دانشگاه صنعتی امیرکبیر
دانشکده مهندسی کامپیوتر و فناوری اطلاعات

سیستم تشخیص تقلب در بیمه

استاد راهنما : دکتر رضا صفابخش

نگارنده : علیرضا حیدری

۲۵ دی ۱۳۹۶

چکیده

در سبک زندگی امروزه، با توسعه انواع موانع و بلایای طبیعی و غیرطبیعی نیاز است تا راه‌حل هایی جهت بهبود مشکلات حاصل از این نوع بلایا ایجاد شود. صنعت بیمه جهت حل اینگونه بیمه ها ایجاد شد و باعث شد تا با ایجاد قراردادی بین یک بیمه‌گر و بیمه گذار بتوان راحت تر به حل این موارد پرداخت.

صنعت بیمه با عقد قراردادی بین طرفین باعث می‌شود تا بیمه‌گر با پرداخت مبلغی از پرداخت مبلغ بیشتری در هنگام وقوع حادثه پیشگیری کند. این مبلغ را بیمه‌گر از مجموع پول‌های بدست آمده از قراردادهای کلی پرداخت می‌کند.

باوجود این تلاش‌ها ممکن است در بسیاری از مواردی که بیمه‌ها ثبت می‌شوند در مراحل مختلف آن توسط افراد بیمه‌گر تخلفاتی صورت بگیرد که با تشخیص آن می‌توان به ادامه کار صنعت بیمه و افزایش اطمینان طرفین به یکدیگر کمک کرد.

در این مقاله به مواردی که جهت ساخت سیستمی برای تشخیص تقلب در صنعت بیمه به آن‌ها نیازمندیم میپردازیم و سعی می‌کنیم این سیستم بهترین عملکرد ممکن را داشته باشد.

فهرست مطالب

[۱] screening and identification
[۲] investigation
[۳] logistic regression
[۴] order simple
[۵] tree decision

۱ مقدمه

امروزه فروش بیمه اهمیت خود را افزایش داده است و انواع توسعه ها در این زمینه در حال شکل گیري است. با توسعه این صنعت و انتقال موارد آن در بستروب و انواع دیگر بسترها، مشکلات فراوانی در پی آن بوجود می آید. یکی از این موارد امکان تقلب[۶] در این سیستم است به گونه اي که از متدوال ترین نوع این موارد، میتوان به گرفتن خسارت از بیمه، آتش سوزي عمدي، به صورت مکرر اشاره کرد

فروش و ثبت عملیات هاي بیمه‌اي به صورت داده منجر به دسترسی داشتن به اطلاعات عملیات هاي بیمه‌اي یک کاربر میشود. میتوان با بررسی داده‌هاي هرکاربر و نحوه رفتار او در رابطه با خرید بیمه و استفاده از آن به نحوه عملکرد پی برد و در صورت وقوع تقلب، آن را تشخیص داد. تشخیص این تقلب به الگوریتم‌هاي خاص خود و داده کاوي نیازمند است.

در ادامه سعی می‌شود تا الگوریتم‌های لازم شناسایی و استفاده شوند و بتوان از آن ها در سیستم های تشخیص تقلب استفاده کرد.

---

[۶] fraud

۲ صنعت بیمه

بیمه[۷] سازوکاری است که طی آن یک بیمه‌گر، بنا به ملاحظاتی تعهد می‌کند که زیان احتمالی یک بیمه‌گذار را در صورت وقوع یک حادثه در یک دوره زمانی خاص، جبران نماید یا خدمات مشخصی را به وی ارائه دهد؛ بنابراین، بیمه یکی از روشهای مقابله با ریسک است.

طی یک قرارداد بیمه، ریسک مشخصی از یک طرف قرارداد (که بیمه‌گذار نامیده می‌شود) به طرف دیگر (که بیمه‌گر نامیده می‌شود) منتقل می‌گردد. بنا به تعریف، بیمه‌گر شخصی حقوقی است که در مقابل دریافت حق بیمه از بیمه‌گذار، جبران خسارت یا پرداخت مبلغ مشخصی را در صورت بروز حادثه تعهد می‌کند. در مقابل، بیمه‌گذار شخصی حقیقی یا حقوقی است که با پرداخت حق بیمه، جان، مال یا مسوولیت خود یا دیگری را تحت پوشش بیمه قرار می‌دهد.

