطلاعات مورد نیاز در مورد مشتری درخواست‌کننده‌ی وام، معیار

[1] credit scoring
[2] Liu (2001)
[3] capital reservation
[4] finance

کمی جهت قبول یا رد درخواست اعتباری تعیین گردد. با تعریف متغیرهای کمی راجع به خصوصیات وام‌گیرندگان در مدل فوق، می‌توان به ارزیابی عینی و واقع‌بینانه‌تری در مورد وضعیت اعتباری آنها دست یافت و تا حد امکان از برداشت‌های ذهنی و توأم با خطا پرهیز کرد.

در این مقاله ابتدا به تعریف برخی از مفاهیم اساسی استفاده‌شده در مدل‌های امتیازدهی اعتباری خواهیم پرداخت، سپس به مطالعه تجربی و روش تحقیق اشاره خواهد شد. همچنین چگونگی استفاده متغیرها در مدل‌های امتیازدهی اعتباری توضیح داده می‌شود. به این صورت که به معرفی کامل چگونگی انتخاب مشاهدات و داده‌ها و فراگرد انتخاب متغیرها خواهیم پرداخت. سپس بهترین مدل ارزیابی ریسک اعتباری معرفی گردیده و قدرت تمایز سیستم رتبه‌بندی اندازه‌گیری می‌شود. بخش بعد به بررسی روش دوم امتیازدهی یعنی CART اختصاص خواهد داشت و در نهایت این دو روش از نظر میزان دقت در پیش‌بینی مشتریان مورد ارزیابی قرار می‌گیرند.

۱- مروری بر ادبیات موضوع؛ مبانی نظری

۱-۱- تعریف امتیازدهی اعتباری (CS)

صاحب‌نظران گوناگون تعاریف نسبتاً مشابهی درباره CS ارایه کرده‌اند. فلدمن[1]، CS را فراگرد تخصیص یک معیار کمّی به‌صورت واحد یا نمره، به یک قرض‌گیرنده بالقوه، جهت ارایه‌ی تخمینی از عمل‌کرد آتی وی می‌داند[2]. به گفته‌ی استاندارد و پورز[3]: «سیستم رتبه‌بندی اظهارنظر در مورد ارزش اعتباری یک بده‌کار با توجه به اوراق بدهی یا سایر تعهدات مالی وی بر اساس عوامل ریسک است». به‌نظر مودیز[4] رتبه‌بندی عبارت از «اظهارنظر در مورد توانایی آتی بده‌کار و تعهد حقوقی صادرکننده‌ی[5] آن برای انجام پرداخت‌های به‌موقع اصل و بهره، روی یک اوراق بهادار با درآمد ثابت مشخص» است.[6]

[1] Feldman
[2] Khavein, Frame, J. White (2001)
[3] Standard & Poors
[4] Moody's
[5] issuer
[6] Sinkey (1992)

۱-۲- تعریف وام بد و قصور[1]

یکی از نکات و مسائل اساسی در ساخت یک مدل ریسک اعتباری تعریف قصور است که مدل براساس آن شکل می‌گیرد. با توجه به اهمیت موضوع، کمیته بال[2]، تعاریف زیر را برای قصور ارایه می‌دهد:

قصور به حالتی اطلاق می‌شود که برای بده‌کار، یک یا بیشتر از یکی از حوادث زیر اتفاق افتاده باشد.

- در توانایی بده‌کار برای پرداخت تعهداتش شامل اصل، بهره یا کارمزد تردید وجود دارد.

- بده‌کار بیش از ۹۰ روز از هرگونه تعهد اعتباری خویش، سپری کرده باشد.

- بده‌کار با تشکیل پرونده، اعلام ورشکستگی کرده باشد.[3]

۱-۳- افق زمانی

یکی دیگر از مفاهیم اساسی و بسیار کلیدی در مدیریت ریسک اعتباری تعیین افق زمانی است. همچنان‌که کمیته بال توضیح می‌دهد، بیشتر بانک‌ها بر اساس عادت، افق زمانی یک‌ساله را در فراگرد مدل‌سازی ریسک اعتباری مورد استفاده قرار می‌دهند. دلیل آنان برای استفاده از این نگرش آن است که افق یک‌ساله به بهترین وجه دوره‌ای را که طی آن موارد ذیل اتفاق می‌افتد را در بر می‌گیرد:

الف- زیان حاصل از ریسک مشتری در پرتفولیو اعتباری به دوره بعد منتقل نمی‌شود.

ب- اطلاعات در مورد بده‌کاران جدید فاش می‌شود.

پ- داده‌های مربوط به قصور منتشر می‌شود.

ت- بودجه‌ریزی داخلی و سرمایه و صورت‌های مالی سهام‌داران تهیه می‌شود.

ث- تمدید اعتبارات به‌صورت معمول مورد ارزیابی مجدد قرار می‌گیرد.

همان‌طورکه گفته شد، بسیاری از نهادها از افق زمانی یک‌ساله استفاده می‌کنند تا اکسپوژر ریسک اعتباری[4] را اندازه بگیرند. به‌نظر می‌رسد افق زمانی برای ارزیابی

---

[1] bad loan and default
[2] Basel committee
[3] Hayden (2003)
[4] credit risk exposure

دقت مدل‌ها و نیز جهت جواب‌گویی به نیازهای قانونی و اقتصادی یک متغیر مهم است. توانایی مدل‌های ارزیابی اعتباری با توجه به طبیعت دوگانه‌شان، شدیداً به طول افق زمانی حساس است.[۱] لیکن افق زمانی بلندمدت نیز به‌ویژه موقعی‌که تصمیمات در مورد تخصیص وام‌های جدید اتخاذ می‌گردد مورد علاقه است.

۱-۴- روش‌های برآورد مدل امتیازدهی

مدل‌های مورد استفاده برای CS به دو گروه اصلی تقسیم‌بندی می‌شوند:

گروه اول: مدل‌های CS پارامتریک

الف- مدل احتمال خطی[۲]

ب- مدل‌های لاجیت و پروبیت[۳]

ج- مدل تحلیل ممیزی

د- شبکه‌های عصبی[۷] (NN)

گروه دوم: مدل‌های CS غیرپارامتریک

الف- برنامه ریزی خطی

ب- درخت‌های رده‌بندی[۴] (الگوریتم‌های افراز بازگشتی[۵])

ج- مدل نزدیک‌ترین همسایه‌ها[۶]

د- فراگرد تجزیه و تحلیل سلسله مراتبی[۸] (AHP)

هـ- سیستم‌های خبره[۹] (ES)

کلاس دیگری از روش‌های هوش مصنوعی نیز به‌طور گسترده در سیستم‌های امتیازدهی مورد استفاده قرار گرفته که الگوریتم‌های ژنتیک[۱۰] نامیده شده است[۱۱]. در این تحقیق با توجه به محدود بودن حجم اطلاعات در دسترس در مورد مشتریان حقوقی بانک کارآفرین و با استناد به نتایج مطالعات سایر کشورها از دو روش رگرسیونی لاجیت و درخت‌های رده‌بندی (که آن‌را CART می‌نامیم) استفاده می‌شود.

[1] Gallati (2003)
[2] linear probability model
[3] logit and probit model
[4] classification trees
[5] recursive portioning algorithms
[6] k-nearest neighbors
[7] neural network
[8] analytical hierarchy process
[9] expert systems
[10] genetic algorithms
[11] Kiss (2003)

۱-۵-مرور ادبیات تجربی روش CART در امتیازدهی اعتباری

روش رگرسیونی لاجیت یکی از روش‌های مدل‌سازی آماری است که به دلیل قابلیت کاربرد گسترده آن در زمینه‌های گوناگون و به‌ویژه در اعتبار سنجی مشتریان برای خواننده شناخته شده‌تر است. لیکن به دلیل جدیدتر بودن روش CART، مناسب است نظریه و همچنین کاربرد این روش را در مبحث امتیازدهی اعتباری توضیح دهیم[۱].

روش درخت‌های طبقه‌بندی و رگرسیونی[۲] که به اختصار آن را CART می‌نامیم در دهه‌ی ۸۰ توسط بریمن، فریدمن، السون، استن[۳]، در مقاله‌ای تحت عنوان «درخت‌های طبقه‌بندی و رگرسیونی» در سال ۱۹۸۴ منتشر شد. در این روش برای ساخت یک درخت، تصمیم به نمونه‌ای احتیاج داریم که نمونه‌ی آموزشی نامیده می‌شود. نمونه‌ی آموزشی برای سیستم‌های امتیازدهی به‌صورت تصادفی از بین اطلاعات مشتریان گذشته انتخاب می‌شود. هر چه این اطلاعات از پراکندگی کمتری برخوردار باشند، مدل به‌دست آمده مشتریان جدید را به‌صورت دقیق‌تری طبقه‌بندی می‌کند.

مدل CART یا مدل‌های طبقه‌بندی شبیه آن توسط بسیاری از محققان دیگر با موفقیت مورد استفاده قرار گرفته است. فریدمن، آلتمن و کائو[۴] مسأله طبقه‌بندی شرکت‌ها را به‌وسیله این روش و DA[۵] بررسی نمودند. مارایز، پاتل و والفسون[۶] قابلیت استفاده درخت‌های طبقه‌بندی و مدل‌های لاجیت را در وام‌دهی تجاری[۷] مطالعه نمودند. کیم و سرینویاسن[۸] کاربرد این روش را در وام‌دهی صنعتی با مدل لاجیت و همچنین با MDA[۹] مقایسه نمودند. اخیراً هم لی، چیو، چائو و لو[۱۰] (۲۰۰۴) نتایج مقایسه‌ای بین شبکه‌های عصبی و CART را ارایه کرده‌اند و نشان دادند هنگامی که حجم داده‌ها و تعداد متغیرها زیاد است روش CART راه حل ساده‌تر و کاراتری برای

---

[۱] تئوری مربوط به روش CART در بخش دوم مطالعه تجربی به تفصیل توضیح داده شده است.

