دوفصلنامه پژوهش‌های بیمه‌ای، سال اول، شماره ۲، پاییز و زمستان ۱۳۹۵
صفحات ٦١ تا ٨٠

ارزیابی الگوریتم‌های درخت تصمیم، بیز ساده و رگرسیون لجستیک در کشف تقلبات بیمه اتومبیل

دکتر آتوسا گودرزی*
سجاد جنت بابایی**

چکیده

از منظر علوم اقتصادی و با مراجعه به آمار و اطلاعات، تخلفات مالی در صنعت بیمه به‌صورت فزاینده‌ای در حال تبدیل‌شدن به یک مسئله جدی و قابل تأمل است. یکی از روش‌های مناسب جهت ارزیابی و مدل‌سازی تخلفات و تقلبات، تکنیک‌های داده‌کاوی است که می‌تواند نقش مهمی در کشف و پیش‌بینی تقلبات مالی ایفا نماید. این شیوه برای آشکار نمودن حقایق پنهان در ورای حجم انبوهی از داده‌ها بکار گرفته می‌شود. شرکت‌های بیمه می‌توانند الگوی پنهان در داده‌ها را کشف کرده و با مدل‌سازی و طراحی الگوهای مناسب اقدامات جدی در راستای کاهش تقلبات، مدیریت ریسک‌ها و ایفای تعهدات به عمل‌آورند.

در این مقاله، تکنیک‌های رایج جهت کشف تقلب در بیمه‌های اتومبیل (شامل رگرسیون لجستیک، درخت تصمیم و دسته بندی بیز ساده) مورداستفاده قرار می‌گیرد. ابزارهای داده‌کاوی به‌طور معمول با هدف تعمیم مدل‌های کشف ادعاهای تقلبی و ارائه پیش‌بینی به کار گرفته می‌شوند. متغیرهای توضیحی که در سه مدل این مطالعه مورداستفاده قرار می‌گیرند، عبارت‌اند از: سابقه بیمه‌ای، تعداد ادعاهای خسارت، تأخیر در اعلان خسارت، سن، جنس، کروکی و مبلغ خسارت (شکوری ۱۳۹۰) داده‌های موردنیاز از

* دانشیار، موسسه آموزش عالی بیمه اکو، دانشگاه علامه طباطبائی، (نویسنده مسئول) atousagoodarzi@yahoo.com
** دانش آموخته کارشناسی ارشد علوم محاسبات و برنامه‌ریزی بیمه، موسسه آموزش عالی بیمه، s.janatbabaei@gmail.com

یکی از شرکت‌های بزرگ بیمه، پس از اخذ مجوزهای لازم، جمع‌آوری شده است. داده‌ها به دو بخش تقسیم شده است. از بخش نخست برای ساخت مدل و از بخش دوم برای دسته‌بندی استفاده شده است. یافته‌های این مطالعه مؤید آن است که مدل رگرسیون لجستیک دقت بیشتری برای پیش‌بینی کل ادعاها(تقلبی و غیر تقلبی) نسبت به دو مدل دیگر، یعنی درخت تصمیم و روش بیز داشته است.

کلمات کلیدی: داده‌کاوی، تقلب، بیمه اتومبیل، درخت تصمیم، بیز ساده، رگرسیون لجستیک.

طبقه‌بندی JEL: C11, C21, G22.

مقدمه

شرکت‌های بیمه در سراسر جهان به‌طور مداوم منابع هنگفتی را در قالب جبران خسارت ادعاهای تقلبی از دست می‌دهند که در صورت صرف این منابع در فعالیت‌های تخصصی و سرمایه‌گذاری می‌توانستند خدمات مناسب‌تری به مشتریان خود ارائه نمایند. عوامل فروش می‌توانند داده‌کاوی را با اقدامات مربوط به کشف تقلب، ترکیب و دقت عملکرد خود را افزایش دهند که به این طریق نیاز به نیروی انسانی نیز کاهش می‌یابد. نتیجه این اقدامات درنهایت می‌تواند به کمینه نمودن خسارات منجر شود.

افرادی که مرتکب تقلب می‌شوند، می‌کوشند تا منفعت بیشتری از بیمه‌نامه (فراتر از خسارت واردشده)، دریافت کنند. معمولاً در دو موقعیت ممکن است بیمه‌شده اقدام به تقلب نماید. نخست در شرایطی که یک شخص به‌صورت عمدی تلاش می‌کند تا خسارتی را ایجاد و یا در گزارش میزان وقوع آن اغراق نماید. موقعیت دوم، زمانی است که بیمه‌شده با علم به وجود پوشش بیمه، احتیاط کمتر و یا حتی بی‌احتیاطی می‌کند.

براین‌اساس، می‌توان اهم هزینه‌های وارده ناشی ازاین‌گونه تقلبات را به شرح زیر طبقه‌بندی نمود:

- به‌طور میانگین خانواده‌ها حق بیمه بالاتری برای جبران هزینه‌های ناشی از تقلب می‌پردازند.
- قیمت کالاها به دلیل پرداخت حق بیمه بیشتر(به سبب هزینه‌های مربوط به تقلب‌های بیمه‌ای)، افزایش می‌یابد.
- هزینه‌های بیمه سلامت به دلیل وقوع تقلب در ادعاهای خسارت به‌ویژه در مناطقی که پوشش نامحدود هستند، افزایش پیدا می‌کند.
- عموم بیمه‌شده‌ها با دقت بیشتری موردبررسی قرار می‌گیرند و زمان رسیدگی و تسویه مطالبات طولانی می‌شود.
- به ازای هر واحد پولی که برای تقلب بیمه‌ای هزینه می‌شود؛ سودآوری شرکت بیمه به‌طور مستقیم متأثر می‌شود.
- با استخدام واحدهای بازرسی، هزینه‌های نیروی انسانی شرکت‌های بیمه افزایش

می‌یابد.

