



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 راه حل 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 بیمه گر و بیمه گذار بتوان راحت تر به حل این موارد پرداخت.

صنعت بیمه با عقد قراردادی بین طرفین باعث می شود تا بیمه گر با پرداخت مبلغی از پرداخت مبلغ بیشتری در هنگام وقوع حادثه پیشگیری کند. این مبلغ را بیمه گر از مجموع پول های بدست آمده از قراردادهای کلی پرداخت می کند.

با وجود این تلاش ها ممکن است در بسیاری از مواردی که بیمه ها ثبت می شوند در مراحل مختلف آن توسط افراد بیمه گر تخلفاتی صورت بگیرد که با تشخیص آن می توان 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 آن ها نیازمندیم می پردازیم و سعی می کنیم این سیستم بهترین عملکرد ممکن را داشته باشد.

فهرست مطالب

۲	۱ مقدمه
۳	۲ صنعت بیمه
۳	۱۰۲ انواع بیمه
۴	۲۰۲ تقلب در صنعت بیمه
۶	۳ مراحل شناسایی تقلب
۶	۱۰۳ شناسایی و غربال‌گری ^۱
۷	۲۰۳ تحقیق و بررسی ^۲
۷	۳۰۳ مذاکره با بیمه‌گذار یا طرح دعوی
۸	۴ فراگیری ماشین و داده‌کاوی
۹	۵ مطالعه تجربی
۹	۱۰۵ متغیرهای مورد استفاده در مدل
۱۰	۲۰۵ روش رگرسیون لجستیک ^۳

^۱ screening and identification
^۲ investigation
^۳ logistic regression

۱ مقدمه

امروزه فروش بیمه اهمیت خود را افزایش داده است و انواع توسعه ها در این زمینه در حال شکل گیری 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^۴ در این سیستم است به گونه ای که از متدوال ترین نوع این موارد، میتوان به گرفتن خسارت از بیمه، آتش سوزی عمدی، به صورت مکرر اشاره کرد

فروش و ثبت عملیات های بیمه ای به صورت داده منجر به دسترسی داشتن به اطلاعات عملیات های بیمه ای یک کاربر میشود. میتوان با بررسی داده های 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 الگوریتم های خاص خود و داده کاوی نیازمند است.

در ادامه سعی می شود تا الگوریتم های لازم شناسایی و استفاده شوند و بتوان از آن ها در سیستم های تشخیص تقلب استفاده کرد.

fraud^۴

۲ صنعت بیمه

بیمه^۵ 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 (بیمه‌گر) تعهد می‌کند در ازای پرداخت وجه یا وجوهی از طرف دیگر (بیمه‌گذار) در صورت وقوع یا بروز حادثه خسارت وارده بر او را جبران نموده یا وجه معینی را بپردازد. متعهد را بیمه‌گر، طرف تعهد را بیمه‌گذار و وجهی را که بیمه‌گذار 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 بیمه‌های 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

^۵Insurance

- 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 :

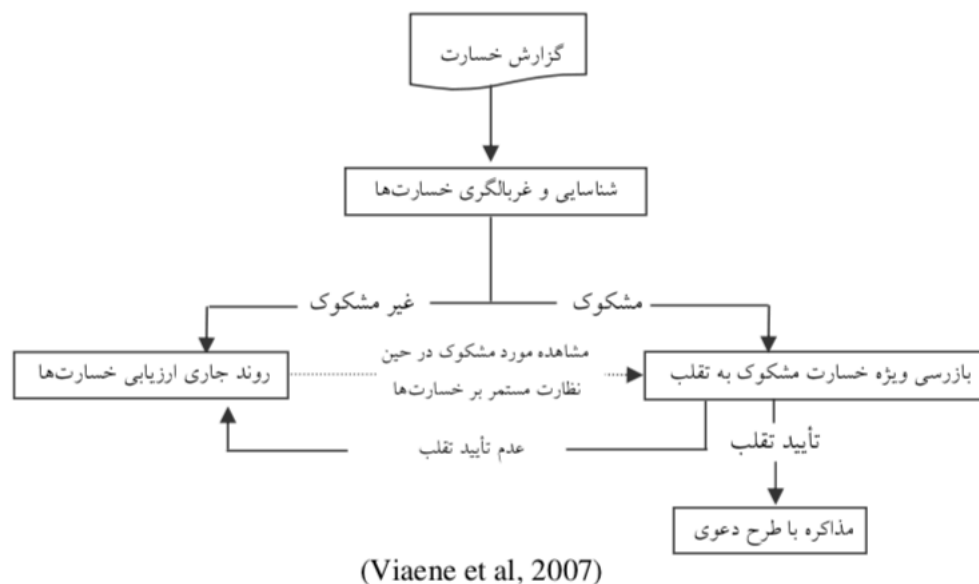
- اگرچه بیشتر پرونده‌ها و دعاوی بیمه‌ای درست و صحیح است، تقریباً نیمی از پرسش‌شوندگان احتمال تقلبی بودن یک ادعا را رد نکرده‌اند.
- احتمال وقوع تقلب بیمه‌ای بیشتر از سایر سوء استفاده‌هاست.
- در میان افراد شرکت‌کننده در تحقیق، در خصوص درست یا نادرست بودن اقداماتی مانند خریدن مال مسروقه یا رانندگی در حالت 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 به عنوان مصادیق تقلب در نظر می‌گیرند. بعضی از کلاهبرداری‌ها در صنعت بیمه کاملاً 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 حق بیمه‌ای که برای سالیان متمادی به شرکت بیمه پرداخت نموده است درصدد بهره‌برداری از فرصت برآید و با تجمع کلیه زیان‌های پیشین با خسارت فعلی سعی در کسب موقعیت مالی بهتر کند. مورد دوم که ممکن است منجر به 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 با 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 منجر به عدم پرداخت خسارت می‌گردد.



