نشریـه علمـی حسابداری مدیریت
دوره ۱۵ / شماره ۲ (پیاپی ۵۳) / تابستان ۱۴۰۱
صفحه ۱۴۱ تا ۱۵۳

کاربرد یادگیری بدون نظارت در کشف تقلبات بیمه اتومبیل (الگوریتم جنگل ایزوله)

فربد خانی‌زاده [۱]
فرزان خامسیان [۲]
مریم اثنی‌عشری ✉ [۳]

تاریخ دریافت: ۱۴۰۰/۱۱/۱۴ تاریخ پذیرش: ۱۴۰۱/۰۵/۰۱

چکیده

استراتژی شرکت‌های بیمه در مقابله با تخلفات و تقلبات، بسیار حائز اهمیت می‌باشد. نداشتن چنین برنامه‌ای برای جلوگیری از تقلبات بیمه‌ای و پرداخت سریع خسارت بیمه‌گذاران، ممکن است در کوتاه مدت موجب رضایت مشتریان و افزایش پورتفوی شرکت‌ها گردد؛ اما در بلندمدت عواقب ناگواری را برای صنعت بیمه به همراه دارد. به‌عبارت دیگر، هزینه پرونده‌های تقلب خسارت در طول زمان به‌صورت افزایش حق بیمه و غیرمستقیم به بیمه‌گذاران منتقل می‌گردد. هدف از این مطالعه، ارائه مکانیزمی به شرکت‌های بیمه جهت کشف تقلب است. دستیابی به این هدف از طریق الگوریتم بدون نظارت و جهت کشف ناهنجاری آشکار در مجموعه داده می‌باشد. استفاده از الگوریتم مزبور به علت تجمیعی بودن آن باعث افزایش دقت در تشخیص پرونده‌های مشکوک به تقلب و کاهش موارد مثبت کاذب می‌گردد. بر اساس نتایج مقاله خسارت وارده به راننده مقصر، نوع و کاربری خودرو، جنسیت زیان‌دیده از مهمترین شاخص‌ها در کشف پرونده‌های مشکوک به تقلب هستند.

واژه‌های کلیدی: الگوریتم بدون نظارت، جنگل ایزوله، کشف تقلب، بیمه خودرو

[۱] استادیار، گروه پژوهشی بیمه‌های اموال و مسئولیت، پژوهشکده بیمه، تهران، ایران. khanizadeh@irc.ac.ir
[۲] استادیار، گروه پژوهشی عمومی بیمه، پژوهشکده بیمه، تهران، ایران. khamesian@irc.ac.ir
۳ استادیار، گروه پژوهشی بیمه‌های اموال و مسئولیت، پژوهشکده بیمه، تهران، ایران. (نویسنده مسئول): esnaashari@irc.ac.ir

۱- مقدمه

تقلب بیمه‌ای عملی است که با هدف کلاهبرداری از بیمه‌گر، برای کسب منافع مالی انجام می‌گیرد. تقلب بیمه‌ای از زمان شکل‌گیری بیمه به‌عنوان بنگاه‌های تجاری وجود داشته و سالانه میلیاردها دلار، هزینه را به شرکت‌های بیمه تحمیل نموده است. تقلبات بیمه‌ای انواع گوناگونی دارد و در تمام حوزه‌های بیمه‌ای رخ می‌دهد و طیف گسترده‌ای از ادعاهای اغراق‌آمیز تا تصادف‌ها و خسارت‌های تعمدی را در برمی‌گیرد. این تقلب‌ها سبب افزایش هزینه‌های بیمه‌گر و در پی آن، افزایش مبلغ حق‌بیمه می‌شود؛ از این‌رو تقلبات بیمه‌ای، به ضررکلیه بیمه‌گذاران خواهد بود. امروزه با وجود پیشرفت‌های فراوان در شناسایی این تقلب‌ها، هزینه‌های ایجادشده برای شرکت‌های بیمه‌ای در اثر این کلاهبرداری‌ها در حال افزایش است. در گذشته، پیامدهای مالی تقلبات بیمه‌ای در حدی نبود که ارزش بررسی و تلاش برای یافتن راه‌حل‌های ممکن را داشته باشد، اما به تازگی این وضع دگرگون شده و شرکت‌های بیمه، اهمیت مطالعه و ضرورت بررسی بیشتر در خصوص کشف تقلبات بیمه‌ای و عوامل موثر بر آنها در حوزه‌ها و رشته‌های بیمه‌ای مختلف را درک کرده‌اند. هزینه سالیانه صنعت بیمه آمریکا برای خسارت‌های تقلبی رشته اتومبیل در حدود ۲۹ میلیارد دلار می‌باشد و باید توجه داشت که به دلیل افزایش هزینه‌های شرکت، مبلغ این خسارت‌های تقلبی، در حق‌بیمه کلیه افراد دارای بیمه خودرو سرشکن می‌شود. سازمان ملی جرائم بیمه[1] آمریکا، تخمین می‌زند که برای جبران کلاهبرداری در بیمه، حدود ۲۰۰ تا ۳۰۰ دلار به حق‌بیمه هر فرد اضافه شود. در سال‌های اخیر تکنیک‌های داده‌کاوی و یادگیری ماشین، مانند شبکه‌های عصبی مصنوعی، منطق فازی و الگوریتم‌های ژنتیک، به دلیل توانمندی بالایی که در مدل کردن مسائل پیچیده دارند، به ابزار رایج در کشف تقلب تبدیل شده‌اند.

با توجه به مطالب بیان شده، هدف اصلی این مقاله، بررسی و کشف الگوهای رایج تقلب در رشته بیمه خودرو است و اهداف فرعی آن، به شرح زیر هستند.

- استخراج شاخص‌های موثر در وقوع تقلب؛
- شناسایی انواع تقلبات با توجه به پرونده‌های موجود؛
- تحلیل داده‌ها با استفاده از روش‌های مناسب و مختلف داده‌کاوی، استخراج مدل‌ها به فراخور داده‌ها؛
- استفاده از الگوریتم‌های یادگیری ماشین جهت افزایش دقت در کشف تقلب.