به موجب قانون بیمه ایران، بیمه عبارت است از قراردادی که به موجب آن یک طرف (بیمه گر) تعهد می‌کند در ازای پرداخت وجه یا وجوهی از طرف دیگر (بیمه گذار) در صورت وقوع یا بروز حادثه خسارت وارده بر او را جبران نموده یا وجه معینی را بپردازد. متعهد را بیمه گر، طرف تعهد را بیمه گذار و وجهی را که بیمه گذار به بیمه‌گر می‌پردازد حق بیمه و آنچه را که بیمه می‌شود موضوع بیمه نامند.

۱.۲ انواع بیمه

در یک تقسیم‌بندی کلی بیمه به دو دسته بیمه های اجتماعی و بیمه‌های بازرگانی تقسیم‌بندی می‌شود. مبحث تقلب به طور عمومی در بیمه‌های بازرگانی مطرح می‌شود که از انواع آن به موارد زیر می‌توان اشاره کرد:

- بیمه آتش‌سوزی
- بیمه حمل و نقل
- بیمه مسافرتی – بیمه سفر
- بیمه عمر
- بیمه حوادث
- بیمه بدنه اتومبیل

---

[۷] Insurance

- بیمه شخص ثالث
- بیمه درمان
- بیمه سرطان
- بیمه کشتی
- بیمه هواپیما
- بیمه مهندسی
- بیمه پول
- بیمه مسوولیت
- بیمه اعتباری

۲.۲ تقلب در صنعت بیمه

در سال ۲۰۰۲، موسسه تحقیقاتی فرانک به سفارش انجمن بیمه‌گران بریتانیا، تحقیقی با شرکت ۲۰۰۰ نفر انجام داد. هدف اصلی این تحقیق سنجش دیدگاه مردم در خصوص ادعاهای تقلبی در صنعت بیمه بود. هدف دیگری که از طراحی این تحقیق دنبال می‌شد، این بود که تقلب و سوءاستفاده از بیمه را جزو اقدامات خلاف قانون در جامعه مطرح کند. نتایج این تحقیق نشان می‌دهد که بخشی از تقلب و سوء استفاده در بیمه، ناشی از ناآگاهی و عدم شناخت مردم درباره چیزی است که درست است. بیشتر کسانی که در این تحقیق مورد پرسش قرار گرفته‌اند، درباره آنچه که رفتار درست تلقی می‌شود، اطلاع دقیقی نداشته‌اند. نتایج این تحقیق نشان داد که :

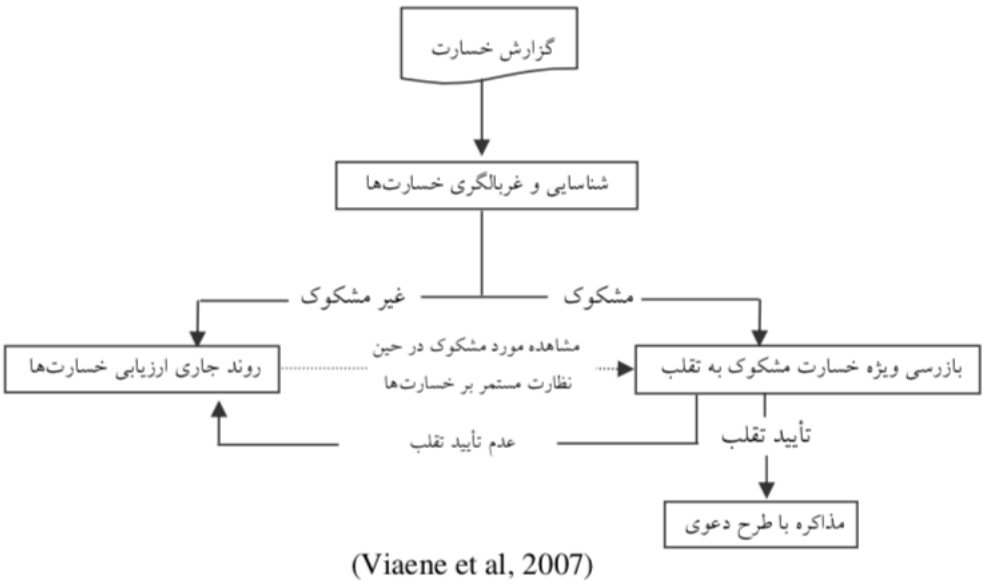
- اگرچه بیشتر پرونده‌ها و دعاوی بیمه ای درست و صحیح است، تقریبا نیمی از پرسش‌شوندگان احتمال تقلبی بودن یک ادعا را رد نکرده‌اند.
- احتمال وقوع تقلب بیمه‌ای بیشتر از سایر سوء استفاده‌هااست.
- در میان افراد شرکت کننده در تحقیق، در خصوص درست یا نادرست‌بودن اقداماتی مانند خریدن مال مسروقه یا رانندگی در حالت مستی، دیدگاه‌های متفاوتی وجود دارد.