- شرکت‌های بیمه‌ای که به‌طور مؤثر از تقلب جلوگیری به عمل نمی‌آورند، ممکن است ظرفیت‌های رقابت‌پذیری خود را از دست دهند؛ خصوصاً زمانی که نرخ‌ها به دلیل تقلب افزایش یابد.

داده‌کاوی تکنیک‌های مختلفی را برای استخراج اطلاعات از داده‌ها فر اهم می‌کند. لذا، شرکت‌های بیمه برای کشف روندها و مدل‌ها در میان انبوهی از داده‌ها از تکنیک‌های داده‌کاوی استفاده می‌نمایند(رابرت[1] ۲۰۱۰). استفاده از علم داده‌کاوی در حوزه تقلب چه کمکی در کنترل این هزینه‌ها می‌تواند داشته باشد؟ در صورت وجود و جمع‌آوری داده‌های تقلبات چگونه می‌توان از آن در کاهش این تخلفات در آتی استفاده نمود؟

در این مقاله پس از مقدمه در بخش ادبیات تحقیق به ساختار کلی تقلب در بازارهای مالی و تحقیقات پیشین در این زمینه پرداخته شده است. در بخش سوم مبانی نظری به کار گرفته‌شده در این تحقیق براساس تکنیک‌های مورداستفاده ارائه شده است. در بخش چهارم پس از معرفی متغیرهای مدل با استفاده از داده‌های اخذشده از یکی از شرکت‌های بزرگ بیمه به کمک نرم‌افزار SPSS Clementine خروجی‌های سه روش آنالیز شده‌اند و نهایتاً در بخش پنجم به ارائه نتیجه‌گیری و پیشنهادها براساس روش‌های مورداستفاده در مدل پرداخته شده است.

ادبیات تحقیق

در این مقاله تصویری از روش‌های دسته‌بندی برای کاربرد کشف تقلب بیمه‌ای با استفاده از تکنیک‌های داده‌کاوی ارائه می‌شود. این ساختار در شکل ۱ انعکاس یافته است(نگای[2] و دیگران ۲۰۱۱).

1. Robert
2. Ngai et al

شکل ۱. چارچوب کلی کاربرد داده‌کاوی در کشف تقلبات مالی

تحقیقات متعددی در رشتهٔ بیمه اتومبیل در کشورهای مختلف جهان انجام شده حذف و با توجه به تنوع حجم و نوع داده‌ها، روش‌های مختلفی برای کشف تقلبات بیمه معرفی شده است. این روش‌ها می‌توانند در دو طیف با ناظر یا بی ناظر طبقه‌بندی شوند. در روش‌های با ناظر، نمونه‌هایی از موارد تقلبی و غیر تقلبی موجود است. بر این اساس مدلی طراحی می‌شود که قابلیت تشخیص و پیش‌بینی تقلبی بودن یا نبودن نمونه‌های جدید را دارا باشد. این روش برای بررسی تقلباتی کاربرد می‌یابد که از قبل وجود داشته‌اند. روش‌های بی ناظر به دنبال کشف نمونه‌هایی هستند که کمترین شباهت با نمونه نرمال را نشان دهند.

ویسبرگ و دریگ[1] یک مدل رگرسیون خطی چندگانه را برای انتخاب شاخص های مختلف تقلب طراحی نمودند. بلادجی و دیون[2] نیز مطالعاتی را با استفاده از داده‌های بیمه اتومبیل برای کانادا به انجام رساندند. آرتیس و دیگران[3] از مدل‌های لوجیت

1. Weisberg & Derrig
2. Belhadji & Dionne
3. Artis et al

چندگانه و مدل‌های لوجیت چندگانه تودرتو برای کشف تقلب بیمه استفاده کردند. فوآ و همکاران[1] با ترکیب الگوریتم‌های شبکه های عصبی پس انتشاری مدل تقلبات بیمه را طراحی و معرفی نمودند. براکت و همکارانش[2] علاوه بر شناسایی خسارت‌های تقلبی و دسته‌بندی آنها از روش تحلیل مؤلفه های اصلی برای انتخاب مهمترین شاخصهای مؤثر بر تقلبات بهره برداری کردند. رخا[3] دسته‌بندی بیز ساده و درخت تصمیم مبتنی بر الگوریتم‌های پیش‌بینی و آنالیز الگوهای تقلب را مورد استفاده قرار داد. وی اجرای مدل‌های متریک که از ماتریس اغتشاش حاصل می شد را مبنای مطالعه خود قرار داد.

مبانی نظری

بیمه اتومبیل در میان رشته های مختلف بیمه‌ای، سهم بالایی در پرتفوی حق بیمه دریافتی و حجم خسارت دارد. از لحاظ رخداد تقلب این رشته بیشتر از سایر رشته ها در معرض وقوع است به طوریکه که عدم توجه به این تقلبات هزینه‌های شرکت بیمه را افزایش و درآمدهای آنها را کاهش می‌دهد و در نهایت سبب افت عملکرد شرکت بیمه می‌شود. بررسی فرایند های تقلب با دریافت گزارش ادعای خسارت آغاز می‌شود و پس از بررسی‌های لازم در صورت نداشن شواهدی مبنی بر ارتکاب تقلب طبق روال عادی برای پرداخت خسارت ارجاع داده می‌شوند. لذا برای بررسی، پرونده‌ها به دو دسته ی جعلی و غیر جعلی تقسیم می‌شوند. دسته‌بندی داده‌های مستخرج از پایگاه های اطلاعاتی یکی از موارد مهم در فرایند آنالیز داده‌ها است که می‌تواند برای استخراج مدل‌ها و کلاس‌بندی داده‌ها مورد استفاده قرار گیرد. این روش، شیوه موثری برای ایجاد درک مناسب از داده‌های انبوه محسوب می‌شود(ویلسون[4] ۲۰۰۵). در طی مراحل انجام فرایند داده‌کاوی سه بخش مهم وجود دارد:

- حذف داده‌ها: داده‌های بی ارزش و غیر قابل استفاده حذف می‌شوند.
- فشرده کردن داده‌ها: این مرحله با کد گذاری داده‌ها انجام می‌شود.
- کشف الگوها: الگوهای موجود در پایگاه داده‌ها از جمله دسته‌بندی و.... کشف

1. Phua et al
2. Brockett et al
3. Rekha et al
4. Wilson

می‌شوند.