[2] classification and regression trees
[3] Breiman, Freidman, Olshen, Stone
[4] Frydman, Altman, and Kao
[5] discriminant analysis
[6] Walfson Marias and Patel
[7] commercial lending
[8] Srinivasan and Kim
[9] multivariate discriminant analysis
[10] Tian-Shyug Lee, Chih-Chou Chiu, Yu-Chao Chou and Chi-Jie Lu

حل مسائل ارایه می‌کند. همه این مطالعات به اتفاق عنوان داشتند که روش طبقه-بندی درختی، جداسازی بهتری از سایر روش‌ها فراهم می‌آورد. ایشان این عملکرد را به ماهیت غیر پارامتریک بودن روش طبقه‌بندی درختی نسبت می‌دهند[1].

۲- روش تحقیق

۲-۱- مدل لاجیت

همان‌طورکه گفته شد، چند روش رگرسیونی برای ایجاد مدل‌های امتیازدهی وجود دارد. لیکن برخی از آنها (مثل احتمال خطی) در زمینه تئوریک و در عمل برای به‌کارگیری در CS دارای مشکلاتی هستند. برای حل مشکلات مربوط به مدل احتمال خطی، محققین به جستجوی گزینه‌های جای‌گزین پرداختند. در این مدل ممکن است که احتمال برآورد شده، خارج از بازه [0,1] باشد و لذا باید یک تبدیل مناسب یافت که وقوع احتمال در این بازه را حتمی سازد. تعداد زیادی از محققین رگرسیون لاجیت را ترجیح می‌دهند زیرا این روش از دقت مناسبی برخوردار است. در این مدل، متغیر وابسته یک متغیر دوانتخابی یا دوگانه[2] است[3].

۲-۱-۱- مشاهدات و متغیرها

از ۴۴۸ مشاهده‌ای که به‌طور تصادفی از مشتریان حقوقی بانک کارآفرین مورد استفاده قرارگرفت تعداد ۳۶۲ مشاهده در دسته‌ی مشتریان خوش‌حساب و ۸۶ مشاهده در گروه مشتریان بدحساب قرار داشت[4]. لازم به توضیح است که این حجم، نمونه‌ی مربوط به مشتریان بانک کارآفرین بین سال‌های (۱۳۷۸ تا ۱۳۸۴ه.ش) است که اقدام به اخذ تسهیلات از بانک کارآفرین نموده‌اند. مبلغ تسهیلات این مشتریان از حد شعب بانک خارج بوده و به این منظور تصمیم‌گیری در مورد آنها به‌صورت متمرکز و در اداره اعتبارات بانک صورت گرفته است. تعداد مشاهداتی که وارد بانک اطلاعاتی شد حدود ۱۰۶۸ پرونده اعتباری بود. لیکن به دلیل ناقص بودن برخی از این پرونده‌ها (به‌ویژه نبود

---

[1] Kiss (2003)
[2] binary
[3] http://www.cs.uk.n1/docs/vakken/ida/idahc8.pdf
[۴] منظور از مشتریان خوش‌حساب مشتریانی است که (طبق تعریف بال۲) یا هیچ‌گونه تأخیری در پرداخت اقساط خود نداشته‌اند و یا حداکثر ۳ ماه تأخیر دارند. درحالی‌که مشتری بد حساب دارای حداقل ۳ ماه تأخیر است.

ترازنامه و صورت سود و زیان سال قبل تاریخ اخذ تسهیلات آنها)، تنها ۴۴۸ مشاهد وارد مدل‌سازی و تحلیل شد. این تعداد مشاهده دربرگیرنده متغیرهای رگرسیون یا متغیرهای توضیحی همچون نسبت‌های مالی و اقلام مهّم مندرج در صورت‌های مالی (یعنی ترازنامه و صورت سود و زیان) همچنین ویژگی‌های پایه‌ای مشتری است. این متغیرها شامل ۳۳ نسبت مالی و ۱۶ متغیر مربوط به اطلاعات پایه‌ای غیرترازنامه‌ای است. بر اساس روش گام به گام[1] (که قادر به انجام روش انتخاب رو به جلو و حذف رو به عقب[2] است) متغیرهای مستقل وارد رگرسیون لاجیت شدند.

با تحلیل آماره‌هایی که در مدل رگرسیونی محاسبه شده‌اند یکی از بهترین مدل‌های تخمین زده شده، انتخاب شد که شامل متغیر وابسته عمل‌کرد مشتری (خوش‌حساب یا بد حساب بودن) و یازده متغیر مستقل، «سابقه عملکرد مشتری در سیستم بانکی»، «وضعیت سوددهی آن»، «دارایی به شاخص ضمنی[3]»، «فروش به دارایی»، «حساب‌های دریافتنی به بدهی‌ها»، «دارایی سریع به فروش»، «موجودی نقد به فروش»، «سرمایه در گردش به فروش»، «دوره وام»، «سابقه‌ی ارتباط با بانک کارآفرین» و «بدهی بانکی به کل» بدهی است.

[1] stepwise procedure
[2] forward selection and backward elimination
[3] GDP deflator

جدول ۱: نسبت‌های مالی برگزیده برای استفاده در مدل امتیازدهی اعتباری[1]

ردیف	نسبت حساب‌داری	عامل ریسک	فرضیه	علامت اختصاری
۱	بدهی‌ها به دارایی‌ها	اهرمی	+	DET.AST
۲	حقوق صاحبان سهام به دارایی‌ها	اهرمی	−	EQU.AST
۳	بدهی‌های بلندمدت به دارایی‌ها	اهرمی	+	LDET.AST
۴	بدهی بانکی به دارایی‌ها	اهرمی	+	BDET.AST
۵	بدهی بانکی به (دارایی‌ها منهای بدهی بانکی)	اهرمی	+	BDET
۶	بدهی بانکی به بدهی‌ها	اهرمی	+	BDET.DET
۷	دارایی‌های جاری به بدهی‌های جاری	نقدینگی	−	LIQU
۸	دارایی‌های جاری به بدهی‌ها	نقدینگی	−	Cast.totD
۹	سرمایه در گردش به دارایی‌ها	نقدینگی	−	Wcap.Ast
۱۰	موجودی نقد به دارایی‌ها	نقدینگی	−	Csh.Ast
۱۱	سرمایه در گردش به فروش خالص	نقدینگی	−/+	Wcap.Sal
۱۲	موجودی نقد به فروش خالص	نقدینگی	−/+	Csh.Sal
۱۳	دارایی‌های جاری به خالص به فروش	نقدینگی	−/+	Cast.Sal
۱۴	دارایی‌های سریع به خالص به-فروش	نقدینگی	−/+	quAst.Sal
۱۵	بدهی بانکی کوتاه‌مدت به بدهی بانکی	نقدینگی	−	sh.lng.Det
۱۶	موجودی نقد به بدهی‌های جاری	نقدینگی	−	csh.cDet
۱۷	سرمایه در گردش به بدهی‌های جاری	نقدینگی	−	Wcap.Cdet
۱۸	نسبت سریع	نقدینگی	−	Qui
۱۹	حساب‌های دریافتنی به بدهی‌ها	فعالیت	−	rAcc.Det
۲۰	حساب‌های پرداختنی به فروش خالص	فعالیت	+	pAcc.Sal
۲۱	حساب‌های دریافتنی به فروش	فعالیت	−	rAcc.Sal
۲۲	فروش خالص به دارایی‌ها	گردش	−	Sal.Ast
۲۳	سود عملیاتی به دارایی‌ها	گردش	−	Opr.Ast
۲۴	EBIT به فروش خالص	سوددهی	−	EBT.sal
۲۵	سود ناخالص به سود عملیاتی	سوددهی	−	Gprof.Ast
۲۶	(EBIT+ سود بهره) به دارایی‌ها	سوددهی	−	Prof.ast
۲۷	خالص سود به فروش خالص	سوددهی	−	Prof.Sal
۲۸	خالص سود به سود عملیاتی	سوددهی	−	Prof.Opr
۲۹	سود انباشته به دارایی‌ها	سوددهی	−	Aprof.Ast
۳۰	بازده حقوق صاحبان سهام	سوددهی	−	ROE
۳۱	دارایی‌ها به شاخص قیمت مصرف-کننده	اندازه	−	Ast.CPI
۳۲	فروش خالص به شاخص قیمت مصرف-کننده	اندازه	−	Sel.CPI
۳۳	لگاریتم دارایی‌ها به شاخص ضمنی	−	−	Ast.Gdp

[1] در این جدول ستون عامل ریسک گروه نسبت‌های مالی و ستون فرضیه بیان‌گر رابطه‌ی نظری انتظاری بین نسبت‌های مالی و نکول است

جدول ۲: متغیرهای کمی و غیرکمی برای استفاده در مدل امتیازدهی اعتباری

ردیف	متغیر	علامت اختصاری
۱	مجموع آخرین بدهی و تعهدات به سیستم	TOT.DET
۲	گردش حساب بستان‌کار ۶ ماهه	SIX.TR
۳	مانده‌ی حساب جاری	CURR
۴	تعداد مجوز	LIC
۵	کد زمینه فعالیت	COD
۶	فرم حقوقی	LEG.TYP
۷	کد بخش اقتصادی	ECN.PAR
۸	تعداد کارکنان	STFF
۹	جمع مساحت استیجاری	RENT
۱۰	جمع مساحت ملکی	OWN
۱۱	سابقه شرکت	E.YER
۱۲	سابقه ارتباط با بانک کارآفرین	R.YER
۱۳	طول دوره وام (ماه)	L.PERD
۱۴	سابقه‌ی کار(سال)	HIST.JOB
۱۵	سابقه‌ی عمل‌کرد در سیستم بانکی	HIST.COD
۱۶	وضعیت سوددهی	PROF.COD

۲-۱-۲- تصفیه نسبت‌ها و متغیرهای مالی کاندید برای دست‌یابی به متغیرهای اصلی

از آنجایی که بسیاری از متغیرهایی که به‌عنوان کاندید در نظر گرفته شده‌اند، از صورت‌های اصلی مالی و اطلاعات پایه‌ای آن استخراج می‌شوند ممکن است به‌صورت دو به دو با همدیگر همبستگی داشته باشند. اگر این متغیرها در رگرسیون وارد شوند به دلیل وجود هم‌خطی باعث بی‌معنی‌شدن سایر ضرایب از طریق بالارفتن واریانس کوواریانس بین ضرایب و در نتیجه، کاهش کارایی تخمین‌زن‌های مدل می‌شوند. لذا این موضوع می‌بایست در واردکردن متغیرهای مستقل در مدل رگرسیون و بررسی معنی‌داری کلی، در مدل انتخابی در نظر گرفته شود. بدین‌منظور ماتریس همبستگی بین متغیّرهای کاندید تشکیل گردید. با توجه به این ماتریس ملاحظه شد که بسیاری از

متغیرهای توضیحی با یک‌دیگر همبستگی شدید مثبت و معنی‌دار، دارند. با مدنظر داشتن این نکته مدل‌هایی برآورد می‌گردند که متغیرهای توضیحی آن، از حداقل همبستگی
دو به دو برخوردار باشند. بدین ترتیب روشی که برای ورود متغیرها در مدل رگرسیون انتخاب شد انتخاب رو به جلو[1] متغیرها بود.