۲- پیشینه پژوهش

از دهه ۹۰ تا کنون تحقیقات بسیاری در زمینه کشف خسارت‌های تقلبی در رشته بیمه خودرو انجام شده است. برای مثال می‌توان به تحقیق (ویسبرگ و دریگ[2]، ۱۹۹۸) اشاره کرد. در این تحقیق نویسندگان به بررسی شاخص‌های تاثیرگذار در شناسایی پرونده‌های مشکوک پرداختند و کارایی روش‌ها و شیوه‌های ارزیابان خسارت در کشف چنین پرونده‌هایی را بررسی نموده‌اند. (بلهادجی و دیون[3]، ۱۹۹۷) همچنین تحقیقاتی را در این زمینه با استفاده داده‌های خسارت بیمه خودرو کشور کانادا انجام داده‌اند. (کیومینز و تنیسن[4]، ۱۹۹۲) در تحقیقی خصوصیات وضعیت تعادلی بازار بیمه را که در آن بیمه‌گذاران فرصت‌طلب ممکن است ادعاهای کلاهبرداری ارائه دهند را بررسی نموده‌اند. محققان دیگری مانند (دریگ و استاسزیوسکی[5]، ۱۹۹۵؛ ویزبرگ و دریگ[6]، ۱۹۹۸؛ براکت و همکارانش[7]، ۱۹۹۸) در مقالات خود تکنیک‌هایی برای شناسایی خسارت‌های تقلبی و دسته‌بندی کلاهبرداری‌ها ارائه دادند. (آرتیس و همکاران[8]، ۲۰۰۲) عملکرد مدل‌های انتخاب باینری را برای کشف تقلب در بازار بیمه خودرو اسپانیا برای سال‌های ۱۹۹۶ تا ۱۹۹۳ تجزیه و تحلیل کردند. آنها روشی برای اصلاح طبقه‌بندی نوع خسارت معرفی کردند. در مطالعه دیگری که توسط (اسکویی و همکاران ۱۳۹۹) بر روی داده‌های صنعت بیمه ایران انجام گرفت، از طریق الگوریتم‌های بانظارت یادگیری ماشین تاثیر ویژگی‌های خودرو در پیش‌بینی ریسک خسارت مالی در رشته بیمه شخص ثالث بررسی گردید که این موضوع در تحقیق دیگری که توسط همین تیم نویسندگان انجام گرفت اساس کشف تقلب به وسیله الگوریتم‌های با نظارت بود. استفاده از الگوریتم‌های با نظارت در کشف تقلب در سایر بازارهای مالی مانند بازار بورس و کلاهبرداری‌های مالیاتی نتایج خوبی داشته است. از آن جمله می‌توان به کاربرد یادگیری عمیق، شبکه‌های

عصبی و درخت تصمیم در تحقیقات (جوادیان کوتنائی و همکاران، ۱۳۹۹؛ تاراسی و همکاران، ۱۳۹۸) اشاره کرد.

روش‌های نوین داده کاوی امروزی جایگزین رویکردهای کلاسیک آماری و ریاضی شده و باعث افزایش شناسایی در کشف تقلب و رضایت بنگاه‌های اقتصادی در این زمینه گردیده است (ابودوکو و هار[۹]، ۲۰۱۹؛ سورینو و پنگ[۱۰]، ۲۰۲۱؛ سوبودهی و پانیگراهی[۱۱]، ۲۰۱۸؛ لیو[۱۲] و همکاران، ۲۰۲۰؛ گوپدارزی و جانات بابایی[۱۳]، ۲۰۱۷؛ تیواری[۱۴] و همکاران، ۲۰۲۱؛ عزیز[۱۵] و همکاران، ۲۰۲۲). 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(پول هول[۱۶] و یاروویی[۱۷]، ۲۰۱۹؛ گوپتا[۱۸] و همکاران، ۲۰۲۱؛ روخسار[۱۹] و همکاران، ۲۰۲۲).

۳- روش‌شناسی پژوهش

در این مقاله از روش یادگیری بدون نظارت (هاستی[۲۰] و همکاران، ۲۰۰۹؛ خامسیان و همکاران، ۲۰۲۱) برای تحلیل داده‌ها استفاده شده است. شایان ذکر است الگوریتم‌های استفاده شده، در فضای زبان برنامه‌نویسی پایتون کدنویسی و اجرا شده‌اند.

در الگوریتم‌های بدون نظارت، متغیر هدف نداریم و خروجی الگوریتم بر اساس الگوی درون داده‌ها مشخص می‌شود. در این نوع الگوریتم با مفهوم ناهنجاری[۲۱] روبرو می‌شویم. ناهنجاری‌ها الگوهایی هستند که دارای مشخصات و ویژگی‌های متفاوتی به نسبت نمونه‌های عادی هستند (چندولا[۲۲] و همکاران، ۲۰۰۹؛ روف[۲۳] و همکاران، ۲۰۱۹؛ پانگ[۲۴] و همکاران، ۲۰۲۱). تشخیص ناهنجاری‌ها از اهمیت قابل توجهی برخوردار است و غالباً اطلاعات کاربردی و حیاتی را در حوزه‌های مختلف فراهم می‌کند. بهترین مثال برای این نوع از الگوریتم‌ها، خوشه‌بندی یک جمعیت با داشتن اطلاعات شخصی و خریدهای مشتریان می‌باشد که به صورت خودکار آنها را به گروه‌های همسان و هم‌ارز تقسیم کنیم. در این دسته از یادگیری، تنها ورودی (x) را داریم و خروجی از پیش تعیین شده نیست. در واقع اینجا ناظری وجود ندارد تا به الگوریتم در یادگیری کمک کند و مدل مجبور است خودش ساختار مخفی داده بدون برچسب را پیدا کند (لیزون[۲۵]، ۲۰۱۵).

ما در این پژوهش از میان الگوریتم‌های بدون نظارت، الگوریتم جنگل ایزوله را که در ادامه به‌صورت مختصر معرفی می‌شود، جهت کشف تقلبات مشکوک استفاده می‌کنیم. این الگوریتم نقاط قوت زیادی دارد؛ از جمله مهمترین آنها می‌توان از تجمیعی بودن آن نام برد. تجمیعی بودن الگوریتم سبب افزایش دقت در تشخیص پرونده‌های مشکوک به تقلب و کاهش موارد مثبت کاذب می‌گردد. ویژگی دیگر الگوریتم جنگل ایزوله قابلیت و توانایی زیاد آن در تشخیص ناهنجاری‌ها است.