کلاهبرداری در بیمه اتومبیل از روش‌های مختلفی صورت می‌گیرد، برخی از شرکت‌ها اغراق در اعلام میزان خسارت و برخی دیگر سایر فعالیت‌های هدفمند، مانند تصادفات ساختگی، اسناد جعلی و ارائه اطلاعات نادرست را به‌عنوان مصادیق تقلب درنظر می‌گیرند. بعضی از کلاهبرداری‌ها در صنعت بیمه کاملا آگاهانه و عمدی است. بیمه‌گذار ممکن است موجبات بروز خسارتی را فراهم آورد تا بدین طریق از محل بیمه‌نامه خود منفعتی کسب کند.

به‌طور کلی، بیمه‌گذاران در دو موقعیت مرتکب تقلب می‌شوند: مورد اول، شرایطی است که در آن، فرد آگاهانه سعی در ایجاد خسارت یا اغراق در میزان و نوع خسارت دارد؛ به عنوان مثال، در یک سانحه تصادف ممکن است فرد بیمه‌گذار با توجه به حق‌بیمه‌ای که برای سالیان متمادی به شرکت بیمه پرداخت نموده است درصدد بهره‌برداری از فرصت برآید و با تجمیع کلیه زیان‌های پیشین با خسارت فعلی سعی در کسب موقعیت مالی بهتر کند. مورد دوم که ممکن است منجربه خسارت‌های جعلی گردد، مواردی است که بیمه‌گذار به صرف داشتن بیمه‌نامه احتیاط کمتری می‌کند. بدین معنی که گرچه ممکن است شخص قصد ایجاد خسارت یا اغراق در میزان آن را نداشته باشد، با این حال اقدام به انجام فعالیت‌هایی می‌کند که در صورت نداشتن بیمه‌نامه، این فعالیت‌ها را انجام نمی‌داد.

۳ مراحل شناسایی تقلب

یک مدل معمول و رایج برای تشخیص تقلب در نمودار ۱ قابل مشاهده است. مراحل شامل شناسایی و غربال‌گری، تحقیق و بررسی، مذاکره با بیمه‌گذار یا طرح دعوی است که در روند ارزیابی خسارت ها اجرا می‌شود. روند ارزیابی خسارت‌ها با رخداد یک حادثه و اعلام گزارش به شرکت بیمه آغاز و با پرداخت یا عدم پرداخت خسارت پایان می‌یابد. عواملی چون عدم تمایل به ارائه اطلاعات صحیح از نشانه‌های کلاهبرداری است که در صورت اثبات تخلف منجربه عدم پرداخت خسارت می‌گردند.



(Viaene et al, 2007)

شکل ۳٬۱: مراحل تشخیص تقلب

۱.۳ شناسایی و غربال‌گری[۸]

این مرحله جهت شناسایی و تفکیک خسارت‌های مشکوک به تقلب است. خسارت‌هایی که از این مرحله گذر می‌کنند، طبق روال معمول و با حداقل هزینه‌های اداری ارزیابی می‌شوند، اما خسارت‌هایی که این امر مستلزم صرف زمان، هزینه و نیروی انسانی بیشتر است. بدون وجود سیستم های هوشمند، بررسی خسارت ها تنها براسا