توجه داریم که ملاک ما برای گزینش بهترین مدل، علاوه بر معنی‌داریِ ضرایب آن بررسی میزان لگاریتم احتمال، معیار آکائیک یا بیزین شوارتز، شبه R^2 رگرسیون و نیز نیکویی برازش آن است. پس از آنکه بهترین مدل رگرسیون تک‌متغیره گزینش شد، سایر متغیرها در این مدل وارد گردیده و آن‌گاه بهترین مدل لاجیت دومتغیره برگزیده می‌شود. این‌کار را همچنان برای مدل‌های سه متغیره، چهار متغیره، و غیره ادامه می‌دهیم تا جایی که معیارهای خوب بودن مدل رگرسیونی جدید از آخرین مدل گزینش‌شده، بهتر نباشد. این روش انتخاب متغیرهای توضیحی، روش انتخاب رو به جلو تحت رهیافت گام به گام[2] نامیده می‌شود[3].

۲-۱-۳- معرفی بهترین مدل ارزیابی ریسک اعتباری

بر اساس روش گام به گام که در فوق بدان اشاره شد، متغیرهای مستقل وارد رگرسیون لاجیت شدند. با تحلیل آماره‌هایی که در مدل رگرسیونی محاسبه شده‌اند یکی از بهترین مدل‌های تخمین‌زده شده، مدلی است که در جدول مشاهده می‌شود. همان‌طورکه ملاحظه می‌گردد مدلی که برآورد شده است، شامل متغیر وابسته prform1 به‌عنوان عمل‌کرد مشتری و یازده متغیر مستقل histcod، profcod، astgdp، salast، racccdet، quastsal، cshsal، wcapsal، lperd، ryer و bdetdet است. این متغیرها به‌ترتیب بیان‌گر سابقه عمل‌کرد مشتری در سیستم بانکی، وضعیت سوددهی آن، لگاریتم دارایی به شاخص ضمنی، فروش به دارایی، حساب‌های دریافتنی به بدهی‌ها،

[1] forward looking

[2] البته ما در این تحقیق از هر دو روش انتخاب روبه جلو و حذف روبه عقب استفاده نمودیم. نکته جالب توجه آن‌است که هر دو روش به یک مدل نهایی منتهی شد که در جدول ۳ آمده‌است.

[3] سبزواری، ۱۳۸۴

دارایی‌های سریع به فروش، موجودی نقد به فروش، سرمایه در گردش به فروش، دوره‌ی وام، سابقه‌ی ارتباط با بانک کارآفرین و بدهی بانکی به کل بدهی است.

جدول ۳: خروجی‌های مدل لاجیت برآوردشده نهایی

متغیر وابسته	ضرایب	انحراف استاندارد خطا	مقدار z استاندارد	مقدار p-value
سابقه عملکرد مشتری در سیستم بانکی	./۶۷	./۱۷	۳/۸	۰
وضعیت سوددهی	./۷۰	./۴۰	۱/۷۵	../۰۸
دارایی به شاخص ضمنی	-۰./۱۷۷	./۰۳	-۵	۰
فروش به دارایی	-./۴	./۱۷	-۲/۳۵	./۰۱
حساب‌های دریافتنی به بدهی‌ها	-۱/۵	./۵۱	-۲/۹۷	۰
دارایی سریع به فروش	-./۰۹	./۰۳	-۲/۵۸	./۰۱
موجودی نقد به فروش	./۳۲	./۱۲	۲/۶۷	۰
سرمایه در گردش به فروش	./۱۴	./۰۵	۲/۸۷	۰
دوره وام	./۰۴	۰۰	۴/۷۳	۰
سابقه‌ی ارتباط با بانک کارآفرین	-۰./۶۴	./۱۴	-۴/۵۸	۰
بدهی بانکی به کل بدهی	۱/۲	./۵۶	۲/۱۷	./۰۳

نسبت‌های فروش به دارایی، دارایی‌های سریع به فروش، موجودی نقد به فروش و سرمایه درگردش به فروش از گروه نسبت‌های نقدینگی است. نسبت بدهی بانکی به بدهی از گروه نسبت‌های بدهی است. نسبت حساب‌های دریافتنی به بدهی‌ها از گروه نسبت‌های فعالیت است. ضمن این‌که دارایی‌ها به شاخص ضمنی[۱] بیان‌گر قیمت واقعی دارایی‌ها بوده و لذا از ارقام اصلی ترازنامه خواهد بود. متغیرهای «سابقه عمل‌کرد در

[۱] این شاخص (GDP deflator) از تقسیم تولید ناخالص داخلی اسمی به تولید ناخالص داخلی به قیمت ثابت محاسبه می‌شود.

سیستم بانکی»، «وضعیت سوددهی»، «طول دوره‌ی وام و سابقه‌ی ارتباط با بانک کارآفرین» از متغیرهای اساسی غیر ترازنامه‌ای هستند که به صورت متغیر مجازی یا خود متغیر در نظر گرفته‌شده‌اند، به این صورت که در متغیر histcod به شرکت‌هایی که در سیستم بانکی کشور خوش‌حساب بوده‌اند و هیچ‌گونه چک برگشتی یا اینکه اصلاً سابقه‌ای نداشته‌اند، عدد صفر و آنهایی که در سیستم بانکی بدحساب بوده‌اند عدد یک اختصاص یافته است. در قالب متغیر مجازی profcod به شرکت‌هایی که سود خالص مثبت دارند عدد صفر و به آنهایی‌که سود خالص صفر یا منفی دارند عدد یک اختصاص داده شد. طول دوره‌ی وام نیز برابر زمان سررسید آخرین قسط وام به‌صورت ماه است. سابقه ارتباط مشتری با بانک کارآفرین نیز به‌صورت سال و از تاریخ افتتاح حساب تا تاریخ دریافت محاسبه شده است. از گروه نسبت‌های بازار (مربوط به قیمت سهام در بورس) متغیری در رگرسیون قرار نگرفته است چرا که تعداد بسیار اندکی از مشتریان با شکل سهامی عام از بانک کارآفرین وام گرفته‌اند.

آماره‌های معمول رگرسیون را در جدول(۳) مشاهده می‌کنیم. در ستون اول متغیر وابسته و متغیرهای مستقل قرار گرفته است. ستون دوم شامل ضرایب متغیرهای مستقل و ستون سوم انحراف استاندارد این ضرایب، ستون چهارم آماره نرمال استاندارد هر ضریب و ستون پنجم مقدار p-value معنی‌دار بودن یا نبودن متغیرهای رگرسیون را مشخص می‌نماید. با توجه به این‌که فاصله اطمینان ۹۵ درصد در این مطالعه، معیار معنی‌داری قرارگرفته‌است و از آنجایی‌که مقدار p-value برای همه متغیرهای مستقل کمتر از ۵ درصد است و تنها برای متغیرprofcod، ۸ درصد است، لذا ۱۰ متغیر مدل فوق در سطح ۹۵ درصد اطمینان و profcod در سطح ۹۰ درصد اطمینان اختلاف معنی‌دار از صفر دارند. لگاریتم احتمال رگرسیون ۱۶۱/۷۸- و معیار آکائیک ۰/۷۸ است.

۲-۱-۴- آزمون نیکویی برازش

نرم‌افزارهای آماری و اقتصادسنجی دو آزمون نیکویی برازش هاسمر- لمشوف[1] و اندروز[2] را انجام می‌دهد. ایده‌ی اصلی این دو آزمون آن است که مقادیر برازش‌شده مورد انتظار را با مقادیر واقعی هر گروه مقایسه می‌کند. اگر اختلافات بزرگ باشد، مدل را رد می‌کنیم چراکه برازش نامناسبی برای داده‌ها فراهم می‌کند. به‌طورخلاصه 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=1 باشد، گروه‌بندی می‌کند. آماره χ^2 در پایین جدول (۴) گزارش شده است. مقدار آماره هاسمر- لمشوف (۳/۹۱) از ۱۶ کمتر است لذا این آزمون نیز نیکویی مدل برازش شده را تأیید می‌نماید.

جدول ۴: خروجی آزمون نیکویی برازش

آزمون نیکویی برازش	
تعداد مشاهدات	۴۴۸
تعداد گروه‌ها	۱۰
مقدار آماره هاسمر- لمشوف	۳/۹۱
Prob>chi2	./۸۶۵

۲-۱-۵- تحلیل اثر نهایی و کشش

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

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

[1] Hosmer-Lemeshow (1989)
[2] Anrdews (1988a,1988b)

در این رابطه $f(x)=\frac{dF(x)}{dx}$ تابع چگالی $F(x)$ است. توجه داشته باشیم که β_j به‌وسیله‌ی عامل f که خود بستگی به مقادیر همه‌ی 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.