۳-۱- الگوریتم جنگل ایزوله

اکثر رویکردهای موجود برای تشخیص ناهنجاری، سبدی از موارد عادی را ایجاد می‌کنند، سپس مواردی را که مطابق با مشخصات عادی نیستند، شناسایی می‌کنند. به عنوان مثال می‌توان به شیوه‌های آماری، طبقه‌بندی و خوشه‌ای به ترتیب در منابع (ابی و همکاران[۲۶]، ۲۰۰۶؛ وانگ[۲۷] و همکاران، ۲۰۱۹؛ اسمیتی[۲۸]، ۲۰۲۰؛ الگوشیری[۲۹] و همکاران، ۲۰۲۰) اشاره کرد.

در واقع در مدل‌های اشاره شده ابتدا رفتارهای نرمال، بررسی و شناسایی می‌شود و پس از آن، مواردی که از رفتار طبیعی انحراف و فاصله دارند به عنوان رفتار ناهنجار انتخاب می‌شوند. این رویکرد شاید برای مجموعه داده‌هایی با نمونه‌های محدود مشکلی ایجاد نکند؛ اما برای مجموعه داده‌های بزرگ، هم از لحاظ اختصاص حافظه و هم زمان محاسبه، باعث ایجاد یک چالش بزرگ می‌شود. در رویکرد جنگل ایزوله این مشکل از طریق ایزوله کردن نقاط نمونه، قابل حل می‌باشد. به طور کلی الگوریتم‌های جنگل به علت آنکه در زیرمجموعه الگوریتم‌های تجمعی[۳۰] قرار دارند به علت کارایی بالاترشان نسبت به الگوریتم‌های غیر تجمعی، از محبوبیت و کاربرد بیشتری برخوردار هستند. بایاس و واریانس کم، اگرچه اغلب در جهت‌های متضاد حرکت می‌کنند، دو ویژگی اساسی مورد انتظار برای ارزیابی یک مدل هستند. اساسا، برای حل یک مسئله، به

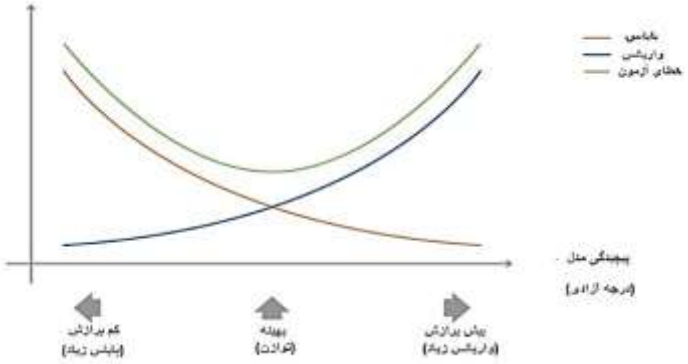
دنبال مدلی هستیم که دارای درجات آزادی کافی برای کشف الگوهای پیچیده در مجموعه داده‌ها باشد. از طرف دیگر برای اجتناب از ظهور واریانس زیاد، درجات آزادی نباید زیاد بزرگ باشند. این موضوع همان رابطه‌ی شناخته شده توازن یایاس-واریانس[۳۱] می‌باشد.

در تئوری یادگیری تجمعی، یادگیرنده‌های ضعیف[۳۲] (یا مدل‌های پایه) را مدل‌هایی می‌نامیم که از ترکیب آنها به عنوان بلوک‌های اصلی در طراحی مدل‌های پیچیده تر استفاده می‌شود. اغلب اوقات، این مدل‌های پایه به تنهایی عملکرد چندان مناسبی از خود نشان نمی‌دهند. دلیل این امر یا بایاس یالا (مدل‌های با درجه آزادی کم) یا واریانس بیش از حد (مدل‌های با درجه آزادی بالا) می‌باشد.

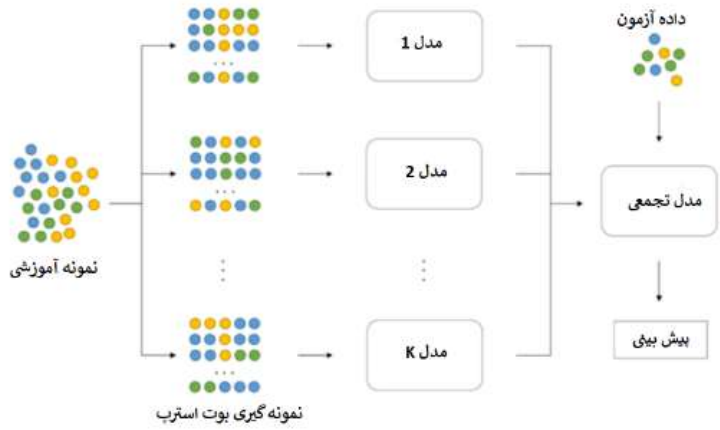
در روش‌های تجمع سعی بر آن است که بایاس و/یا واریانس یادگیرنده‌های ضعیف را با ترکیب تعدادی از آنها کاهش دهیم تا یک یادگیرنده قوی (یا مدل تجمعی) ایجاد کنیم که به عملکرد بهتری دست یابد.

الگوریتم‌های تجمعی به دو گروه اصلی بگینگ[۳۳] و بوستینگ[۳۴] تقسیم می‌شوند. با توجه به آنکه الگوریتم جنگل ایزوله در دسته الگوریتم‌های بگینگ قرار می‌گیرد به توصیف بیشتر این گروه می‌پردازیم.

الگوریتم بگینگ مخفف بوت استرپ اگریگیتینگ[۳۵] می‌باشد. تعریف بوت استرپ در علم آمار، نمونه گیری با جایگذاری از یک نمونه اصلی به دفعات زیاد است. در حقیقت در این روش از یک نمونه ثابت با حجم محدود به دفعات زیاد نمونه گیری مجدد البته با جایگذاری انجام داده می‌شود تا در نهایت بتوان با استفاده از نتایج کلیه‌ی دفعات نمونه گیری، در مجموع به یک توزیع نمونه‌ای دست یافت. پس از انجام نمونه‌گیری‌های مختلف (به عنوان مثال K نمونه)، K مدل پایه مختلف بر روی هر یک از نمونه‌ها آموزش داده می‌شوند و در نهایت یک مدل نهایی را تشکیل می‌دهند.