[۸] screening and identification

اطلاعات موجود درمورد بیمه‌گذار و خسارت وارده ممکن است. اما ازآنجا که معمولا جستجوی دستی در پرونده‌ها و موارد مشابه گذشته، بسیار مشکل و زمان‌بر است، کارشناسان خسارت باید براسا اطلاعات بسیار محدود و اغلب با اتکا به تجربیات به تصمیم‌گیری بپردازند. معمولا داده‌ها از سه طریق قابل دستیابی اند:

- گردآوری داده‌ها در مرحله صدور بیمه‌نامه از طریق فرم‌هایی که توسط بیمه‌گذاران پر می‌شوند، اطلاعاتی درمورد بیمه‌گذاران و اتومبیل بیمه شده از قبیل تاریخ تولد، نشانی، نوع اتومبیل، تاریخ اخذ گواهینامه رانندگی، نوع کاربری اتومبیل و ... که غالبا عوامل موثر در شناسایی ریسک تحت پوشش و تعیین نرخ مناسب در محاسبه حق بیمه است را برای بیمه‌گر فراهم می‌آورد. همچنین این اطلاعات در آینده به همراه جزئیات خسارت در تکمیل پروفایل مشتریان استفاده می‌شود.
- گردآوری داده‌ها در مرحله ارزیابی خسارت‌ها که توسط کارشناسان مربوطه جهت پرداخت خسارت استفاده می‌شود، داده‌هایی از قبیل زمان، مکان، شرح وقوع و علت حادثه، شاهدان و مشخصات اتومبیل‌های ثالث(نوع، سال ساخت، سازنده) و ... را در اختیار شرکت بیمه قرار می‌دهند.
- گردآوری داده‌های موجود در پایگاه داده‌هایی که در صنعت اتومبیل اطلاعات مربوط به خودروها، مدل‌های آنها و هزینه خرید و تعمیر قطعات مختلف را در اختیار کارشناسان خسارت قرار می‌دهند. به کمک چنین پایگاه‌های داده‌ای، ارزیابان خسارت می‌توانند به سرعت مبالغ پرداخت را محاسبه کنند.

۲.۳ تحقیق و بررسی[۹]

تشخیص تقلبی‌بودن ادعا برعهده ارزیاب خسارت است که وی براساس تجربه، توانایی و خلاقیت خود این فرآیند را انجام می‌دهد. براساس تحقیقات تنیسن و سالساس[۱۰] روش‌های رایج رسیدگی خسارت‌ها عبارت‌اند از: بازدید از محل، بررسی پیشینه، گزارش‌های واحدهای ویژه بازرسی و نظارت بر فعالیت‌های بیمه‌گذار

۳.۳ مذاکره با بیمه‌گذار یا طرح دعوی

اکثریت شرکت‌های بیمه ترجیح می‌دهند به همان روش‌های سنتی به بازرسی خسارت‌ها جهت تشخیص تقلب بپردازند، ولی با این‌حال در برخی از موارد نیاز به دادگاه خواهد بود. اما دعوی قضایی و بازرسی های ویژه معمولا مستلزم صرف هزینه و زمان زیادی است. معمولا شرکت‌های بیمه به‌دلیل تاثیری که ممکن است طرح دعوی در دادگاه و شکست احتمالی در آن، بر شهرت شرکت در بازار داشته باشد تمایلی به طرح دعوی در دادگاه‌ها ندارند.

[۹] investigation

[۱۰] Salsas, & Tennyson ۲۰۰۲

۴ فراگیری ماشین و داده‌کاوی

با توجه به فراوانی اطلاعات در امروزه و در عصر حاضر مدیریت این داده های فراوان نیازمند دانش جدیدی است. اطلاعات فراوانی در قالب پایگاه‌های داده ذخیره شده است که تبدیل آن‌ها به دانش مورد نیاز جهت تصمیم‌گیری، نیازمند ابزارهای جدیدی است. روش‌های آماری برای تحلیل داده‌ها بیشتر برپایه استخراج شاخص‌های کمی استوار است. اگرچه این روش‌ها به صورت غیرمستقیم ما را به دانش مورد نیاز جهت تصمیم‌گیری سوق می‌دهند، اما درنهایت تفسیر نتایج آنها نیازمند تحلیل‌های انسانی است. روش‌های نوین تحلیل داده باید به دانش لازن و قابلیت تصمیم‌گیری براساس داده‌ها تجهیز شوند. جهت دستیابی به این هدف* محققین به ارائه ایده‌های جدیدی از فراگیری ماشین[۱۱] پرداخته اند. با توجه به این ایده‌ها وظیفه فراگیری ماشین، تبدیل داده‌ها به دانش تصمیم‌گیری خواهد بود. همچنین براساس این ایده‌ها، ضرورت پیدایش یک حوزه تحقیقاتی جدید که داده‌کاوی[۱۲] نام گرفته به وجود آمده است.