روش‌های دسته‌بندی می‌تواند براساس مؤلفه ها و معیار های زیر طبقه‌بندی و ارزیابی شود:

- **دقت:** توانایی یک دسته‌بندی مشخص برای پیش‌بینی درست یک داده جدید یا مشاهده نشده ی قبلی، رانشان می‌دهد.
- **سرعت:** سرعت هزینه‌های محاسباتی در تثبیت و استفاده از دسته‌بندی کننده‌ها یا پیش‌بینی کننده‌ها است.
- **توانایی:** توانایی دسته‌بندی کننده برای ایجاد پیش‌بینی های صحیح با داده‌های اغتشاش یا ازدست‌رفته مشخص را منعکس می‌نماید.
- **مقیاس‌پذیری:** توانایی ایجاد کلاس‌های کارآمد با داده‌های انبوه را نشان می‌دهد.
- **تفسیرپذیری:** سطح مفهوم‌رسانی توسط کلاس‌ها می‌باشد، که یک موضوع ذهنی و کمتر قابل‌دسترسی است(کامینس و تنیسون[1] ۱۹۹۲).

رگرسیون لجستیک

رگرسیون لجستیک یکی از ابزارهای مهم داده‌کاوی است و اغلب در مواقعی کاربرد می‌یابد که متغیر پاسخ دوبخشی (باینری) باشد(اودد و رکاچ[2] ۲۰۱۰). در این روش ضرورت دارد متغیر پاسخ عدد صحیح یا نسبی باشد. در این مقاله متغیر وابسته اسمی بوده و مقدار یک، برای تقلبی بودن یک ادعا و مقدار صفر، برای غیر تقلبی بودن آن منظور شده است. شکل کلی مدل رگرسیون لجستیک در رابطه ذیل ارائه شده است.

$$P_i = P(Y|X_i) = \frac{\exp(b_i X_i)}{1 + \exp(b_i X_i)}$$

که در آن P_i احتمال تقلبی بودن یک ادعاست، $_{Y=1}$ به شرط وجود تعدادی از متغیرهای مستقل استد. b_i ضرایبی هستند که باید برآورد شوند. لگاریتم طبیعی این احتمال یا لوجیت $_Y$ به قرار زیر است:

$$\text{Logit } Y = b_0 + b_1 X_1 + \ldots . b_n X_n$$

1. Cummins & Tennyson
2. Oded & Rokach

درخت تصمیم

درخت تصمیم یک ابزار پیش‌بینی با استفاده از مشاهدات است که نمایش نموداری از نتایج ممکن را ارائه می‌دهد. درخت تصمیم می‌تواند با الگوریتم‌های مختلف فراگیری ماشین مانند ID3، CART و C4.5 طراحی شود. پیش‌بینی ها به‌وسیله برگ‌ها و شاخه‌ها به صفات موردبررسی متصل می‌شوند. درواقع یک مجموعه از گره‌های تصمیم با شاخه‌ها به هم متصل می‌شوند و از ریشه گره گسترش پیدا می‌کنند تا برگ و گره‌های نهایی ادامه می‌یابد. الگوریتم رده‌بندی و درخت رگرسیونی(CART) برای شرایط باینری کاربرد دارد. برای هر گره تنها دو شاخه وجود دارد. درخت تصمیم حاصل از این الگوریتم، برای هر تصمیم یک کاوش کامل از تمام متغیرهای موجود و انشعاب مقادیر فراهم و براساس معیار زیر با خرد کردن‌های بهینه، انتخاب می‌شود. اگر $\Phi(s|t)$ معیار خوبی این خرد کردن برای متغیر s در گره t باشد:

$$\Phi(s|t) = 2P_L P_R \sum_{j=1}^{\#class} |P(j|t_L) - P(j|t_R)|$$

$t_L = t$ شاخه چپ گره

$t_R = t$ شاخه راست گره

$$P_L = \frac{t_L \text{ تعداد مشاهدات در}}{\text{تعداد مشاهدات در داده های آموزشی}}$$

$$P_R = \frac{t_R \text{ تعداد مشاهدات در}}{\text{تعداد مشاهدات در داده های آموزشی}}$$

$$P(j|t_L) = \frac{t_L \text{ در } j \text{ تعداد رده}}{t \text{ تعداد مشاهده در گره}}$$

نقطه بهینه زمانی به‌دست می‌آیند که معیار فوق برای تمام حالت‌های ممکن خرد کردن در گره مشخص، حداکثر شود (لاروس[1] ۲۰۰۵).

بیز ساده

این روش دسته‌بندی برای داده‌های با حجم بالا دارای دقت و سرعت بالایی است. در رده‌بندی بیز ساده فرض بر این است که احتمال رخداد یک صفت روی احتمال سایر صفت‌ها، بی‌تأثیر است. در تئوری بیز استخراج احتمال پسین با استفاده از احتمال پیشین امکان‌پذیر است. اگر θ پارامتریک توزیع نامعلوم باشد و $P(\theta)$ احتمال پیشین آن باشد، آنگاه:

$$P(\theta|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|\theta)\,P(\theta)}{P(\boldsymbol{X})}$$

در مواقعی که مجموعه‌ای از پیشامدهای θ (جعلی و غیر جعلی) وجود داشته باشد، از فرضیه حداکثر احتمال استفاده خواهیم کرد.

$$\theta_{NB} = \arg\max_{\theta} \prod_{i=1}^{n} f(X_i|\theta).p(\theta)$$

که در آن X_i متغیرهای مستقل است (شکوری ۱۳۹۰).