یکی از تحلیل‌های دیگری که از مدل لاجیت می‌توان استخراج کرد و تقریباً شبیه اثر نهایی است، حساسیت متغیر وابسته به تغییر در هر یک از متغیرهای مستقل مدل است. در واقع تحلیل حساسیت محاسبه کشش متغیر وابسته نسبت به متغیرهای مستقل خواهد بود که معیار بهتری برای تعیین اهداف ما خواهد بود و درست به‌همین دلیل نیز در این قسمت کشش محاسبه شده است[۱].

جدول ۵: خروجی محاسبه کشش‌ها بعد از تخمین مدل لاجیت

متغیر	میزان کشش	مقدار prob
سابقه‌ی عمل‌کرد مشتری در سیستم بانکی	./۶۶	۰
وضعیت سوددهی	./۰۸	./۱۵۳
دارایی به شاخص ضمنی	-./۰۱۷	۰
فروش به دارایی	-./۰۳۹	./۰۱۵
حساب‌های دریافتنی به بدهی‌ها	-./۱۵	./۰۰۲
دارایی سریع به فروش	-./۰۰۹	./۰۰۵
موجودی نقد به فروش	./۰۳۱	./۰۰۴
سرمایه در گردش به فروش	./۰۱۴	./۰۰۲
دوره وام	./۰۰۴	۰
سابقه‌ی ارتباط با بانک کارآفرین	-./۰۶۲	۰
بدهی بانکی به کل بدهی	./۱۱	./۰۳۳

[۱] سبزواری (۱۳۸۴)

برطبق جدول (۵)، متغیر وابسته (خوش حسابی یا بدحسابی مشتری) به‌ترتیب به نسبت حساب‌های دریافتنی به بدهی‌ها، بدهی بانکی به کل بدهی‌ها و وضعیت سوددهی حساسیت بیشتر و به ترتیب به طول دوره‌ی وام و دارایی‌های سریع به فروش کشش و حساسیت کمتری دارد. برای مثال یک درصد تغییر در نسبت حساب-های دریافتنی به بدهی‌ها، در نقطه‌ی میانگین (x=./۳۴) باعث کاهش احتمال قصور به اندازه ۱۵/. درصد می‌گردد. این درحالی است که یک درصد تغییر در طول دوره وام در نقطه‌ی میانگین (x=۱۴/۰۹) تنها باعث افزایش میزان قصور به اندازه ۰/۰۰۴ درصد می‌گردد. ناگفته پیداست اگر، بانک بخواهد بر اساس این مدل و میزان کشش‌های فوق به مشتری خود وام دهد، ابتدا می‌بایست به متغیرهای با کشش بالا توجه بیشتری داشته باشد، چراکه تأثیر بیشتری در قصور یا عدم قصور آن دارد.

۲-۱-۶- تفسیر متغیرها و ضرایب آنها

در بخش‌های قبل نتیجه آزمون‌های اصلی مربوط به مدل لاجیت به اختصار مورد تحلیل قرارگرفت. همان‌طورکه دیدیم همه آزمون‌ها صحت برآورد را تأئید کردند. در این قسمت به ارزیابی و بررسی نقش هر کدام از متغیرهای توضیحی در پیش‌بینی قصور می-پردازیم.

- سابقه‌ی عمل‌کرد مشتری در سیستم بانکی: منظور وضعیت مشتری در سیستم بانکی به‌صورت تعیین خوش‌حساب یا بدون سابقه و یا بد حساب بودن آن است. این عامل به-صورت متغیر مجازی وارد رگرسیون شد. بدین طریق که به مشتریان خوش‌حساب و مشتریان بدون سابقه عدد ۰ و به مشتریان بدحساب عدد ۱ اختصاص یافت. انتظار بر آن است که اگر مشتری جدید بانک کارآفرین در سیستم بانکی کشور از عمل‌کرد نامناسبی برخوردار باشد، احتمال بدحساب بودن وی، در بانک کارآفرین نیز افزایش یابد. مثبت و معنی‌دار بودن ضریب این متغیر، این موضوع را تصدیق می‌نماید.

- وضعیت سوددهی: با توجه به تعریفی که از این متغیر مجازی ارایه کردیم و همچنین ضریب مثبت و معنی‌دار آن این استنباط صورت می‌گیرد که شرکت‌هایی که سود خالص صفر یا منفی دارند احتمال قصور بیشتری خواهند داشت. بدیهی است که ادبیات ریسک اعتباری نیز این موضوع را تأئید می‌کند.

- دارایی‌ها به شاخص ضمنی: این متغیر همان‌طورکه گفته شد، بیان‌گر ارزش واقعی دارایی‌ها است و انتظار داریم با احتمال قصور رابطه‌ی منفی داشته باشد. با توجه به ضریب منفی آن در جدول (۳)، انتظار ما برآورده می‌شود.

- نسبت فروش به دارایی: این نسبت بیان‌گر میزان فعالیت و کارایی شرکت است. بنابراین انتظار می‌رود با افزایش میزان گردش دارایی‌های شرکت احتمال قصور آن نیز کاهش یابد. از جدول (۳) نیز مشخص است که این نسبت با قصور رابطه‌ی معکوس دارد.

- نسبت حساب‌های دریافتنی به بدهی‌ها: با توجه به تئوری مالی و ریسک اعتباری، رابطه‌ی این نسبت با بدحساب بودن مشتری معکوس است. از آنجا که ضریب این متغیر منفی و با معنی است لذا مدل تجربی با تئوری در تطابق خواهد بود.

- نسبت دارایی‌های سریع به فروش: از نظر تئوریک این نسبت می‌تواند با احتمال قصور رابطه‌ی مثبت یا منفی داشته باشد. در مطالعه‌ی تجربی ما این رابطه منفی است.

- نسبت موجودی نقد به فروش: از نظر تئوریک این نسبت می‌تواند با احتمال قصور، رابطه‌ی مثبت یا منفی داشته باشد. در مطالعه‌ی تجربی ما این رابطه، مثبت است.

- نسبت سرمایه در گردش به دارایی‌ها: این نسبت از گروه نسبت‌های نقدینگی است و با توجه به ادبیات موضوع ریسک اعتباری می‌تواند با احتمال قصور رابطه‌ی منفی یا مثبت داشته باشد. این امر با عنایت به ضریب مثبت و معنی‌دار این متغیر در مدل رگرسیون تجربی، قابل توجیه است.

- طول دوره‌ی وام: طول دوره‌ی وام به‌صورت ماه وارد مجموعه‌ی متغیرها گردیده است. از حیث نظری هر چه قدر طول دوره وام بیشتر شود، ریسک آن نیز بیشتر می‌شود[1]. بنابراین مدل تجربی باتوجه به داشتن ضریب مثبت با تئوری امور مالی و ریسک سازگار است.

- تعداد سال‌های ارتباط با بانک کارآفرین: علامت منفی متغیر بیان‌گر ارتباط منفی افزایش تعداد سال‌های ارتباط شرکت با بانک و احتمال قصور آن است.

[1] این موضوع بیان‌گر نمودار ساختار زمانی نرخ بهره با شیب مثبت است.

- نسبت بدهی‌های بانکی به کل بدهی: انتظار بر آن است که هرچه قدر نسبت بدهی‌های بانکی (یا وام‌های) شرکت به کل بدهی‌های آن بیشتر باشد، ریسک اعتباری آن نیز به نسبت بیشتر شود. این موضع با تأئید معنی‌داری ضریب و نیز علامت مثبت آن قابل استناد خواهد بود.

۲-۱-۷- جدول پیش‌بینی مورد انتظار

این جدول طبقه‌بندی درست و نادرست را بر اساس یک قاعده پیش‌بینی که کاربر آن‌را تعریف کرده است نشان می‌دهد. برای دیدن نتایج این جدول ابتدا باید مقدار برش (دراینجا ۰/۳) را مشخص کرد. جدول (۶) نتیجه طبقه‌بندی را بر مبنای مقادیر برش مفروض نشان می‌دهد.

جدول ۶: خروجی پیش‌بینی مورد انتظار

	مشتریان خوش‌حساب	مشتریان بدحساب	کل
P(DEP=1)<=./3	۳۲۰	۴۰	۳۶۰
P(DEP=1)>./3	۴۶	۴۲	۸۸
کل	۳۶۶	۸۲	۴۴۸
صحیح	۳۲۰	۴۲	۳۶۲
درصد صحیح	۸۷/۴۳	۵۱/۲۲	۸۰/۸۰
ناصحیح	۱۲/۵۷	۴۸/۷۸	۱۹/۲

طبقه‌بندی درست موقعی حاصل می‌شود که احتمال پیش‌بینی‌شده کمتر یا برابر مقدار برش (دراینجا ۰/۳) بوده وDEP یا متغیر وابسته برابر صفر باشد یا موقعی‌که احتمال پیش‌بینی‌شده بزرگ‌تر از مقدار برش بوده و DEP برابر یک است. در مدل ما تعداد ۳۲۰ مشاهده دارای DEP=0 (قصور نکرده) و تعداد ۴۶ مشاهده با DEP=1 (قصورکرده) به درستی به‌وسیله‌ی مدل پیش‌بینی‌شده است. نکته‌ی قابل توجه در این ادبیات آنست که درصدی از مشاهدات که در آنها DEP=1 است و به‌درستی پیش‌بینی شده‌اند

حساسیت[1] نامیده می‌شوند، درحالی‌که درصدی از مشاهدات که در آنها DEP=0 است و به‌درستی پیش‌بینی شده‌اند مشخصه[2] نامیده می‌شود.