داده کاوی فرآیند کشف الگوها در داده‌ها است. این فرآیند باید خودکار یا نیمه‌خودکار باشد. الگوهای شناسایی شده باید معتبر بوده و برای ما مزایایی از جمله مزایای اقتصادی داشته باشند. همچنین داده ها باید همواره در قالب کمیت های معتبر ارائه شوند. استفاده از مدل های ریاضی برای شناسایی تقلب، این امکان را به متخصصین شرکت‌های بیمه می‌دهد که با صرف زمان و هزینه کمتری تشخیص دهند که ادعای خسارت اعلام شده از لحاظ آماری مشکوک به تقلب است یا خیر. در ادامه سه روش رگرسیون لجستیک، بیز ساده و درخت تصمیم‌گیری که از ابزارهای رایج داده‌کاوی است معرفی و با استفاده از این روش‌ها مدل‌هایی برای شناسایی و دسته‌بندی خسارت‌های تقلبی بر روی داده‌های واقعی تعریف خواهد شد.

---

[۱۱] Learning Machine

[۱۲] Mining Data

۵ مطالعه تجربی

برای ساختن یک مدل ریاضی، نیاز به داده‌هایی از هر دو دسته ادعاهای جعلی و غیرجعلی داریم. در این بخش از این اطلاعات استفاده میکنیم و با سه الگوریتم ذکر شده به شناسایی موردهای دارای تقلب و مشکوک به آن می‌پردازیم.

۱.۵ متغیرهای مورد استفاده در مدل

در هریک از سه مدل مورد استفاده در شناسایی تقلب که در قبل بیان شد، جعلی یا غیرجعلی بودن یک پرونده، به عنوان متغیر وابسته درنظرگرفته می‌شود. مقدار ۱ برای متغیر وابسته به معنای جعلی‌بودن پرونده خسارت و مقدار ۰ به معنای غیرجعلی بودن آن پرونده است. در این مطالعه، فرآیند شناسایی تقلب بااستفاده از شش متغیر مستقل صورت گرفته است. برای انتخاب متغیرها از روش ترکیبی استفاده شده است. روش ترکیبی از ترکیب دو روش پیش‌رونده و پس‌رونده تشکیل شده است. در این روش در هر مرحله از ورود متغیرها به مدل، متغیر که کمترین ارتباط را با متغیر وابسته داشته باشد، حذف و متغیری که بیشتری ارتباط را داشته باشد، انتخاب می‌شود. همچنین برای بهترشدن نتایج از نظرات کارشناسان بیمه اتومبیل نیز بهره گرفته شده است.

اولین متغیر مستقل، سابقه بیمه‌ای هریک از بیمه‌گذاران در شرکت بیمه است. این متغیر به این دلیل انتخاب شده است که انتظار می‌رود احتمال ارتکاب تقلب توسط بیمه‌گذارانی که سابقه بیمه‌ای بیشتری در یک شرکت بیمه دارند، کمتر باشد. بنابراین یک رابطه معکوس بین این متغیر و متغیر وابسته وجود دارد.

دومین متغیر مستقل، تعداد ادعاهای خسارت بیمه‌گذاران در طول دوره سابقه بیمه است. تعداد ادعاهای خسارت بیشتر توسط یک بیمه‌گذار می‌تواند به این معنا باشد که بیمه‌گذار از بیمه‌نامه به‌منظور مقاصد سودجویانه استفاده کرده باشد. ازاین‌رو بین این متغیر و متغیر وابسته یک رابطه مستقیم وجود خواهد داشت.