شکل ۱- نمودار توازن بایاس-واریانس



شکل ۲- نمودار الگوریتم تجمیعی (بگینگ)

برای بیان ریاضی شکل ۲ و الگوریتم بگینگ، فرض کنید که K نمونه بوت استرپ با اندازه B ایجاد شده است:

$$\{z_1^1, z_2^1, \dots, z_B^1\}, \{z_1^2, z_2^2, \dots, z_B^2\}, \dots, \{z_1^K, z_2^K, \dots, z_B^K\}$$

در مجموعه‌های بالا z_b^k بیانکر b امین مشاهده از k امین نمونه بوت استرپ می‌باشد. سپس K یادگیرنده‌ی ضعیف $w_1, w_2, \dots, w_k$ بر روی هر یک از مجموعه نمونه‌های داده برازش می‌شوند و در نوعی فرآیند میانگین گیری یک مدل تجمعی با واریانس کمتر به دست می‌آید:

$$s_K(.) = \frac{1}{K}\sum_{k=1}^{K} w_k$$

$$s_K(.) = arg\max_j [card(k|w_k = j)]$$

روابط بالا به ترتیب روش‌های رسیدن به مدل تجمعی در مسائل رگرسیون و طبقه بندی می‌باشد.

در الگوریتم جنگل ایزوله که یک مدل تجمعی است بر اساس مسیری که منجر به ایزوله شدن یک مشاهده از نمونه می‌شود؛ نقطه را به عنوان ناهنجاری یا یک رفتار عادی در نظر می‌گیرد:

همانطور که در شکل شماره ۳ قابل مشاهده است دو نقطه قرمز و آبی از دو مسیر متفاوت ایزوله شده‌اند و بر اساس الگوریتم جنگل ایزوله، نقاط ناهنجار به ریشه نزدیک‌تر هستند. بر همین اساس الگوریتم مزبور نقطه قرمز را به عنوان یک نمونه ناهنجار در نظر می‌گیرد.

از لحاظ ریاضی به هر یک از مشاهدات مجموعه داده‌ها یک امتیاز ناهنجاری اختصاص داده که از رابطه زیر محاسبه می‌شود:

$$S(x,n) = 2^{-\frac{E(h(x))}{c(n)}}.$$

n، تعداد نمونه و $c(n)$، میانگین طول مسیرهای ناموفق در درخت

جستجوی دودویی بوده و از رابطه زیر محاسبه می‌شود.

$$c(n) = 2H(n-1) - 2\left(\frac{n-1}{n}\right)$$

در معادله بالا $H(i)$ عدد هارمونیک بوده که می‌توان آن را با ثابت اویلر به صورت زیرتخمین زد.

$$ln(i) + 0.5772156649$$

و $h(x)$ طول مسیر منتهی به نقطه x می‌باشد. از روابط بالا حالت‌های زیر منتج می‌شود.

- وقتی $E(h(x)) \to c(n)$ آنگاه $s \to 0.5$
- وقتی $E(h(x)) \to 0$ آنگاه $s \to 1$
- وقتی $E(h(x)) \to n-1$ آنگاه $s \to 0$

تابع s یک تابع یکنوا نسبت به $h(x)$ و ارتباط بین آنها در نمودار شماره ۲ مشخص می‌باشد. با استفاده از امتیاز ناهنجاری، می‌توانیم ارزیابی‌های زیر را انجام دهیم.

- اگر نمونه‌ها خیلی نزدیک به ۱ باشند، قطعاً ناهنجاری هستند.
- اگر نمونه‌ها خیلی کوچکتر از ۰/۵ باشند، کاملاً امن هستند و می‌توان آنها را به عنوان موارد عادی در نظر گرفت.
- اگر برای تمام موارد ۰/۵ ≈ s، آنگاه ناهنجاری مشخصی وجود ندارد.

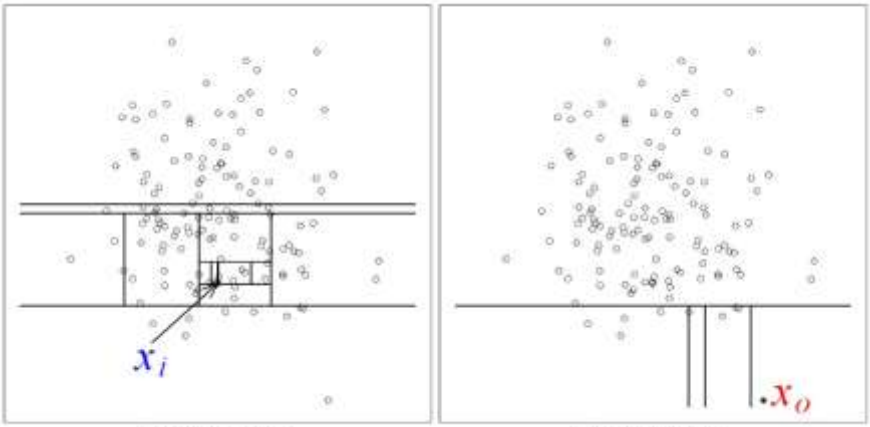
برای آگاهی و مطالعه در ارتباط با جزئیات بیشتر و دقیق‌تر از الگوریتم جنگل ایزوله می‌توان به مقاله (لیو وهمکاران[۳۶]، ۲۰۰۸) مراجعه کرد.

پیش از آنکه به استفاده از الگوریتم معرفی شده بپردازیم، اطلاعاتی در ارتباط با ساختار داده‌ها ارائه می‌گردد.