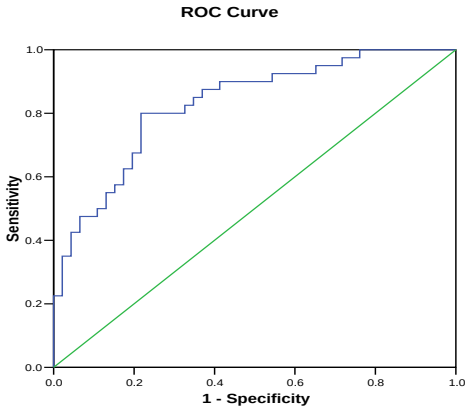
۲-۱-۸- منحنی ROC[3] و بررسی قدرت تمایز مدل تجربی

این منحنی معیار ارزشمندی برای ارزیابی عمل‌کرد مدل‌های طبقه‌بندی و دسته‌بندی فراهم می‌آورد. منحنی ROC شاخص مجازی برای ارزیابی دقت آزمایش است. منطقه زیر منحنی بیان‌گر این است که، احتمال قصور برای یک مشتری بدحساب از احتمال قصور برای مشتری خوش‌حساب که هر دو به‌طور تصادفی گزینش شده‌اند بالغ گردد. در نمودار ROC در محور عمودی آن sensitivity و بر روی محور افقی 1-specificity مشاهده می‌گردد، این دو محور را با توجه به موضوع امتیازدهی اعتباری توضیح می‌دهیم.

Sensitivity یا حساسیت؛ عبارت است از: درصدی از مشتریان بدحساب که نتیجه امتیازدهی آنان نمره‌ای بزرگتر از مقدار برش C است.

Specificity یا مشخصه: سهمی از مشتریان خوش‌حساب است که نتیجه‌ی امتیازدهی آنان، نمره‌ای بزرگ‌تر از مقدار برش C است و یا به عبارت دیگر در مورد آنان دچار خطای نوع اوّل شده‌ایم[4].

نمودار۱: منحنی ROC برای مدل نهایی



[1] sensitivity
[2] specificity
[3] receiver operating characteristic
[4] http://www.FirstKnow.It

جدول ۷: نتیجه محاسبات مربوط به منحنی ROC برای مدل نهایی

سطح معنی داری ۹۵ درصد		ارزش prob	انحراف خطا	محدوده
مرز بالایی	مرز پایینی			
/۹۰	/۷۵	۰	/۰۴۴	/۸۲۹

در نمودار (۱) منحنیROC را برای مدل امتیازدهی اعتباری مشتریان حقوقی بانک مشاهده می‌کنیم. محدوده‌ی[1] زیر منحنی (۸۲۹/.) بیان‌گر احتمال آن است که نمره‌ی امتیازدهی، برای یک مشتری قصورکرده از نمره یک، مشتری قصورنکرده‌ای که به‌صورت تصادفی انتخاب شده‌است، بیشتر شود. معنی‌داری مجانبی[2] برای مدل برآورد شده برابر صفر است لذا آزمون فوق معنی‌دار است. درحالتی‌که رفتار مشتریان به‌صورت تصادفی حدس زده می‌شود، احتمال درست پیش‌بینی کردن برابر ۰/۵ است و این درحالی است که در مدل فوق، محدوده‌ی زیرمنحنی برابر ۸۲۹/. است. بنابراین استفاده از نتایج مدل امتیازدهی بهتر از حدس زدن رفتار آتی مشتریان به‌صورت کاملاً تصادفی است[3].

۲-۲- درخت‌های طبقه‌بندی

درخت‌های طبقه‌بندی یک روش داده‌کاوی است که با، استفاده از ساختاری به نام درخت‌های تصمیم داده‌های جدید را طبقه‌بندی می‌کند. درخت‌های تصمیم از سؤالاتی تشکیل شده که، نمونه آموزشی را به بخش‌های کوچک و کوچکتر تقسیم می‌کند. این امر تا جایی ادامه پیدا می‌کند که مجموعه‌ای از افراد باقی بمانند که آنها را به هیچ طریقی نتوان در گروه‌های جدا قرار داد. در روش CART سؤال‌ها فقط به‌صورت بله یا خیر هستند. به‌عنوان مثال، سؤال می‌تواند به‌صورت «آیا شرکت سهامی عام است یا خیر» و یا «آیا شرکت در سیستم بانکی خوش‌حساب و یا بدحساب شناخته می‌شود»؟ مطرح شود. الگوریتم استفاده از CART برای تعیین خصوصیات و سؤال‌های تفکیک-

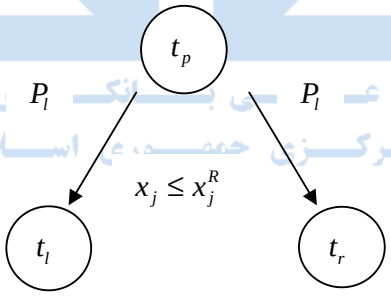
[1] area
[2] asymptotic significant
[3] Engelmann, Hayden, and T. Dirk (2003)

کننده، به‌صورت بازگشتی است، به‌طوری‌که 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 دسته‌بندی داده‌ها می‌نماید. هرسؤال داده‌های باقیمانده را به دو گروه تقسیم می‌کند که بیشترین تفاوت را با هم داشته باشند و از طرفی داده‌های قرارگرفته در هر گروه بیشترین همگنی را با سایر داده‌های هم‌گروه خواهند داشت. این فراگرد برای هر قسمت باقیمانده، ازداده‌ها تکرار خواهد شد[1].

۲-۲-۱- روش طبقه‌بندی از طریق درخت‌های طبقه‌بندی

فرض کنید t_p گره اصلی (مادر) و t_r و t_l به‌ترتیب گره‌های فرعی، سمت راست و چپ گره اصلی باشند. نمونه‌ی آموزشی ماتریسی از متغیرهاست که آن‌را با ماتریس X نشان می‌دهیم این ماتریس M متغیر دارد که هر کدام را به‌صورت x_j نمایش می‌دهیم. تعدادکل مشاهدات را هم N می‌نامیم. بردار کلاس Y در برگیرنده‌ی تعداد N مشاهده است که به‌هرکدام از آنها یکی از K کلاس مدنظر را نسبت داده‌ایم. یک درخت دسته-بندی براساس قاعده‌ی تفکیک بنا می‌شود که این قاعده تفکیک، تقسیم داده‌ها به قسمت‌های کوچک و کوچکتر را امکان‌پذیر می‌کند. می‌دانیم که هربار داده‌ها (درهرگره) به دو بخش تقسیم می‌شوند. در هر قسمت داده‌هایی قرار می‌گیرند که بیشترین همگنی را داشته باشند.

نمودار ۲: الگوریتم تفکیک در روش CART



[1] Steele(1995)

که در آن t_p ، t_r و t_l به‌ترتیب گره اصلی، گره فرعی راست و گره فرعی چپ هستند. x_j متغیر **j** ام و x_j^R بهترین مقدار x_j برای تفکیک است. بیشترین همگنی یک گره فرعی با تابعی به نام تابع ناخالصی $i(t)$ تعیین می‌شود. برای هرکدام از حالت-های تفکیک چپ و راست میزان ناخالصی گره اصلی ثابت باقی می‌ماند، یعنی برای تمام حالت‌های $x_j \le x_j^R$ و $J = 1, \cdots, M$ 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. به‌عبارت 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 آن‌را با $\Delta i(t)$ نشان می‌دهیم.

$$\Delta i(t) = i(t_p) - E[i(t_C)] \quad (۱)$$

که در آن t_c گره‌های سمت چپ و راست گره اصلی هستند. فرض کنید p_l, p_r به ترتیب احتمال‌های گره‌های سمت چپ و راست باشند. به‌این معنی که یک داده با چه احتمالی در گره سمت راست و با چه احتمالی در گره سمت چپ قرار خواهد گرفت. بنابراین می‌توان نوشت:

$$\Delta i(t) = i(t_p) - p_l i(t_l) - p_r i(t_r) \quad (۲)$$

بنابراین در هر گره در روش **CART** به حل یک مسأله ماکزیمم‌سازی به‌صورت زیر خواهیم پرداخت.

$$\underset{x_j \le x_j^R,\ j=1....M}{\arg\max} \quad [i(t_p) - p_l i(t_l) - p_r i(t_r)] \quad (۳)$$

معادله (۳) نشان می‌دهد که براساس الگوی **CART**، تمام حالتهای ممکن بین متغیرها در ماتریس **x** برای یافتن بهترین معادله تفکیک $x_j \le x_j^R$ جستجو می‌شود با این قید که مقدار تغییر در تابع ناخالصی کل $\Delta i(t)$ ماکزیمم شود. سؤال مهم دیگر این است که تابع ناخالصی چگونه تعریف می‌شود. در تئوری تعدادی تابع ناخالصی وجود دارد که هر کدام برای موارد خاصی به‌کار می‌روند و دارای مزایای متفاوتی هستند. اما

در عمل فقط از دو مورد آنها در اغلب موارد استفاده می‌شودکه عبارتند از الگوی تفکیک جینی[1] و الگوی تفکیک توئینگ[2].

۲-۲-۲- روش طبقه‌بندی از طریق درخت‌های رگرسیونی

در درخت‌های رگرسیونی دسته‌بندی صورت نمی‌گیرد، در عوض بردار پاسخی مانند **Y** تعریف می‌شود که مقدار پاسخ هر متغیر موجود در ماتریس **X** در آن نگهداری می‌شود. با توجه به‌این‌که در درخت‌های رگرسیونی از کلاس‌های پیش‌فرض استفاده نمی‌کنیم بنابراین از الگوهای جینی و توئینگ خبری نیست. در درخت‌های رگرسیونی الگوریتم تفکیک براساس کمینه‌کردن مربع خطاها صورت می‌گیرد. به‌این-معنی که مقدار انتظاری واریانس دو گره‌ای که از تفکیک به‌دست می‌آید، باید کمترین مقدار یا مینیمم باشد. رابطه آن را به‌صورت زیر نشان می‌دهیم.

$$\underset{x_j \le x_j^R,\ j=1....M}{\arg\min} \quad [P_l \mathrm{Var}(Y_l) + P_l \mathrm{Var}(Y_l)] \qquad (۴)$$

که $\mathrm{Var}(Y_r), \mathrm{Var}(Y_l)$ بردارهای پاسخ برای گره‌های فرعی چپ و راست هستند. اگر به مشاهدات مربوط به کلاس **k** عدد یک و به سایر مشاهدات عدد صفر نسبت دهیم واریانس نمونه برابر $p(k|t)[1-p(k|t)]$ خواهد بود که همان احتمال تابع توزیع دو جمله‌ای است. بنابراین با توجه به تعداد کلاس‌های موجود **k** می‌توانیم تابع ناخالصی را به‌صورت زیر اندازه‌گیری کنیم.

$$i(t) = 1 - \sum_{k=1}^{K} p^2(k \mid t) \qquad (۵)$$

همان‌طورکه ذکر شد، ساختن بزرگترین درخت ممکن به معنی اعمال قاعده تفکیک به آخرین داده‌های باقیمانده، در نمونه مورد آموزش است. در حالت درخت‌های رگرسیونی ممکن است، درخت بیشینه بسیار بزرگ باشد، در این حالت هر مقدار بردار پاسخ، ممکن است به یک گره جداگانه متعلق باشد[3].