سومین متغیر مستقل، فاصله زمانی بین وقوع حادثه تا اعلام خسارت به شرکت بیمه از سوی بیمه‌گذار است. فرض شده است که هرچه این فاصله زمانی طولانی‌تر باشد، احتمال تقلب بیشتر افزایش خواهد یافت. بنابراین یک رابطه مستقیم بین این متغیر و متغیر وابسته وجود خواهد داشت.

چهارمین متغیر مستقل، وضعیت کروکی خسارت رخ داده است. مقدار ۱ برای این متغیر به معنی نداشتن کروکی و مقدار ۰ به معنی داشتن کروکی است. این متغیر به این‌دلیل انتخاب شده است که با حضور پلیس در صحنه حادثه، شانس تقلب از قبیل صحنه‌سازی کاهش می‌یابد.

۲.۵ روش رگرسیون لجستیک[۱۳]

زمانی که متغیر وابسته، متغیری کیفی با دو سطح باشد، مدل‌های رگرسیون معمولی قابل استفاده نیستند. در این‌گونه موارد معمولا از رگرسیون لجستیک استفاده می‌شود. مدل رگرسیون لجستیک به‌این‌صورت تعریف می‌شود:

$$logit(\frac{p}{۱-p}) = \beta_۰ + \beta_۱ X_۱ + ... + \beta_d X_d \qquad (۵^{'}۱)$$

که در این مدل هاX متغیر های وابسته و p احتمال مشاهده مقدار ۱ برای متغیر وابسته به شرط مشاهده مقادیر متخلف x است.

ضرایب رگرسیونی در این حالت با فرض دوجمله‌ای بودن توزیع متغیر وابسته از روش حداکثر درست‌نمایی برآورد می‌شوند. باتوجه به اینکه در این تحقیق متغیر وابسته(وضعیت پرونده خسارت) یک متغیر دو سطحی است، از رگرسیون لجستیک برای تشخیص جعلی یا غیرجعلی بودن پرونده‌های خسارت استفاده شده است. با استفاده از رگرسیون لجستیک پیش‌رو متغیرهایی که نقش مهم‌تری در تعیین وضعیت پرونده خسارت داشتهآند، شناسایی و وارد مدل شده‌اند. در گام اول مبلغ کل پرونده خسارت و مقدار ثابت در مدل قرار گرفته‌اند. در گام های دوم و سوم به ترتیب متغیر های فاصله زمانی وقوع حادثه تا اعلام خسارت و نوع خسارت به مدل افزوده شده‌اند.

مشاهده‌شده		برآوردشده		
		وضعیت پرونده		درصد صحیح
		جعلی	غیرجعلی	
وضعیت پرونده	جعلی	۳۰	۶	۸۸/۳
	غیرجعلی	۴	۳۲	۸۸/۹
کل				۸۶/۱

شکل ۵٬۱: دقت مدل در شناسایی وضعیت پرونده‌های خسارت با استفاده از رگرسیون لجستیک

[۱۳] logistic regression

۳.۵ روش بیز ساده[۱۴]

بیز ساده، شکل بسیار مقدماتی از مدل احتمال بیزی است. احتمال رخداد هریک از نتایج نهایی، براساس احتمالات رخداد متغیرهای مستقل به شرط رخداد همان نتیجه به‌دست‌می‌آید. فرض ما بر این است که احتمال رخداد هریک از متغیرهای مستقل به شرط رخداد یک نتیجه نهایی خاص، مستقل از احتمال رخداد سایر متغیرهای مستقل به شرط رخداد همان نتیجه باشد. عملکرد بیز ساده دسته کننده[۱۵] بر فرضیات استقلال قوی استوار است. یعنی اینکه احتمال رخداد یک صفت روی احتمال سایر صفت‌ها بی‌تاثیر است. تئوری بیز امکان محاسبه احتمال پسین را برمبنای احتمالات پیشین فراهم می‌کند. در مدل احتمال بیز اگر h یک پیشامد و D مشاهدات باشد آنگاه خواهیم داشت:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \qquad (۵.۲)$$