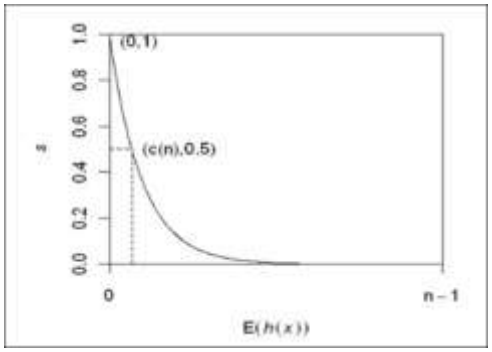
۳-۲- داده‌های پژوهش

جامعه آماری تحقیق شامل ۵۰ هزار نمونه از داده‌های خسارت مربوط به رشته بیمه شخص ثالث می‌باشد. همچنین مجموعه داده مذکور شامل ۱۵ متغیر مستقل نهایی می‌باشد. توضیحات و دسته‌بندی متغیرها در جدول شماره ۱ ارائه گردیده است.

قبل از ورود متغیرهای مورد نظر به مدل، نیاز است که بر روی داده‌ها پیش‌پردازش انجام پذیرد. پیش‌پردازش داده‌ها از گام‌های مهم فرایند داده‌کاوی است که میزان دقت نتایج به‌دست آمده، تا حد زیادی به اجرای درست آن بستگی دارد. در همین راستا برای پیش‌پردازش داده‌ها، اقدامات به شرح زیر انجام گردید.



شکل۳- افرازبندی الگوریتم جنگل ایزوله



نمودار ۲- رابطه بین $\boldsymbol{h(x)}$ و s

جدول ۱: متغیرهای مستقل در مجموعه داده

نام متغیر	کد انگلیسی متغیر
استان وقوع حادثه	CtyNam
محدوده وقوع حادثه	InCty
نوع وسیله نقلیه	MapCarTypCod
سن مقصر حادثه	Age
سن زیان‌دیده	AgeLoser
نوع گواهی‌نامه	IsLcnsFit
نوع کاربری	UsgCod
جنسیت زیان‌دیده	IsMale
جنسیت راننده مقصر حادثه	CusMaleCod
فاصله حادثه تا اعلام خسارت	Days
ساعت وقوع حادثه	Hour
نوع گروه‌بندی وسیله نقلیه	CarGrpCod
نوع خسارت زیان‌دیده	ThrLosTyp
کد نوع حادثه	AcdTypCod
کد استان	CtyNam

- **استخراج و تفکیک اطلاعات**

برای برخی از متغیرها، تنها بخشی از اطلاعات وارد شده مورد استفاده بود و نیاز بود این بخش از سایر اطلاعات به طور جداگانه استخراج شود.

- **ادغام متغیرها**

برخی از متغیرها به تنهایی حاوی اطلاعات مفیدی نبوده و از ترکیب و ادغام آن با متغیر دیگر، اطلاعات قابل تحلیل به‌دست می‌آمد.

- **حذف داده‌های دارای نویز**

در برخی از متغیرها، یک داده به دو دسته متفاوت تعلق داشت؛ به‌عنوان مثال در برخی موارد خودرو وانت یک بار به عنوان سواری و یک بار به عنوان بارکش ثبت شده بود که تا حد امکان، اینگونه اطلاعات دارای نویز از داده‌ها حذف گردید.

- **تبدیل کلیه مقادیر متغیرها به کدهای عددی**

از بین متغیرهای مستقل سن مقصر حادثه، سن زیان‌دیده، ساعت وقوع حادثه، فاصله زمانی وقوع حادثه تا اعلام خسارت، متغیر عددی و سایر متغیرها، متغیر اسمی می‌باشند. در همین راستا برای متغیرهای اسمی، کدگذاری انجام گردید.

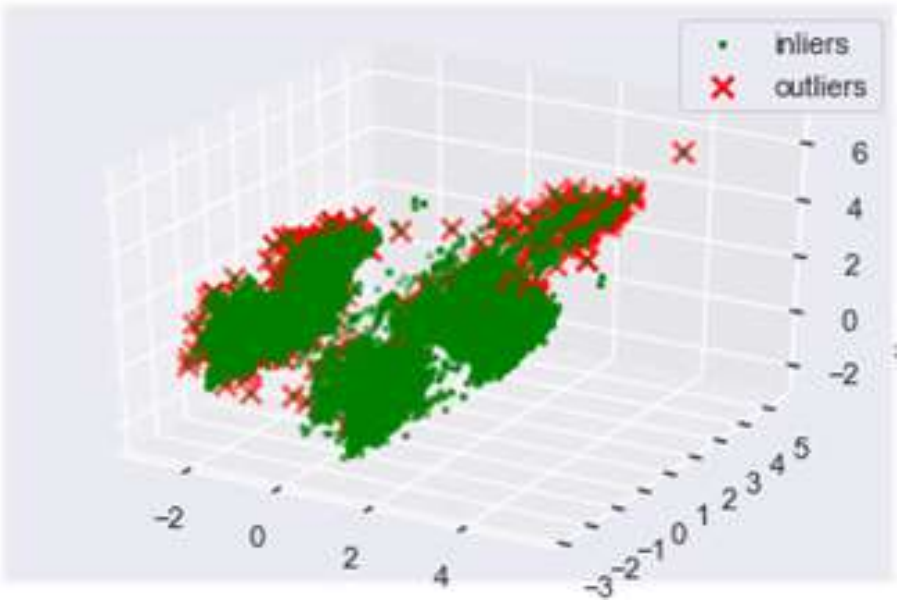
- **گروه‌بندی متغیرهای مستقل عددی**

برای یکسان‌سازی متغیرها و استفاده بهینه از الگوریتم‌های یادگیری ماشین، متغیرهای عددی به گروه و دسته‌های مختلف تقسیم شدند.