شکل ۳-۱: مراحل تشخیص تقلب

۱۰.۳ شناسایی و غربالگری^۶

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 سیستم‌های هوشمند، بررسی خسارت‌ها تنها براسا

^۶ screening and identification

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

- گردآوری داده‌ها در مرحله صدور بیمه‌نامه از طریق فرم‌هایی که توسط بیمه‌گذاران پر می‌شوند، اطلاعاتی در مورد بیمه‌گذاران و اتومبیل بیمه شده از قبیل تاریخ تولد، نشانی، نوع اتومبیل، تاریخ اخذ گواهینامه رانندگی، نوع کاربری اتومبیل و ... که غالباً 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 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^۷

تشخیص تقلبی بودن ادعا برعهده ارزیاب خسارت است که وی براساس تجربه، توانایی و خلاقیت خود این فرآیند را انجام می‌دهد. براساس تحقیقات تنیس و سالساس^۸ 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 این حال در برخی از موارد نیاز به دادگاه خواهد بود. اما دعوی قضایی و بازرسی‌های ویژه معمولاً مستلزم صرف هزینه و زمان زیادی است. معمولاً شرکت‌های بیمه به دلیل تأثیری 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.

^۷ investigation
^۸ Salsas. & Tennyson ۲۰۰۲

۴ فراگیری ماشین و داده‌کاوی

با توجه به فراوانی اطلاعات در امروزه و در عصر حاضر مدیریت این داده‌های 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 در نهایت تفسیر نتایج آنها نیازمند تحلیل‌های انسانی است. روش‌های نوین تحلیل داده باید به دانش لازم و قابلیت تصمیم‌گیری براساس داده‌ها تجهیز شوند. جهت دستیابی به این هدف* محققین به ارائه ایده‌های جدیدی از فراگیری ماشین^۹ پرداخته‌اند. 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^{۱۰} نام گرفته به وجود آمده است.

داده‌کاوی 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 داده‌ها باید همواره در قالب کمیت‌های معتبر ارائه شوند. استفاده از مدل‌های 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.

Learning Machine^۹
Mining Data^{۱۰}

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 در نظر گرفته می‌شود. مقدار ۱ برای متغیر وابسته به معنای جعلی بودن پرونده خسارت و مقدار ۰ به معنای غیرجعلی بودن آن پرونده است. در این مطالعه، فرآیند شناسایی تقلب با استفاده 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 بهتر شدن 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 به منظور مقاصد سودجویانه استفاده کرده باشد. از این رو 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 این دلیل انتخاب شده است که با حضور پلیس در صحنه حادثه، شانس تقلب از قبیل صحنه‌سازی کاهش می‌یابد.

۲.۵ روش رگرسیون لجستیک^{۱۱}

زمانی که متغیر وابسته، متغیری کیفی با دو سطح باشد، مدل‌های رگرسیون معمولی قابل استفاده نیستند. در این گونه موارد معمولاً از رگرسیون لجستیک استفاده می‌شود. مدل رگرسیون لجستیک به این صورت تعریف می‌شود:

$$\text{logit}\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_d X_d \quad (۵.۱)$$

که در این مدل X متغیرهای 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. با توجه 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 داشته‌اند، شناسایی و وارد مدل شده‌اند. در گام اول مبلغ کل پرونده خسارت و مقدار ثابت در مدل قرار گرفته‌اند. در گام‌های دوم و سوم به ترتیب متغیرهای فاصله زمانی وقوع حادثه تا اعلام خسارت و نوع خسارت به مدل افزوده شده‌اند.

^{۱۱} logistic regression