که در آن P(h) احتمال رخداد h، P(D) احتمال رخداد D، P(D|h) احتمال رخداد D به شرط رخداد h و P(h|D) احتمال رخداد پیشامد h به شرط رخداد D است. در مواردی که مجموعه‌ای از پیشامدهای H وجود داشته باشد و بخواهیم محتمل ترین فرضیه را از میان آنان انتخاب کنیم، از فرضیه حداکثر احتمال[۱۶] استفاده می‌شود که رابطه آن به این شکل است:

$$\begin{aligned} h_{MAP} &= argmax P(h|D) \\ &= argmax \frac{P(D|h)P(h)}{P(D)} \\ &= argmax P(D|h)P(h) \end{aligned} \qquad (۵.۳)$$

۴.۵ درخت تصمیم[۱۷]

درخت تصمیم از ابزارهای داده‌کاوی است که در رده‌بندی داده‌های کیفی استفاده می‌شود. در درخت تصمیم، درخت کلی به‌وسیله خردکردن داده‌ها به گره‌هایی ساخته می‌شود که مقادیری از متغیرها را در خود جای می‌دهند. با ایجاد درخت تصمیم براساس داده‌های پیشین که رده آنها معلوم است، می‌توان داده‌های جدید را دسته بندی کرد. درخت تصمیم دارای قابلیت فهم بالا و سرعت مناسب در یادگیری الگو بوده و می‌توان از آن برای کشف تقلب در شرکت‌های بیمه استفاده کرد.

---

[۱۴] order simple

[۱۵] Classifier Bayes Naive

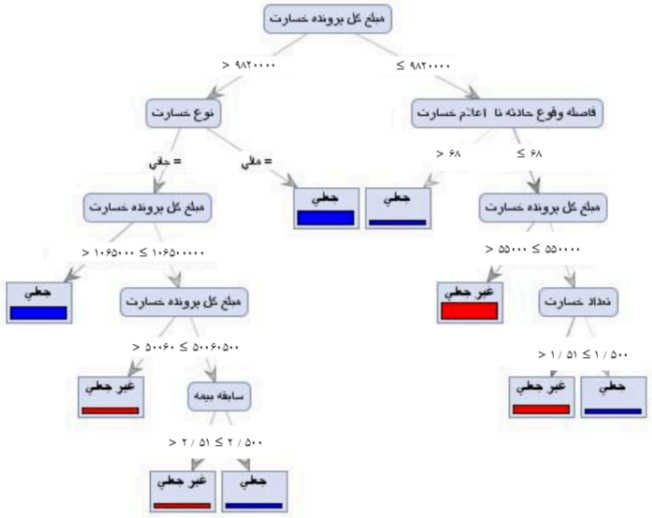
[۱۶] (MAP) hypothesis Posteriori A Maximum

[۱۷] tree decision

مشاهده شده		برآورد شده		
		وضعیت پرونده		درصد صحیح
		جعلی	غیرجعلی	
وضعیت پرونده	جعلی	۳۲	۴	۸۸/۸۹
	غیرجعلی	۳	۳۳	۹۱/۶۷
کل				۹۰/۲۸

شکل ۵٬۲: دقت مدل در شناسایی وضعیت پرونده‌های خسارت با استفاده از مدل بیز ساده

هدف از استفاده از درخت تصمیم در این تحقیق، طبقه بندی داده‌های خسارت جدید در بیمه اتومبیل است. معیارهای مختلفی برای تعیین صفتی که خردکردن داده‌ها باید براساس آن انجام شود، وجود دارد که از آن جمله می‌توان به معیارهای بهره اطلاعاتی [۱۸] ، نسبت بهره[۱۹] و شاخص جینی[۲۰] اشاره کرد.



شکل ۵٬۳: دقت مدل در شناسایی وضعیت پرونده‌های خسارت با استفاده از مدل بیز ساده

[۱۸] Gain Ĭnformation
[۱۹] Ratio Gain
[۲۰] Index Gini

در ادامه با اعمال این مدل بر روی داده‌های اولی، نتایج زیر جهت بررسی دقت مدل به دست آمده است:

مشاهده شده		برآوردشده		
		وضعیت پرونده		درصد صحیح
		جعلی	غیرجعلی	
وضعیت پرونده	جعلی	۳۵	۱	۹۷/۲
	غیرجعلی	۷	۲۹	۸۰/۶
کل				۸۸/۹