[1] Gini
[2] Towing

[1] آیتی گازار (۱۳۸۴)

۲-۲-۳-ایجاد درخت طبقه‌بندی

در این قسمت با استفاده 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 مشاهده‌ی یک احتمال به‌دست می‌آید، در این روش مشاهدات براساس نزدیک‌ترین کلاسِ مرتبط، دسته‌بندی شده و سعی می‌شود مشاهدات موجود در هر کلاس بیشترین شباهت را به یک‌دیگر داشته باشند. در این روش از متغیرها به‌عنوان یک رابطه، استفاده نمی‌شود بلکه از متغیرها فقط به عنوان تفکیک‌کننده استفاده می‌شود. اکنون دسته‌بندی مشتریان را با استفاده از روش ناپارامتریک درخت‌های طبقه‌بندی انجام می‌دهیم. برای این منظور از نرم‌افزار CART مربوط به شرکت سالفورد سیستم[1] استفاده خواهیم کرد.

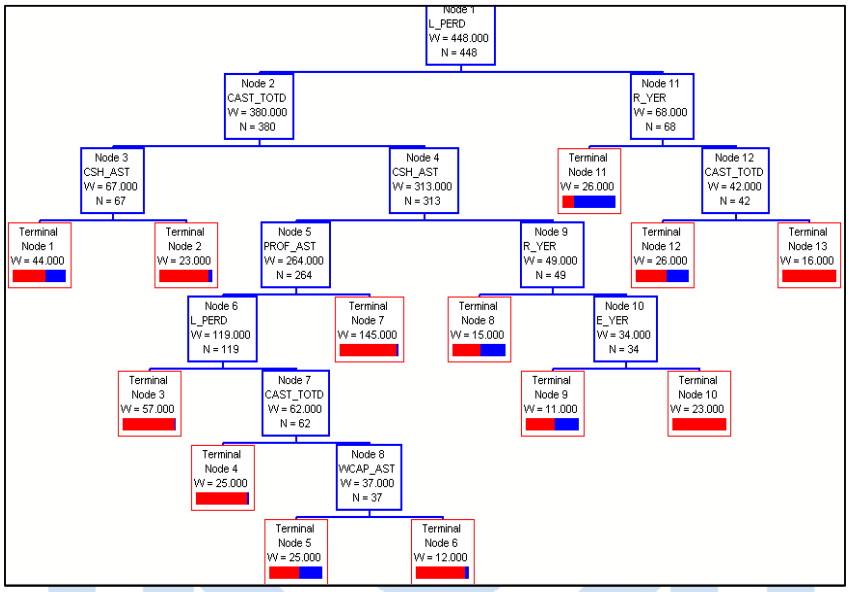
۲-۲-۳-۱- تفکیک و طبقه‌بندی داده‌ها

همان‌طورکه قبلاً گفته‌شد، در این مطالعه تجربی برای برآورد مدل اعتبارسنجی، ۴۹ متغیر شامل ۳۳ نسبت مالی از چهار گروه نسبت اصلی نقدینگی، بدهی یا اهرمی، فعالیت و سوددهی و همچنین ۱۶ متغیر اصلی و مجازی[2] دیگر به‌عنوان متغیرهای کاندید، انتخاب شد. برخلاف مدل‌های آماری و اقتصادسنجی همچون لاجیت در روش طبقه‌بندی درختی محدودیتی در انتخاب و تعیین متغیرهای تفکیکی نداریم. لیکن برای این‌که به درخت بهتری (از نظردقت تفکیک) دست یابیم، باید الگوی مشخصی در انتخاب درخت‌های گوناگون با دقت‌های متفاوت داشته باشیم. بدین منظور سعی شد درخت‌های مختلف با متغیرهای طبقه‌بندی متفاوت مورد آزمایش قرار گیرد. مدل یا درخت نهایی براساس دقت آن به‌ویژه در نمونه آزمون و همچنین رعایت میزان بهینه خطا و تعداد گره‌های نهایی انتخاب گردید.

[1] salford system
[2] Dummy variable

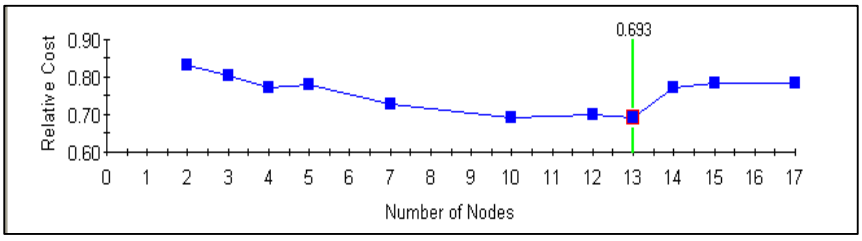
مدل‌های گوناگونی با متغیرهای مختلف مـورد ارزیـابی قـرار گرفـت. در نهایـت متغیرهایی وارد درخت طبقه‌بندی شد که هم در مدل لاجیت تک‌متغیره با وجود عـرض از مبدأ و هم بدون وجود عرض از مبدأ دارای معنی‌داری در سطح ۹۵ درصـد بـود. بـرای ساخت این درخت، ۱۱ متغیر شامل طول دوره وام (l_perd)، سـابقه‌ی ارتبـاط بـا بانـک-کارآفرین (r_yer)، سابقه‌ی تأسیس شرکت (e_yer)، نسبت دارایی جاری به کل بـدهی‌هـا (cast_totd)، سرمایه درگردش به دارایی‌ها (wcap_ast)، نسبت موجودی نقدی به دارایی‌ها (csh_ast)، نسبت حساب‌های دریافتنی به فروش (racc_sal)، نسبت فروش به دارایـی‌هـا (sal_ast)، نسبت سود خالص به دارایی‌ها (prof_ast)، نسبت سود عملیاتی به دارایـی‌هـا (opr_ast) و نسبت لگاریتم داریی‌ها به شاخص ضمنی (ast_gdp). همان‌طورکـه مـشاهده می‌شود پس از این‌که مدل را با شرایط دل‌خواه ساختیم، درختی به شکل زیر ایجاد مـی-شود که بر طبق آن مشاهدات در کلاس‌های مجزا قرار می‌گیرند. در این درخت با توجه به اهمیت متغیرها در تفکیک و کلاس‌بندی، ۷ متغیر نهایی تفکیکی قرار دارد و متغیرهـای دیگر در درخت قرار ندارد.

نمودار ۳: درخت ایجاد شده بر اساس نمونه



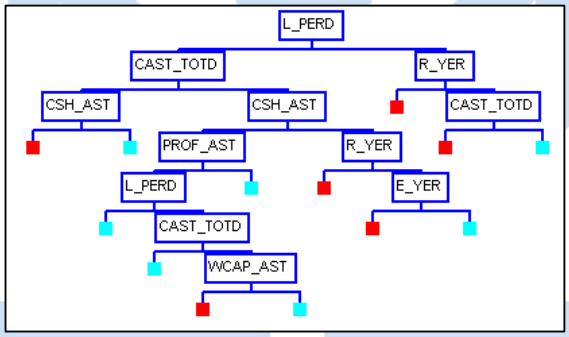
همان‌طورکه در بخش‌های قبل شرح دادیم، ابتدا بزرگ‌ترین درخت ممکن براساس نمونه، ساخته می‌شود و سپس بر اساس پارامترهای تنظیم‌شده، بهترین اندازه‌ی درخت با کمترین مقدار خطا تعیین می‌شود. همان‌طورکه می‌دانیم اندازه‌ی درخت بر اساس گره‌های نهایی تعیین می‌شود که در این مورد تعداد گره‌های نهایی یا به عبارتی طبقات، T= ۱۳است. نمودار (۴) منحنی خطا را برای این درخت نشان می‌دهد که در مقدار T =۱۳ به حداقل می‌رسد. به راحتی قابل مشاهده است که با افزایش یا کاهش تعداد گره‌های نهایی میزان خطا افزایش می‌یابد. برای درخت مورد نظر این مقدار برابر با ۰/۶۹۳ است.

نمودار ۴: منحنی خطا به‌صورت تابعی از گره‌های نهایی درخت



البته با تغییر در پارامترهای مدل نظیر تغییر در الگوی تفکیک و مقدار N_{min} درخت-هایی با ساختار متفاوت به‌وجود می‌آید که در این مورد بهترین حالت، استفاده از الگوی جینی و مقدار $N_{min} = ۱۳$ تشخیص داده شده است. درالگوریتم 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. با افزایش گره-های نهایی میزان خطا کاهش پیدا می‌کند و در حالت درخت بیشینه این مقدار، صفر خواهد بود اما در این حالت تعداد گره‌های نهایی بیان‌گر 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 ذکر شده در بالاست. در درخت تصمیم به-وجود آمده تفکیک بر اساس اهمیت متغیرها در به‌وجود آوردن گروه‌های تفکیکی انجام می‌گیرد. بر این اساس، صورت کلی الگوی تفکیکی و متغیرهای تفکیک‌کننده در نمودار (۵) آورده شده است. اهمیت نسبی این متغیرها در جدول (۸) ارایه شده است.