۴- یافته‌های پژوهش

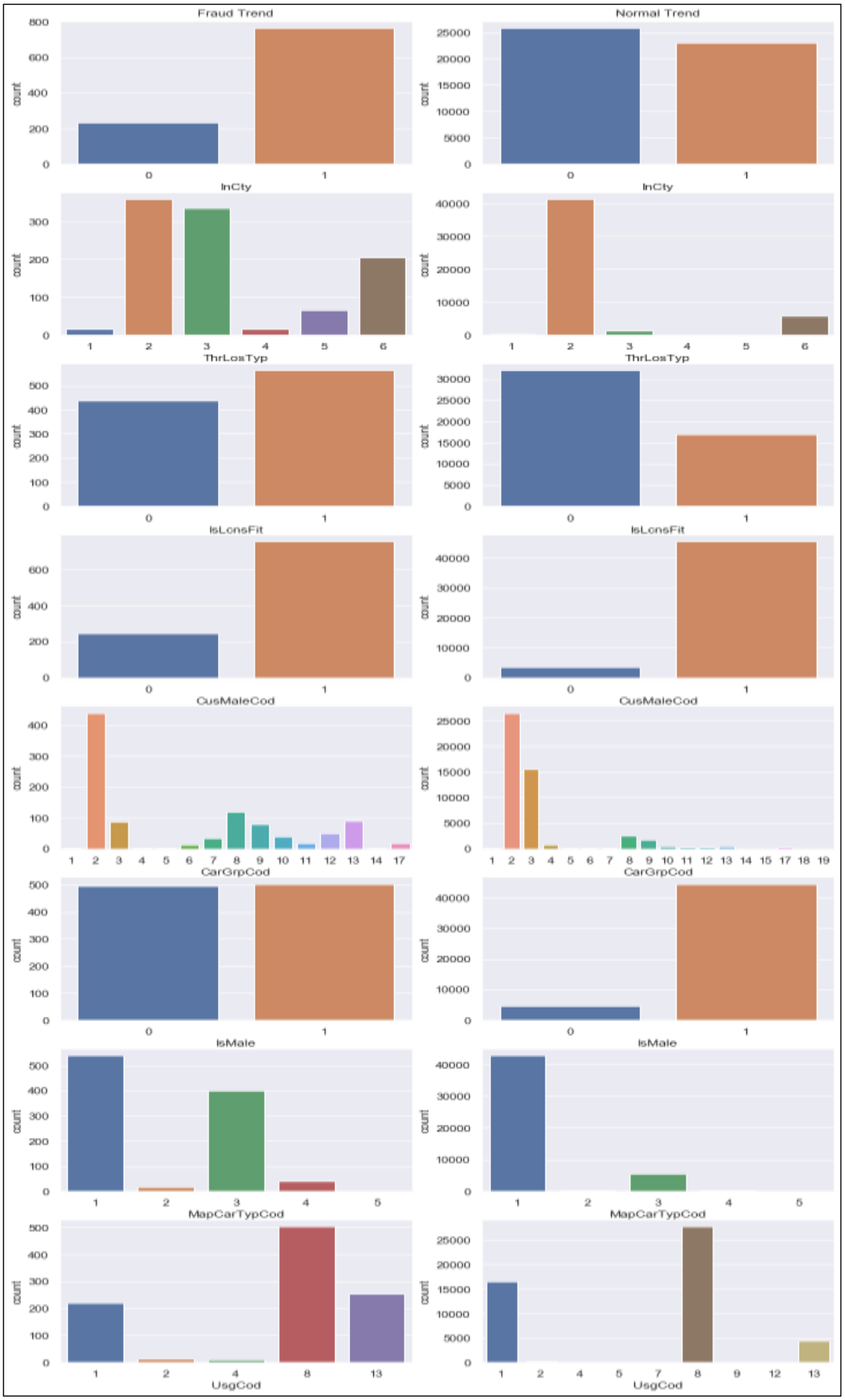
پس از وارد کردن مجموعه داده‌ها به عنوان ورودی الگوریتم، تعداد ۱۰۰۰ نمونه ناهنجاری شناسایی و به‌عنوان خروجی الگوریتم استخراج گردید. جهت نمایش ناهنجاری‌ها از روش تحلیل مولفه‌های اصلی استفاده شده که نتیجه آن در نمودار ۳ ارائه شده است.

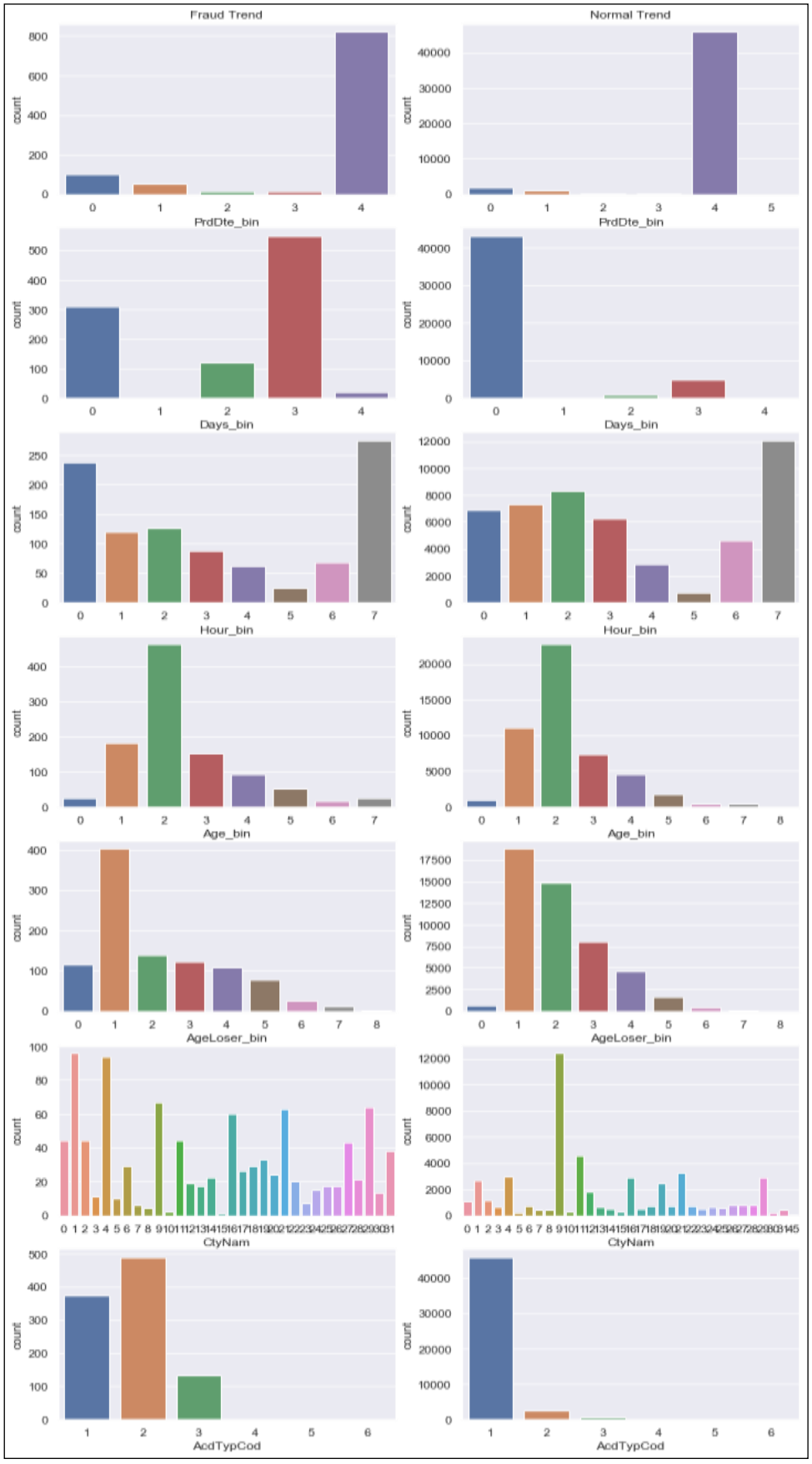
همانطور که در نمودار شماره ۳ مشاهده می‌شود نقاطی که دارای رفتار نامتعارف بوده و به عنوان نمونه‌های ناهنجار شناسایی شده‌اند با رنگ قرمز و نمونه‌های نرمال با رنگ سبز مشخص شده‌اند. در ادامه بر اساس نتایج الگوریتم، به مقایسه روند موجود در داده‌های مربوط به خسارت‌های مشکوک (نقاط ناهنجار) با خسارت‌های سالم می‌پردازیم:



نمودار ۳- نمایش نقاط ناهنجاری و نرمال

نمودار ۴- بررسی ساختار متغیرهای برای خسارت‌های مشکوک و سالم




Fraud Trend
Normal Trend
PrdDte_bin
Days_bin
Hour_bin
Age_bin
AgeLoser_bin
CtyNam
AcdTypCod
count

در نمودارهای بالا تفاوت بین پرونده‌های سالم و پرونده‌های تقلبی برای تمام متغیرهای موجود در تحلیل قابل مشاهده می‌باشد. نتایج بدست آمده از نمودارهای بالا در بخش بعد مورد اشاره قرار گرفته است.

۵- نتیجه‌گیری و بحث

در این مقاله از الگوریتم جنگل ایزوله جهت کشف تقلبات مشکوک استفاده گردید. این الگوریتم متخصصین را جهت کشف الگوهای ناهنجار در پرونده‌های خسارت هدایت می‌کند. براین اساس، نتایج مدل شامل ویژگی‌های موثر در شناسایی موارد نامتعارف استخراج گردیده که می‌تواند به عنوان یک راهنما در اختیار کارشناسان مربوطه قرار گیرد. نتایج به شرح ذیل می‌باشد:

تنها ۲,۵ درصد از پرونده‌های سالم خسارت شامل خسارت وارده به راننده مقصر بوده لیکن در موارد مشکوک به تقلب این عدد به حدود۳۵ درصد می‌رسد. در واقع فراوانی خسارت وارده به راننده مقصر در پرونده‌های مشکوک نسبت به پرونده‌های سالم بیش از ۳۰ درصد می‌باشد. این امر کارشناسان و متخصصان جرایم بیمه ای را نسبت به متغیر "نوع خسارت زیان‌دیده"، به عنوان یکی از تاثیر گذارترین ویژگی‌ها در تشخیص پرونده‌های مشکوک به تقلب، حساس تر می‌کند. در ارتباط با جنسیت فرد مقصر با توجه به آنکه تفاوت چندانی بین پرونده‌های سالم و مشکوک ملاحظه نمی شود در حال حاضر این متغیر تاثیر چندانی در تمایز پرونده‌های مشکوک از سالم ندارد. لیکن زمانی که جنسیت فرد زیاندیده، مدنظر باشد قضیه متفاوت است. در این حالت اختلاف به ۴۲ درصد می‌رسد. در واقع زنان زیاندیده در پرونده‌های مشکوک از فراوانی بالایی برخوردار هستند (۵۰ درصد) در حالیکه در پرونده‌های سالم این عدد تنها ۸ درصد می‌باشد. در واقع می‌توان نتیجه گرفت زنان در باندهای کلاهبرداری مرتبط با بیمه‌های اتومبیل نقش فعالی دارند. از این قشر در حوادث رانندگی به عنوان فرد مصدوم استفاده می‌شود. این امر باعث برانگیختگی بیشتر حس ترحم و دلسوزی شده و می‌تواند بررسی دقیق تر پرونده را تحت الشعاع قرار داده و در تسریع پرداخت خسارت تاثیرگذار باشد. به همین دلیل در پرونده هایی که فرد زیاندیده زن می‌باشد بررسی گسترده تر از طرف کارشناسان و ارزیابان شرکت‌های بیمه توصیه می‌شود. تفاوت فراوانی وسیله نقلیه بارکش در پرونده‌های مشکوک به تقلب نسبت به پرونده‌های سالم نیز به حدود ۳۰ درصد می‌رسد. به عبارتی در میان پرونده‌های سالم و مشکوک به تقلب به ترتیب ۱۱ و ۴۰ در صد از حوادث مربوط به وسیله نقلیه بارکش بوده است. در مجموعه داده‌های در اختیار، نوع وسیله نقلیه به پنج گروه سواری، موتورسیکلت، بارکش، اتوکار و سایر (کشاورزی، راه‌سازیئ ساختمان، حمل زباله و خیابان پاک‌کن‌ها) دسته بندی شده است. با توجه به فراوانی قابل ملاحظه‌ی خودروهای بارکش در پرونده‌های تقلب، از طرفی این متغیر می‌تواند به عنوان شاخصی در شناسایی پرونده‌های مشکوک مورد استفاده قرار گیرد. از طرف دیگر بررسی دلیل این فراوانی و استفاده کلاهبرداران از این نوع خودرو در حوادث رانندگی می‌تواند موضوع با اهمیتی برای تحقیقات آتی باشد.