آنالیز داده‌ها

داده‌های این مقاله از یکی از شرکت‌های بزرگ بیمه اخذ شده که بعد از پالایش و حذف رکوردهای ناقص مشتمل بر۸۰۰ رکورد گردید. لازم به توضیح است با توجه به حساسیت‌های موجود در ارائه اطلاعات و داده‌های ناقص اخذ رکوردهای بیشتر به‌منظور استفاده در مدل‌ها مقدور نبوده است. ازآنجایی‌که برای مدل‌سازی ترکیب نمودن داده‌های جعلی و غیر جعلی ضروری است، نیمی از موارد یعنی ٤٠٠ رکورد از لیست سیاه (که در آن احتمال وقوع تقلب تقریباً یک فرض شده است) در نظر گرفته شد. این گروه از پایگاه داده و انفورماتیک شرکت مربوط پس از مذاکره و کسب مجوز استخراج شده است و شامل پرونده‌هایی بوده است که پس از بررسی‌های لازم رأی به

1. Larose

تقلبی بودن خسارت اعلامی داده شده است. ٤٠٠ رکورد دیگر مربوط به بیمه‌نامه‌های عادی است. براساس معیار دقت، این مطالعه درصدد است بهترین مدل را شناسایی نماید.

کل داده‌ها به دو بخش ۳۰ درصد و ۷۰ درصد تقسیم گردید؛ بر این اساس ۵۵۲ مشاهده مربوط به داده‌های آموزشی[1] و ۲٤۸ مورد دیگر مربوط به مجموعه داده‌های آزمایشی برای اعتبارسنجی است. شایان ذکر است، داده‌های آموزشی برای ساخت مدل و داده‌های آزمایشی[2] برای اعتبار سنجی یا بررسی دقت مدل مورد استفاده قرار می‌گیرند. داده‌های آزمایشی به‌منزله ناظر به الگوریتم وارد و میزان صحت نتایج را ارزیابی می‌کند. کلیه مراحل در نرم‌افزار Spss Clementine انجام شده است. بر این اساس، هفت متغیر توضیحی به همراه متغیر پاسخ در جدول ۱ نشان داده شده است.

نام متغیر	نقش متغیر	نوع متغیر	نام متغیر	نقش متغیر	نوع متغیر
سابقه بیمه‌ای "X_1"	توضیحی	گسسته	مقدار خسارت "X_5"	توضیحی	پیوسته
تعداد ادعای خسارت "X_2"	توضیحی	گسسته	جنس "X_6"	توضیحی	دودویی
تأخیر در اعلان خسارت "X_3"	توضیحی	پیوسته	سن "X_7"	توضیحی	پیوسته
کروکی "X_4"	توضیحی	دودویی	تقلب "Y"	توضیحی	اسمی

جدول ۱. متغیرهای مورداستفاده در مدل‌ها

در هر سه روش مورداستفاده در این مقاله، تقلبی بودن یا نبودن یک متغیر وابسته یا پاسخ در نظر گرفته می‌شود. اولین متغیر مستقل سابقه بیمه‌ای است. در این زمینه ازنظر و تجربه کارشناسان خبره استفاده شده است و به این دلیل برگزیده شده است که

1.Training
2.Testing

انتظار می‌رود احتمال ارتکاب تقلب توسط بیمه‌گذارانی که سابقه بیمه‌ای بالاتری در شرکت بیمه دارند کمتر باشد.

دومین متغیر توضیحی، تعداد ادعاهای خسارت است که بالا بودن این تعداد احتمال تقلبی بودن خسارت را افزایش می‌دهد. متغیر توضیحی بعدی تأخیر در اعلان خسارت است و فرض بر این است که هر چه این دوره طولانی‌تر شود، احتمال تقلب افزایش خواهد یافت. متغیر دیگری که مورد ملاحظه قرارگرفته است، کروکی در صحنه حادثه است. انتظار می‌رود با حضور پلیس در صحنه تصادف، احتمال تقلب تقلیل یابد. مقدار یک برای وجود کروکی و صفر برای عدم وجود آن در نظر گرفته شده است. پنجمین متغیر مستقل، مقدار خسارت است. ازآنجایی‌که شرکت‌های بیمه برای خسارت‌های با مبالغ بالا حساسیت بیشتری نشان می‌دهند؛ به نظر می‌رسد با افزایش این مبلغ احتمال تقلب کاهش می‌یابد. سن راننده در زمان تصادف و جنسیت راننده دو متغیر مستقل دیگری هستند که در این مطالعه در نظر گرفته شده‌اند. با توجه به متغیرهای مستقل اشاره شده، باید خاطرنشان نمود که هدف اصلی این مطالعه کلاس‌بندی و پیش‌بینی متغیر پاسخ براساس متغیرهای توضیحی است.

دسته‌بندی با روش بیز ساده

بعد از برازش مدل با استفاده از این روش، خروجی در جدول ۲ تحت عنوان جدول احتمالات شرطی نشان داده شده است.