شکل ۵٬۴: دقت مدل در شناسایی وضعیت پرونده‌های خسارت با استفاده از مدل بیز ساده

۶ نتیجه گیری

در این مقاله سه روش داده‌کاوی رگرسیون لجستیک، بیز ساده و درخت تصمیم برای ساخت مدل‌هایی جهت شناسایی ادعاهای خسارت تقلبی در بیمه اتومبیل معرفی شدند. در ادامه این روش‌ها بر روی داده‌های واقعی آزمایش و کارایی هر روش سنجیده شد. روش بیز ساده با دقت ۹۰٬۲۸ درصد در شناسایی صحیح جعلی یا غیرجعلی بودن پرونده های خسارت بهترین کارایی را در مقایسه با دو روش درخت تصمیم با دقت کلی ۸۸٬۹ درصد و رگرسیون لجستیک با دقت کلی ۸۶٬۱ درصد داشت. البته باید به این نکته توجه داشت که در مدل بیز ساده برای تشخیص جعلی یا غیرجعلی بودن هر خسارت، شش متغیر و در مدل درخت تصمیم، پنج متغیر حضور دارند. این در حالی است که تصمیم گیری در مدل رگرسیون لجستیک برمبنای سه متغیری است که بیشترین همبستگی را با متغیر وابسته دارند.

۷ پیوست‌ها

۱.۷ پیوست۱. جدول متغیرهای مورد استفاده در مدل

نام اختصاری متغیر	نوع متغیر	شرح متغیر
Y	وابسته	وضعیت پرونده (جعلی یا غیرجعلی‌بودن یک پرونده)
X_1	مستقل	سابقه بیمه‌ای بیمه‌گذاران
X_2	مستقل	تعداد ادعاهای خسارت بیمه‌گذاران در طول دورهٔ سابقهٔ بیمه
X_3	مستقل	فاصله زمانی بین وقوع حادثه تا اعلام خسارت
X_4	مستقل	وضعیت کروکی خسارت رخ‌داده
X_5	مستقل	جانی یا مالی‌بودن خسارت
X_6	مستقل	مبلغ خسارت

۲.۷ پیوست۲. ضرایب متغیرهای مدل و مقادیر P مقدار متناظر

ضرایب هریک از متغیرهای واردشده در مدل رگرسیون لجستیک و همچنین مقدار ثابت مدل همراه با مقادیر p مقدار آن ها در جدول نشان داده شده است.

		B	S.E.	Sig.
گام ۱	X_6	۳/۴۵۴	۱/۱۱۷	۰/۰۰۲
	ثابت مدل	-۱/۲۸۴	۰/۳۹۱	۰/۰۰۱
گام ۲	X_3	۱/۴۴۴	۰/۶۹۹	۰/۰۳۹
	X_6	۳/۰۵۴	۱/۰۷۱	۰/۰۰۴
	ثابت مدل	-۱/۸۹۵	۰/۴۸۴	۰/۰۰۰
گام ۳	X۳	۱/۵۶۰	۰/۵۴۶	۰/۰۰۴
	X۵	۴/۶۲۶	۱/۶۷۴	۰/۰۰۶
	X۶	۶/۸۴۵	۱/۸۹۰	۰/۰۰۰
	ثابت مدل	-۶/۸۳۲	۱/۹۶۷	۰/۰۰۱

۸ منابع

۱. راه‌چمنی، ابوالقاسم ۱۳۸۵، 'تقلب و کلاهبرداری تهدید همیشگی صنعت بیمه'، *فصلنامه آسیا*، ش۳۸، صص۹-۱۶.

2. Artis, M, Ayuso, M & Guillen, M 2002, 'Detection of automobile insurance fraud with discrete choice models and misclassified claims', *Journal of Risk and Insurance,* pp. 325-40.

3. Belhadji, DB & Dionne, G 1997, 'development of an expert system for the automatic detection of automobile insurance fraud', *Risk Management Chair, HEC-Montreal.*

4. Bolton, RJ & Hand, DJ 2002, 'Statistical fraud detection: a review', *Statistical Science*, vol. 17, no. 3, pp. 235–55.

5. Brockett, PL, Xia, X & Derrig, RA 1998, 'Using kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud', *The J. of Risk and Insurance,* pp. 245-74.