نمودار ۵: ساختار متغیرهای تفکیک کننده



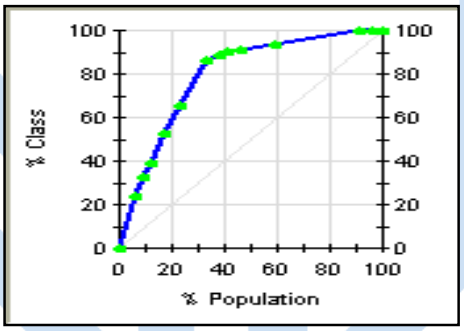
جدول ۸: اهمیت نسبی متغیرها

Variable	Score
L_PERD	100.00
CAST_TOTD	87.65
WCAP_AST	66.81
CSH_AST	57.24
PROF_AST	50.37
E_YER	50.14
R_YER	31.98
OPR_AST	27.73
SAL_AST	22.70
RACC_SAL	14.18
AST_GDP	9.75

۲-۲-۳-۲- بررسی منحنی‌های فایده

اولین ویژگی مورد بررسی منحنی‌های به‌دست‌آمده، منحنی فایده[1] است. در این منحنی میزان فایده‌ای که در صورت استفاده از مدل در مقابل روش‌شانسی و بدون استفاده از مدل (یعنی همان احتمال ۵۰ درصد) نشان داده می‌شود. این منحنی برای کلاس یک که نشان‌دهنده‌ی قصور است آمده است. نمودار (۶) سطح زیر خط قطری برابر با ۰/۵ است که در صورت نبودن هیچ مدلی و این‌که مشتریان به‌صورت تصادفی پذیرفته شوند به‌دست می‌آید و به‌عبارتی هیچ فایده‌ای وجود ندارد. اما اگر از روش درخت‌های طبقه‌بندی استفاده کنیم فایده‌ای نسبت به حالت اولیه (بدون استفاده از مدل) به‌دست می‌آوریم. برای کلاس یک این منحنی ترسیم شده‌اند.

نمودار ۶: منحنی فایده برای کلاس یک (مشتریان بد)



جدول ۹: جزئیات مربوط به منحنی فایده مربوط به کلاس یک (مشتریان بدحساب)

Node	Cases Tgt. Class	% of Node Tgt. Class	% Tgt. Class	Cum % Tgt. Class	Cum % Pop	% Pop
11	20	76.923	24.390	**24.390**	**5.804**	5.80
8	7	46.667	8.537	**32.927**	**9.152**	3.34
9	5	45.455	6.098	**39.024**	**11.607**	2.45
5	11	44.000	13.415	**52.439**	**17.188**	5.58
12	11	42.308	13.415	**65.854**	**22.991**	5.80
1	17	38.636	20.732	**86.585**	**32.813**	9.82
2	2	8.696	2.439	**89.024**	**37.946**	5.13

[1] gain curve

در این منحنی برای هر گروه (گره انتهایی) یک نقطه روی منحنی در نظر گرفته می‌شود به‌این ترتیب که در هر گره درصد کلاس‌بندی صحیح در مقابل درصد فراوانی این کلاس در گره رسم می‌شود. درگوشه‌ی سمت چپ پایین، بدترین گره از لحاظ کلاس-بندی و در گوشه‌ی سمت راست بالا بهترین گره آمده است. در بدترین گره تعدادی از مشاهدات وجود دارند که درصد فراوانی آنها، در گره کم بوده و درصد کلاس‌بندی آنها هم ناچیز است و در حقیقت به کلاسی تعلق ندارند. این‌ها همان داده‌های پرت[1] (یا استثناء) هستند. بهترین گره حالتی است که تمام مشاهدات موجود در آن به یک کلاس خاص اختصاص دارند و به‌صورت کامل کلاس‌بندی شده‌اند و به‌عبارتی صددرصد فراوانی موجود در گره به‌طور کامل دارای یک کلاس مشخص است. جدول (۹) جزئیات مربوط به هر نقطه روی این نمودار را تشریح کرده‌است.

درجدول (۹) به‌ترتیب از چپ، ستون اوّل شماره‌ی گره انتهایی، ستون دوم تعداد مشاهداتی که در این گره، درست کلاس‌بندی شده‌اند، ستون سوم درصد این مشاهدات در گره، ستون چهارم درصد مشاهدات کلاس‌بندی‌شده در مقایسه با کل مشاهدات، ستون پنجم درصد تجمعی مشاهدات کلاس‌بندی‌شده نسبت به کل مشاهدات، ستون ششم درصد فراوانی تجمعی در هر گره نسبت به کل مشاهدات، ستون هفتم درصد فراوانی مشاهدات موجود در گره نسبت به کل مشاهدات و ستون هشتم فراوانی مشاهدات موجود در هر گره را نشان می‌دهد. به‌آسانی قابل مشاهده است که اطلاعاتی که از این روش در اختیار ما قرار داده می‌شود به‌مراتب بیشتر از مدل لاجیت است. در حقیقت می‌توانیم در این روش هر مشاهده را ردیابی نماییم و در مورد آن اطلاعات ارزشمندی داشته باشیم.

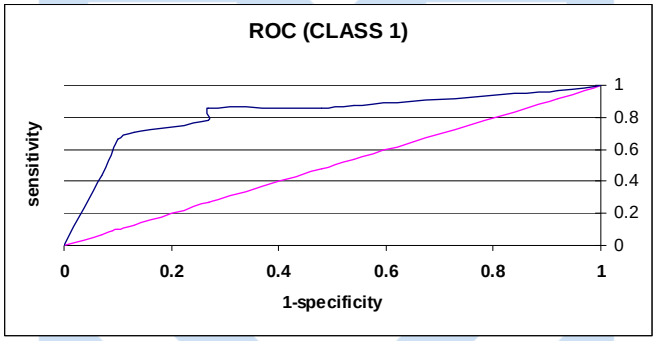
۲-۲-۳-۳- بررسی قدرت تمایز مدل

منحنی بعدی که به آن خواهیم پرداخت و با توجه به آن می‌توانیم صحت کلاس‌بندی مشاهدات را بررسی نماییم منحنی **ROC** است که قبلاً در مورد آن توضیح دادیم. همان-طورکه گفته شد سطح زیر این منحنی نشان‌دهنده‌ی صحت کلاس‌بندی است. هرچه این مقدار به یک نزدیکتر باشد مشاهدات بهتر جداسازی شده‌اند. مقدار ۰/۵ برای سطح

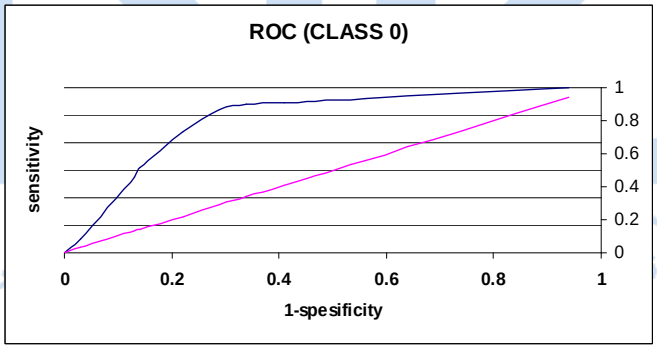
[1] outlier

زیر منحنی به معنی احتمال پنجاه درصد در کلاس‌بندی است که در این صورت استفاده از مدل، بی‌مورد خواهد بود. در مورد این روش برای هر دو کلاس یک و صفر به ازای مقادیر آستانه Sensitivity وSpecificity در جدولی محاسبه و نمودارهای مربوط به آن رسم شده‌است. آن‌چه مشخص است قابل قبول بودن تفکیک مشاهدات با این روش است، چرا که سطح زیرمنحنی برای کلاس یک و صفر تقریباً بیش از ۷۵ درصد است.

نمودار۷: منحنی ROC برای کلاس یک



نمودار ۸: منحنی ROC برای کلاس صفر



در قسمت آخر نتایج به‌دست‌آمده به‌عنوان دقت کلاس‌بندی مشاهدات که توسط نرم-افزار محاسبه شده است را ارایه می‌دهیم. همان‌طورکه در گذشته توضیح دادیم، برای ساختن یک درخت از دو نمونه استفاده می‌شود که یکی به‌عنوان سازنده‌ی مدل و دیگری به‌عنوان آزمایش‌کننده، استفاده می‌شوند. برای این دو نمونه جداول خطای کلاس‌بندی آورده شده است (جداول۱۰و۱۱). طبیعتاً خطای نمونه‌ی آزمایشی از خطای نمونه‌ی آموزشی بیشتراست. به این نوع خطا، خطای درون نمونه‌ای گفته می‌شود. در این جداول ستون اوّل نوع کلاس، ستون دوم فراوانی کلاس، ستون سوم خطای پیش‌بینی و ستون چهارم هزینه‌ی ناشی از عدم کلاس‌بندی صحیح را نشان می‌دهد.[۱]

جدول ۱۰: خطای پیش‌بینی برای نمونه‌ی آموزشی

هزینه	درصد خطا	تعداد مشاهدات اشتباه طبقه‌بندی شده	تعداد مشاهدات	کلاس
./۲۱	۲۰/۷۷	۷۶	۳۶۶	۰
./۱۳	۱۳/۴۱	۱۱	۸۲	۱

جدول۱۱: خطای پیش‌بینی برای نمونه‌ی آزمون

هزینه	درصدخطا	تعداد مشاهدات اشتباه طبقه‌بندی شده	تعدادمشاهدات	کلاس
./۲۵	۲۵/۴۱	۹۳	۳۶۶	۰
./۴۴	۴۳/۹۰	۳۶	۸۲	۱

در جدول (۱۲) دقت نهایی مدل (یا درصد طبقه‌بندی صحیح) نشان داده شده است. این مقادیر برای مشتریان خوش‌حساب تقریباً۷۹ درصد و برای مشتریان بد حساب ۸۶ درصد است.