متغیر پاسخ		سابقه بیمه‌ای *1			تعداد ادعای خسارت			تأخیر در اعلان خسارت *2			کروکی			مقدار خسارت *3			جنس			سن *4		
غ	ج		غ	ج		غ	ج		غ	ج		غ	ج		غ	ج		غ	ج		غ	ج
۰,۵	۰,۵	۱	۰,۱۴۰	۰,۶۵۰	۰	۰,۵۵۰	۰	۱	۰,۹۸۵	۰,۹۸	۰	۰,۶۵۲	۰,۱۹۲	۱	۱	۰,۹۸۸	۱	۰,۸۴۲	۰,۸۴۰	۱	۰,۲۱۰	۰,۲۳۲
		۲	۰,۲۸۸	۰,۳۳۸	۱	۰,۳۸۸	۰,۳۲۸	۲	۰,۰۰۸	۰,۰۰۸	۱	۰,۳۴۸	۰,۸۰۸	۲	۰	۰,۱	۲	۰,۱۵۸	۰,۱۶۰	۲	۰,۱۷۰	۰,۲۲۰
		۳	۰,۱۶۵	۰,۰۱۲	۲	۰,۰۵۲	۰,۴۹۸	۳	۰,۰۰۵	۰,۰۰۲				۳	۰	۰,۰۲				۳	۰,۲۴۰	۰,۱۹۰
		۴	۰,۲۹۵	۰	۳	۰,۰۱	۰,۱۶۸	۴	۰,۰۰۲	۰,۰۰۸										۴	۰,۲۰۰	۰,۲۰۸
		۵	۰,۱۱۲	۰	۴	۰	۰,۰۰۸	۵	۰	۰,۰۰۲										۵	۰,۱۸۰	۰,۱۵۰

*1Value 1 for $x \leq 2.6$, value 2 for $2.6 < x \leq 4.2$, value 3 for $4.2 < x \leq 5.8$, value 4 for $5.8 < x \leq 7.4$, value5 for $x > 7.4$.

*2Value 1 for $x \leq 59.6$, value 2 for $59.6 < x \leq 119.2$, value 3 for $119.2 < x \leq 178.8$, value 4 for $178.8 < x \leq 238$, value 5 for $x > 238.4$.

*3Value 1 for $x \leq 121510000$, value 2 for $121510000 < x \leq 364530000$, value 3 for $x > 364530000$

*4value 1 for $x \leq 29$, value 2 for $29 < x \leq 35$, value3 for $35 < x \leq 41$, value 4 for $41 < x \leq 47$, value for $x > 47$.

ج: جعلی // غ: غیرجعلی

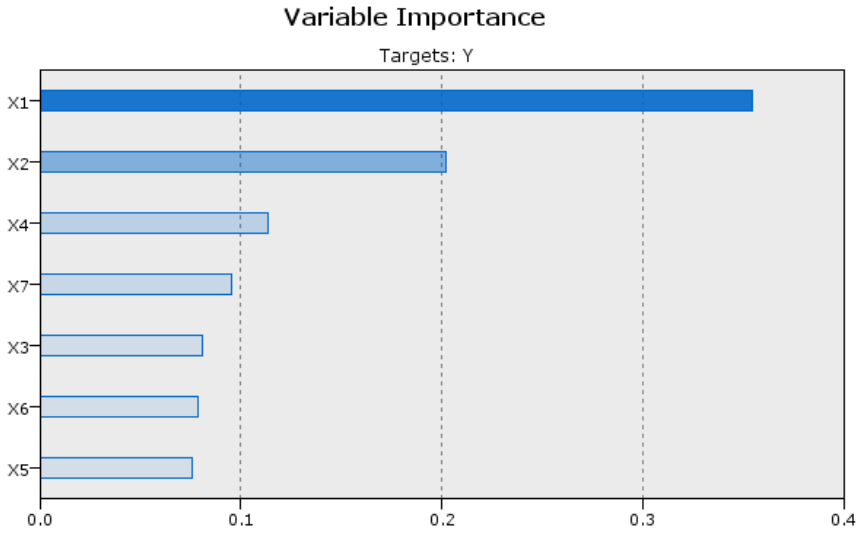
جدول ۲. جدول احتمالات شرطی روش بیز ساده

با استفاده از احتمالات جدول فوق، رده‌بندی داده‌های آزمایشی صورت پذیرفته و نتایج در جدول ۳ نشان داده شده است.

		پاسخ(پیش‌بینی شده)		دقت
		جعلی	غیر جعلی	
متغیر پاسخ (مشاهده شده)	جعلی	۱۲۳	۲	٪۹۸/٤
	غیر جعلی	۱٦	۱۰۷	٪۸۷/۷
کل				٪۹۲,۷٤

جدول ۳. جدول دقت مدل بیز ساده

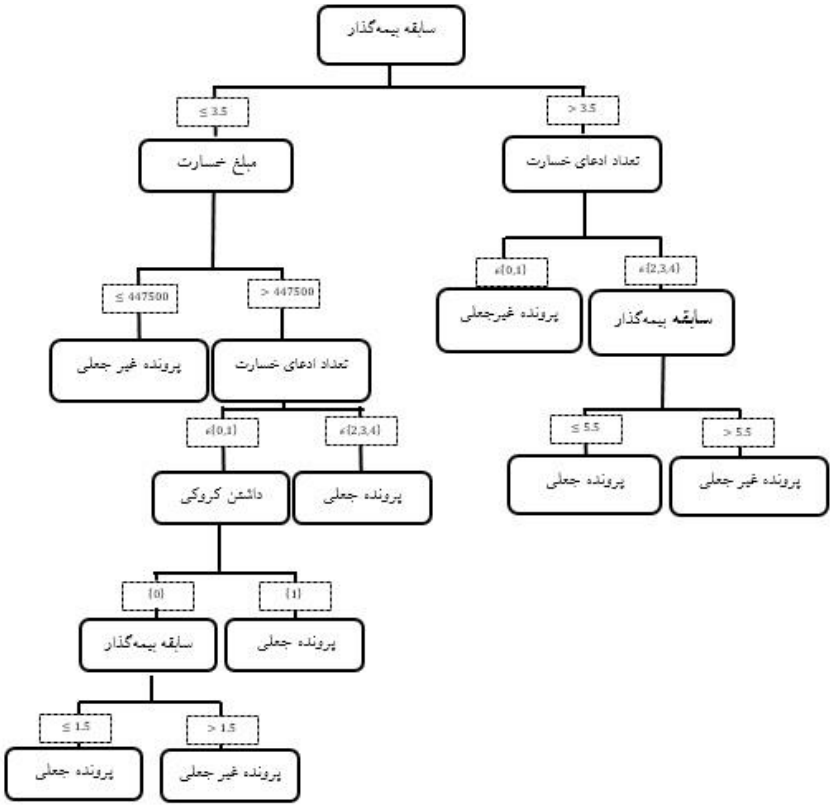
همچنان که ملاحظه می‌شود دقت مدل در شناسایی پرونده‌های جعلی ۹۸/٤ درصد و برای غیر جعلی ۸۷/۷ درصد است، ضمن اینکه دقت مدل برای کل ادعاها اعم از جعلی و غیر جعلی ۹۲/۷٤ درصد می‌باشد. شایان ذکر است، در روش دسته‌بندی، هریک از متغیرها اهمیت متفاوتی در پیش‌بینی متغیر پاسخ دارند. نمودار زیر اهمیت این متغیرها در روش بیز ساده را نشان می‌دهد.



شکل۲. نمودار اهمیت متغیرهای مدل بیز ساده

دسته‌بندی با روش درخت تصمیم

با استفاده از داده‌های مربوطه، درخت تصمیم در شکل ۳ نشان داده شده است.



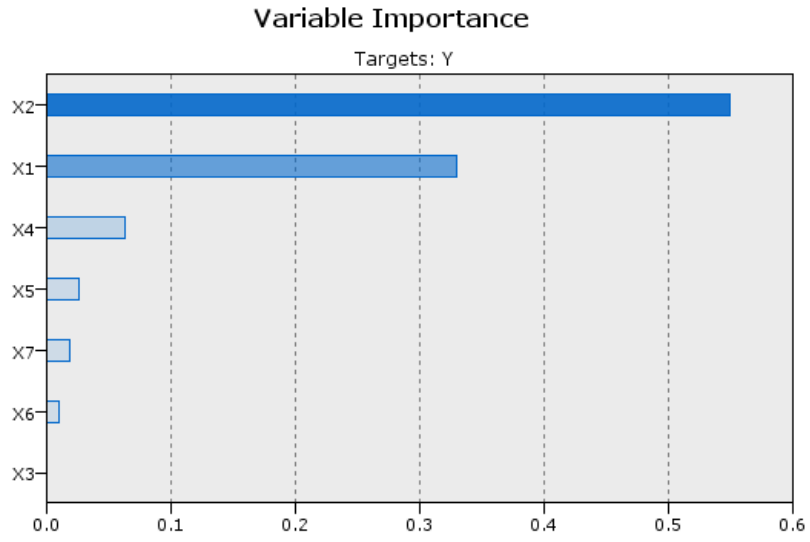
شکل ۳. درخت تصمیم

براساس نمودار فوق می‌توان یک مشاهده نمونه را کلاس‌بندی نمود(پیش‌بینی متغیر پاسخ آن مشاهده). برای مثال، اگر سابقه بیمه‌ای کمتر از ۳/۵ سال و مقدار خسارت بیش از ٤٤٧٥٠٠ ریال باشد می‌توان گفت متغیر پاسخ صفر یا غیر جعلی می‌باشد. به همین صورت قوانین دیگری را می‌توان به‌طور مستقیم از نمودار درخت تصمیم استخراج نمود. سرانجام با به‌کارگیری داده‌های آزمایشی، دقت مدل طبق جدول زیر استخراج می‌شود.

		پاسخ(پیش‌بینی شده)		دقت
		جعلی	غیر جعلی	
متغیر پاسخ (مشاهده شده)	جعلی	۱۱۹	٦	٪۹۵,۲
	غیر جعلی	١٤	۱۰۹	٪۸۸,٦۱
کل				٪۹۲,۷٤

جدول ٤. جدول دقت مدل درخت تصمیم

دقت مدل برای موارد جعلی و غیر جعلی به ترتیب ۹۵/۲ و ۸۸/٦۱ درصد است. ضمن اینکه مدل درخت تصمیم برای کل ادعاها، دقتی معادل ۹۲/۷٤ درصد دارد. اهمیت متغیرها در این روش در قالب نمودار زیر قابل‌ارائه است.



شکل٤. نمودار اهمیت متغیرهای مدل درخت تصمیم

دسته‌بندی با روش رگرسیون لجستیک

در این مرحله رگرسیون لجستیک پیشرو با استفاده از داده‌های آموزشی به کار گرفته شده است (متغیرهایی که تأثیر بیشتری داشته‌اند، به ترتیب وارد مدل شده‌اند). پارامترها

با استفاده از روش حداکثر درستنمایی تخمین زده شده‌اند. این برآوردها در جدول ۵ ارائه شده‌اند.

	متغیر	پارامتر	انحراف استاندارد	معنی‌داری
گام اول	X_2	۸۸۱/۲	۲٤٥/۰	۰,۰۰۰
	ثابت	۳/۲۷٤-	۲۹٦/۰	۰,۰۰۰
گام دوم	X_1	۱/۷۸٤-	۲۱۰/۰	۰,۰۰۰
	X_2	۷۱۷/۳	٤۲۱/۰	۰,۰۰۰
	ثابت	۹۲۱/۰	٤۸۳/۰	۰٥۷/۰
گام سوم	X_1	۱/۹٥۱-	۲٤۹/۰	۰,۰۰۰
	X_2	۸۷٦/۳	٤۷۸/۰	۰,۰۰۰
	X_4	۳۷۰/۲	٤٦۹/۰	۰,۰۰۰
	ثابت	۰/۲۱۰-	٥۸۷/۰	۰۷۲۱/۰

جدول ۵. برآورد پارامترهای مدل رگرسیون لجستیک

در هر گام متغیرهای با بالاترین تأثیر، اضافه و مابقی حذف شده‌اند. در ادامه مقیاس‌های معنی‌داری مدل در جدول ٦ ارائه می‌شود.