[۱] آیتی گازار (۱۳۸۴)

جدول ۱۲: دقت نهایی مدل CART

کلاس واقعی	تعداد مشاهدات	پیش‌بینی صحیح	دقت (درصد)	دقت کل مدل
صفر	۳۶۶	۲۹۰	۷۹	۸۰
یک	۸۲	۷۱	۸۶	

۲-۳- نتایج تجربی

برای برآورد مدل‌های امتیازدهی پارامتریک کارا و مناسب نیاز به تعداد مشاهدات قابل ملاحظه و نسبتاً زیاد است. به‌طوری‌که بررسی‌ها نشان داده، در مطالعات تجربی بانک‌های موفق خارجی، بیش از ده‌ها هزار مشاهده مورد استفاده قرارگرفته است. حجم اطلاعاتی که ما در اختیار داشتیم بسیار اندک بود و به‌نظر می‌رسید که مدل‌های غیرپارامتریک و از جمله‌ی آنها روش CART می‌توانست مشکل کم‌بودن اطلاعات را مرتفع سازد. چراکه اصولاً برخی از محدودیت‌های آماری مربوط به مدل لاجیت را در روش CART که صرفاً یک روش داده‌کاوی است، ندارد.

جدول ۱۳: مقایسه دقت پیش‌بینی مدل لاجیت با روش CART

مدل	وضعیت پرداختی	تعداد مشاهدات واقعی	پیش‌بینی	دقت پیش‌بینی	دقت کل
لاجیت	۰	۳۶۶	۳۲۰	۸۷	۸۰
	۱	۸۲	۴۲	۵۱	
درخت تصمیم	۰	۳۶۶	۲۶۳	۷۱	۸۰
	۱	۸۲	۷۲	۸۷	

همان‌طورکه قبلاً نیز گفته شد و از جدول (۱۳) نیز مشخص است، دقّت روش درخت تصمیم به‌طور متوسط ۸۰ درصد است. بنابراین با مقایسه دقت این روش و دقت مدل لاجیت، می‌توان عنوان داشت که در این مطالعه دقت روش CART در پیش‌بینی صحیح مشتریان بدحساب بیش از۳۰ درصد بهتر از مدل لاجیت عمل می‌کند، لیکن مدل لاجیت در پیش‌بینی صحیح مشتریان خوش‌حساب حدود ۹ درصد

بهتر از روش CART عمل می‌کند. با نگاهی به ستون آخر (یا دقّت کل) مشخص است که دقّت کلی این دو مدل یکسان است.

برای بررسی دقت مدل لاجیت و روش CART در نمونه‌های کوچک‌تر، با استفاده از نرم‌افزار Matlab نمونه‌های تصادفی در اندازه‌های مختلف انتخاب نموده و دقّت پیش-بینی آنها را مورد ارزیابی قرار دادیم. نتایج بررسی‌ها در جدول (۱۴) آمده است. همان-طورکه از جدول مشخص است، روش CART اندکی بهتر از مدل لاجیت عمل می‌کند.

روش‌های آماری (همچون لاجیت) در برابر روش‌های غیرپارامتریک دارای برخی از ایرادات است. یکی از مهمترین ایرادات روش‌های آماری و از جمله روش رگرسیون لاجیت آن است که این‌گونه روش‌ها برپایه اطلاعات تاریخی مشتریان استوار است. در این روش همان‌طورکه دیدیم اطلاعات ترازنامه و سود و زیان سال‌های مالی قبل به عنوان ملاک ارزیابی عمل‌کرد آتی آن درنظر گرفته می‌شود. برای رفع این ایراد، می‌بایست از متغیرهای بازار همچون اطلاعاتی که در بورس اعلام می‌شود بهره‌گرفت. این امر به دلیل آن‌که مشتریان اندکی از بانک کارآفرین جزو شرکت‌های بورسی هستند امکان‌پذیر نیست. همچنین روش‌های آماری و رگرسیونی دارای فروض قوی و محدودکننده هستند. برای نمونه در روش لاجیت رابطه بین ترکیب خطی متغیرهای مستقّل و متغیر وابسته از یک تابع سیگموئید پیروی می‌کند. روش‌های غیرپارامتریک مثل CART از این ایراد مستثنی هستند.

جدول ۱۴: مقایسه دقت پیش‌بینی مدل لاجیت با روش CART در نمونه‌های کوچکتر

حجم نمونه	پیش‌بینی صحیح		دقت پیش‌بینی (درصد)	
	مدل لاجیت	روش CART	مدل لاجیت	روش CART
۱۰۰	۶۸	۷۰	۶۸	۷۰
۲۰۰	۱۴۴	۱۵۰	۷۲	۷۵
۳۰۰	۲۱۹	۲۳۵	۷۳	۷۸
۳۵۰	۲۶۵	۲۷۸	۷۵	۷۹

در این تحقیق ۴۴۸ مشاهده از مشتریان حقوقی بانک کارآفرین، به‌طور تصادفی به-منظور مدل‌سازی مورد استفاده قرارگرفت. تعداد ۳۶۲ مشاهده در دسته‌ی مشتریان

خوش‌حساب و ۸۶ مشاهده، در گروه مشتریان بدحساب قرارگرفت. برای ساختن مدل لاجیت از روش انتخاب روبه جلو و همچنین حذف روبه عقب و برای ایجاد درخت تصمیم از روش جینی و نرم‌افزار CART استفاده نمودیم. مدل لاجیت نهایی شامل ۱۱ متغیر توضیحی است. با توجه به ویژگی منحنی ROC برای مدل استخراج شده، می-توان گفت که، قدرت تمایز این مدل حدود ۳۲ درصد از حالتی که هیچ مدلی وجود ندارد بهتر است. کشش متغیرهای توضیحی نشان می‌دهد که اگر بانک بخواهد براساس این مدل به مشتریان خود وام دهد، ابتدئاً می‌بایست به متغیرهای با کشش بالا، شامل نسبت حساب‌های دریافتنی به بدهی‌ها، بدهی بانکی به کل بدهی‌ها و وضعیت سوددهی توجه بیشتری داشته باشد، چراکه تأثیر بیشتری در قصور یا عدم قصور آنان دارد.

روش دومی که در این تحقیق استفاده گردید روش CART است. در این روش درخت نهایی بر اساس دقّت آن به‌ویژه در نمونه‌ی آزمون و همچنین رعایت میزان بهینه خطا و تعداد گره‌های نهایی انتخاب گردید. در این درخت با توجه به اهمیت متغّیرها، در تفکیک و کلاس‌بندی، ۷ متغّیر نهایی تفکیکی قرار دارد و متغّیرهای دیگر در درخت قرار ندارد. ضمن این‌که اهمّیت عواملی چون دوره وام، نسبت دارایی جاری به کل بدهی و همچنین نسبت سرمایه در گردش به دارایی‌ها از سایر عوامل بیشتر است. منحنی‌های فایده و ROC درخت نشان‌دهنده‌ی منافعی است که در استفاده از این درخت حاصل می‌گردد.

مقایسه مدل لاجیت با روشCART نشان می‌دهد که برای همه مشاهدات دقّت پیش‌بینی دو مدل، تقریباً برابر است، لیکن در نمونه‌های کوچک‌تر دقّت روش CART بیشتر است. که این امر بیانگر مزیّت مدل‌های غیرپارامتریک (همچون CART) نسبت به مدل‌های آماری و پارامتریک (همچون لاجیت) خواهد بود.

منابع و مآخذ

آیتی گازار، حسین. ، ۱۳۸۴، *"مقایسه مدل پارامتریک لاجیت و مدل ناپارامتریک درختهای طبقه‌بندی در فراگرد امتیازدهی اعتباری"*، پایان نامه کارشناسی ارشد، دانشگاه صنعتی شریف.

سبزواری، حسن. ، ۱۳۸۴، *"برآورد و مقایسه مدل امتیازدهی اعتباری پارامتریک لاجیت با روش غیر پارامتریک AHP"*، پایان نامه کارشناسی ارشد، دانشگاه صنعتی شریف.

Liu and Yang (2002) "*A framework of data mining application process for credit scoring*". Arbeitsbericht Nr. 01.

Krahnen, J. P. and M. Weber, 2000, "*Generally accepted rating principles: A primer*", *Journal of Banking and Finance*.

Burak Emela, A. , M. Oralb, A. Reismanb and R. Yolalana, 2003, "*A credit approach for the commercial banking sector*", Socio-Economic Planning Sciences 37, 103–123.

Akhavein, J. , W. S. Frame and J. W. Lawrence, 2001, "*The diffusion of financial innovations*".

Liu, Y. , 2001, "*New issue in credit scoring application*", Nr Arbeitsbericht.

Liu, Y. , 2002, "*The evaluation of classification models for credit scoring*", Arbeitsbericht Nr. 02.

Kiss, F. , 2003, "*Credit scoring process from a knowledge management perspective*", Periodica polytechnica ser.soc.man.scl., VOL. 11, NO. 1, PP. 95–110.

Hayden, E. , 2003, "*Are credit scoring models sensitive with respect to default definition? Evidence from the Austrian market*".

Back, B. , T. Laitinen, K. Sere and M. van Wezel , 1996, "*Choosing bankrauptcy prediction using discriminant analysis, logit and genetic algoritm*", Turku Centre for Computer Science, Technical Report, No 40.

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.

Sinkey, Jr.Joseph F. , 1992, *"Commercial bank financial management"*, 4th edition, Macmillan.

http://www.cs.uk.n1/docs/vakken/ida/idahc8.pdf, "*Logistic regression* ".

Engelmann, B. , E. Hayden and T. Dirk, 2003, "*Measuring the Discriminative Power of Rating Systems*", Discussion paper, Series 2: Banking and Financial Supervision, No 01.

http://www.FirstKnow.It, "*Cumulative Accuracy Profile (CAP) – Gini Coefficient*".

Steele, A. ,1995, "*CREDIT clasification: A comparison of logit model and decision trees*".

Gallati, Reto, 2003, *"Risk management and capital adequacy"*, McGrwHill.