	Likelihood logarithm	Cox & Snell R-square	Nagelkerke R-square
گام اول	۳۹۸/۲۳٦	۰/٤۸٦	۰/٦٤۸
گام دوم	۱۸۰/۱۹٤	۰/٦٥۳	۰/۸۷۱
گام سوم	۱٤۹/۱۲۸	۰/٦۷۲	۰/۸۹۷

جدول٦. معیار برازش مدل رگرسیون لجستیک

نخستین معیار، احتساب دو برابر لگاریتم تابع درستنمایی است. این معیار، دال بر معنی‌داری ضرایب متغیرهای مستقل می‌باشد. کاهش این معیار در جدول، نشان‌دهنده بهبود مدل در طی گام‌های پیش رفته می‌باشد. مقادیر مربع R کاکس وسل و مربع R ناجل کرک نیز محاسبه شده است که درصد تغییرات متغیر وابسته مدل را نشان می‌دهد.

$$\text{logit}(Y) = -0.210 - 1.951\,X_1 + 3.876\,X_2 + 2.370\,X_4$$

مدل فوق برای رده‌بندی ادعاها مورد استفاده قرار می‌گیرد و نتایج آن در قالب جدول ۷ نشان داده شده است.

		پاسخ (پیش‌بینی شده)		دقت
		جعلی	غیر جعلی	
متغیر پاسخ (مشاهده شده)	جعلی	۱۲۱	٤	٪۹٦٫۸
	غیر جعلی	۱۲	۱۱۱	٪۹۰٫۲
کل				**٪۹۳٫۵۵**

جدول۷. دقت مدل رگرسیون لجستیک

ملاحظه می‌شود که مدل رگرسیون لجستیک، دارای دقت ٪۹۳/۵۵ برای کل ادعاها است که بهتر از پیش‌بینی تصادفی ٪۵۰ است. لذا با استفاده از سه متغیر مستقل در مدل لجستیک به نحو مناسب‌تری می‌توان به توضیح مدل پرداخت. مقادیر متغیرهای مستقل می‌توانند برای تخمین احتمال متغیر پاسخ، در مدل فوق‌الذکر جایگذاری شوند. برای نمونه، اگر مقادیر متغیرهای x1 و x2 و x4 به ترتیب برابر ٤و ۳ و ۱ باشد:

$$\text{logit(Y)} = -0.210 - 1.951 * 4 + 3.876 * 3 + 2.370 * 1 = 5.984$$

$$e^{5.984} = 397.0253 \quad P_i = \frac{e^{5.984}}{1+e^{5.984}} = 0.99 \qquad \text{logit(Y)} = 5.984$$

بنابراین؛ احتمال جعلی بودن برای ادعای یک شخص برابر ۰٫۹۹ است. برای مقایسه همان حالت قبلی تنها با این تفاوت که تعداد ادعای خسارت فرد صفر باشد. در این حالت خواهیم داشت:

$$\log \text{odds} = -0.210 - 1.951 \times 4 + 3.876 \times 0 + 2.370 \times 1 = -5.644$$

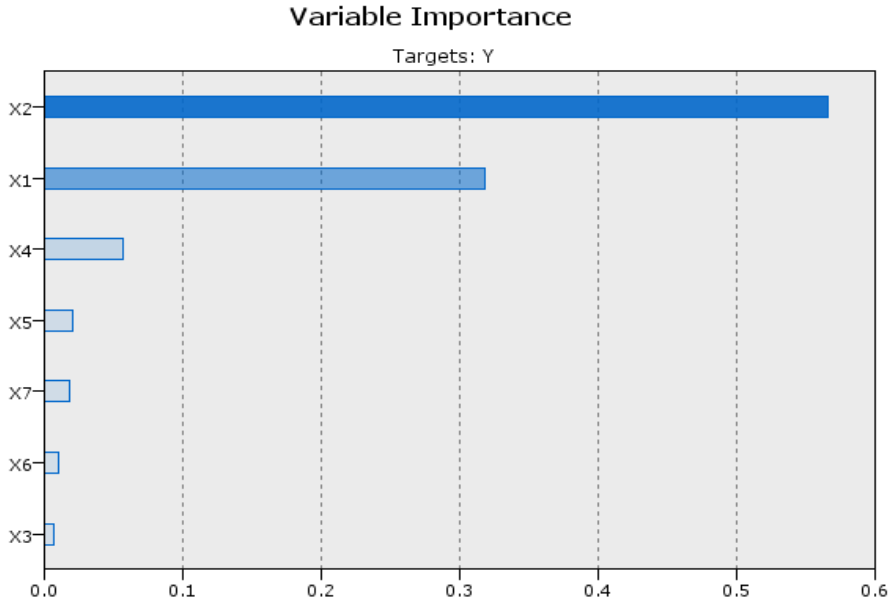
درنتیجه:

$$P_i = \frac{0.0035}{1.0035} = 0.00349$$

بنابراین؛ احتمال اینکه ادعای شخص جعلی باشد به 0/00349 کاهش یافته است. چنانچه میانبر ۵۰:۵۰(احتمال جعلی و غیرجعلی هردو $\frac{1}{2}$) برای حالت دوم در نظرگرفته شود؛ ضرورتی برای پیگیری‌های بیشتر توسط واحد بازرسی وجود ندارد؛ زیرا طبق

جدول احتمالات؛ احتمال جعلی بودن صفر است.

اهمیت متغیرها برای مدل رگرسیون لجستیک در نمودار زیر نشان داده شده است.



شکل ۵. نمودار اهمیت متغیرهای مدل رگرسیون لجستیک

نتایج و پیشنهادها

داده‌کاوی و روش‌های مختلف آن به‌منزله علمی در حال رشد می‌تواند کاربرد وسیعی در صنعت بیمه داشته باشد؛ به‌نحوی‌که استخراج نتایج کاربردی و مدل‌سازی از انبوه داده‌ها خود زمینه‌ای رو به رشد در مدیریت هر چه بیشتر ریسک‌های یک شرکت بیمه است. اعمال سه تکنیک داده‌کاوی بر روی داده‌های واقعی و ارزیابی نتایج آن‌ها مؤید آن است که رگرسیون لجستیک(نسبت به سایر روش‌ها) از دقت بیشتری برخوردار است. از طرف دیگر، از میان متغیرهای مستقل سابقه بیمه‌ای، تعداد ادعاهای خسارت و داشتن کروکی(نسبت به سایر متغیرها) از اهمیت بالاتری برخوردار هستند. یافته‌های این مطالعه تأیید نمود که بیشتر تقلب‌های بیمه اتومبیل در افراد میان‌سال رخ می‌دهد و با افزایش میزان خسارت، احتمال جعلی بودن کاهش می‌یابد.

در روش‌های درخت تصمیم، رگرسیون لجستیک و بیز ساده به ترتیب چهار، سه و

هفت متغیر به کار گرفته شده‌اند. همان‌طور ملاحظه گردید در هر سه مدل متغیرهای سابقه بیمه‌ای، تعداد ادعاهای خسارت و وجود یا عدم وجود کروکی برای صحنه حادثه از اهمیت بالایی برخوردار می‌باشند. شایان‌ذکر است که روش رگرسیون لجستیک به محاسبات کمتری نسبت به سایر روش‌ها نیاز دارد. این در حالی است که در مواقعی که سرعت تصمیم‌گیری اهمیت دارد، روش درخت تصمیم دید شماتیک مناسب‌تری را به‌دست می‌دهد. روش بیز نیز زمانی که احتمال رخداد یک شاخص وابسته به احتمال رخداد سایر متغیرها وابسته نیست، می‌تواند از دقت خوبی برخوردار باشد. از بررسی‌های صورت گرفته برای کشف تقلب این شاخه از صنعت بیمه می‌توان برای مدل‌سازی و کشف تقلبات سایر رشته‌ها نظیر بیمه درمان نیز استفاده نمود. سایر روش‌های شناسایی تقلب در حوزه داده‌کاوی ازجمله سیستم خبره فازی و ماشین بردار پشتیبان می‌تواند در این زمینه استفاده شود.

مدل‌های تقلب، برای مؤثر بودن در کاهش تقلب‌های بیمه‌ای(به‌ویژه در یک بازه زمانی بلندمدت) نیاز به، بروز رسانی مداوم، جهت سازگاری رفتارهای تقلب با الگوهای جدید دارند. در صورت وجود اطلاعات، متغیرهای مستقل دیگری را نیز می‌توان به مدل افزود. ازجمله این متغیرها می‌توان به زمان تصادف، درون‌شهری یا برون‌شهری بودن منطقه تصادف، تعداد مسافران در هنگام حادثه و تعداد سال‌هایی که از زمان اخذ گواهینامه رانندگی گذشته است، اشاره نمود. استفاده ازنظر خبرگان نیز می‌تواند به انتخاب شاخص‌های مناسب تقلب کمک نماید.

منابع

شکوری، مرتضی و همکاران. ۱۳۹۰ . شناسایی تقلب در بیمه اتومبیل با استفاده از داده‌کاوی.، *پژوهشنامه بیمه، شماره ۳، صفحه ۱۰۳-۱۲۸.*

Artis, M., Ayuso, M. and Guillen, M.(2002), "Detection of automobile insurance fraud with discrete choice models and misclassified claims", *Journal of Risk and Insurance*, 325-340.

Belhadji, D. B and Dionne, G. (1997). Development of an expert system for the automatic detection of automobile insurance fraud, *Risk Management* Chair, HEC-Montreal

Bhowmik. Rekha. (2011) Detecting Auto Insurance Fraud by Data Mining Techniques, *Journal of Emerging Trends in Computing and Information Sciences.156-162*.

Brockett, P.L., Xia, X and Derrig, R. A. (1998). Using kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud, *The Journal of Risk and Insurance, 245-74.*

Cummins, J. D and Tennyson, S. (1992). Controlling automobile insurance costs, *Journal of Economic Perspectives, 95-115.*

E.W.T.Ngai, Yong Hu, Y.H,Wong, Yijunchen, xin sun; The application of data mining techniques in financial fraud detection: A classification framework and academic review of literature, *decision support system 50(2011). 559 -569*

Holton Wilson.J.(2005).An Analytical Approach to Detecting Insurance Fraud Using Logistic Regression, *Journal of Finance and Accountancy*.

Larose D.T (2005). *Discovering Knowledge InData*, Central Connecticut university, Canada, Wiley publication

Oded, M and Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook*. Prof, Springer New York Dordrecht Heidelberg London, DOI 10.1007/978-0-387-09823-4.

Phua, C., Alahakoon, D. and Lee, V. (2004). Minority report in fraud detection: *classification of skewed data, Sigkdd Explorations, vol. 6, no. 1, 50-9.*

Roberts, T. (2010). Improving the Defense Lines: The Future of Fraud Detection in the Insurance Industry (with Fraud Risk Models, Text Mining, and Social Networks), *SAS Global forum, Insurance*.

Subelj, L., Furlan, S.,Bajec, M. (2011). An expert system for detecting automobile insurance fraud using social network analysis, *Expert Systems with Applications 38 ,1039–1052*.

Weisberg, H. I and Derrig, R. A. (1993), Quantitative methods for detecting fraudulent automobile bodily insurance claims, *AIB Cost Containment/Fraud Filing, 49-82.*