شایان ذکر است این نتایج می‌تواند به عنوان راهنمایی جهت اهمیت دسترسی به متغیرهای مناسب جهت پیش بینی مدل در اختیار کارشناسان و متخصصین این حوزه قرار گیرد. بدیهی است یکی از نقاط قوت مدل‌های یادگیری ماشین پویا بودن آنها بوده و در همین راستا با گذر زمان مدل مزبور این قابلیت را دارد که متغیرهای جدیدی برای شناسایی پرونده‌های مشکوک ارئه دهد.

یاداشت‌ها

1-. National Insurance Crime Bureau
2-. Weisberg & Derrig
3-. Belhadji & Dionne
4-. Cummins & Tennyson
5-. Derrig & Ostazewski
6-. Weisberg & Derrig
7-. Brockett et al.
8-. Artis et al.
9-. Obodoekwe, & Haar
10-. Severino & Peng
11-. Subudhi & Panigrahi
12-. Liu
13-. Gopdarzi & Janatbabaei
14-. Tiwari
15-. Aziz

16-. Polhul
17-. Yarovyi
18-. Gupta
19-. Rukhsar
20-. Hastie
21-. Anomaly
22-. Chandola
23-. Ruff
24-. Pang
25-. Lison
26-. Abe et al
27-. Wang
28-. Smiti
29-. Alghushairy
30-. Ensemble Learning
31-. Bias-Variance Tradeoff
32-. Weak Learners
33-. Bagging
34-. Boosting
35-. Bootstrap Aggregating (Bagging)
36-. Liu et al

فهرست منابع

* اصغری اسکوئی، محمدرضا؛ خانی‌زاده، فربد و بهادر، آزاده، (۱۳۹۹)، کاربرد داده‌کاوی با استفاده از الگوریتم‌های یادگیری ماشین برای بررسی تاثیر ویژگی‌های خودرو در پیش‌بینی ریسک خسارت مالی در رشته بیمه شخص ثالث، فصلنامه علمی-پژوهشی پژوهشنامه بیمه، ۳۵(۱)، ۳۴-۶۵.
* جوادیان کوتنائی، اکبر؛ عباسعلی پورآقاجان سرحمامی، عباسعلی و حسینی شیروانی، میرسعید (۱۳۹۹)، ارائه مدل شناسایی تقلب مالیاتی بر مبنای ترکیب الگوریتم درخت تصمیم ID3 بهبود یافته و شبکه‌های عصبی پرسپترون چندلایه، نشریه علمی حسابداری مدیریت، ۴۶ (۱۳)، ۵۳-۷۰
* تاراسی، مجتبی؛ بنی طالبی دهکردی، بهاره و زمانی، بهزاد (۱۳۹۸)، پیش بینی گزارشگری مالی متقلبانه از طریق شبکه عصبی مصنوعی(ANN)، نشریه علمی حسابداری مدیریت، ۴۰ (۱۲)، ۶۳-۷۹
* Abe, N., Zadrozny, B., & Langford, J., (2006), Outlier detection by active learning, In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 504-509.
* Alghushairy, O., Alsini, R., Soule, T., & Ma, X. (2020). A review of local outlier factor algorithms for outlier detection in big data streams. Big Data and Cognitive Computing, 5(1), 1.
* Artís, M., Ayuso, M., & Guillén, M., (2002), Detection of automobile insurance fraud with discrete choice models and misclassified claims, Journal of Risk and Insurance, 69(3), 325-340.
* Aziz, R. M., Baluch, M. F., Patel, S., & Ganie, A. H. (2022). LGBM: a machine learning approach for Ethereum fraud detection. International Journal of Information Technology, 1-11.
* Belhadji, B., & Dionne, G., (1997), Development of an Expert System for Automatic Detection of Automobile Insurance Fraud (No. 97-06), Ecole des Hautes Etudes Commerciales de Montreal-Chaire de gestion des risques.
* Brockett, P. L., Xia, X., & Derrig, R. A., (1998), Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud, Journal of Risk and Insurance, 245-274.
* Chandola, V., Banerjee, A., & Kumar, V., (2009), Anomaly Detection: A Survey, ACM Computing Surveys, 41(3), 1-58.
* Cummins, J. D., & Tennyson, S., (1992), Controlling automobile insurance costs, Journal of Economic Perspectives, 6(2), 95-115.
* Derrig, R. A., & Ostaszewski, K. M., (1995), Fuzzy techniques of pattern recognition in risk and claim classification, Journal of Risk and Insurance, 447-482.
* Gopdarzi, A., & Janatbabaei, S., (2017), Evaluation of Three Data Mining Algorithms (Decision Tree, Naive Bayes, Logistic Regression) in Auto Insurance Fraud Detection, Insurance Research, 1(2), 61-80.
* Gupta, R. Y., Mudigonda, S. S., Baruah, P. K., & Kandala, P. K. (2021). Markov model with machine learning integration for fraud detection in health insurance. arXiv preprint arXiv:2102.10978.
* Hastie, T., Tibshirani, R., & Friedman, J., (2009), Unsupervised learning, In The elements of statistical learning (pp. 485-585), Springer, New York.
* Khanizadeh, F., Khamesian, F., & Bahiraie, A., (2021), Customer Segmentation for Life Insurance in Iran Using K-means Clustering, International Journal of Nonlinear Analysis and Applications, 12(Special Issue), 633-642.
* Lison, P., (2015), An introduction to machine learning, Language Technology Group (LTG), 1(35), 1-35.
* Liu, X., Yang, J. B., & Xu, D. L., (2020), Fraud detection in automobile insurance claims: A

statistical review, In Developments of Artificial Intelligence Technologies in Computation and Robotics: Proceedings of the 14th International FLINS Conference (FLINS 2020), 1003-1012.

* Obodoekwe, N., & Haar, D. T. V. D. (2019, February). A comparison of machine learning methods applicable to healthcare claims fraud detection. In International Conference on Information Technology & Systems (pp. 548-557). Springer, Cham.
* Pang, G., Shen, C., Cao, L., & Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. ACM Computing Surveys (CSUR), 54(2), 1-38.
* Polhul, T., & Yarovyi, A., (2019), Development of a method for fraud detection in heterogeneous data during installation of mobile applications, Eastern-European Journal of Enterprise Technologies, 1(2), 65-75.
* Ruff, L., Vandermeulen, R. A., Görnitz, N., Binder, A., Müller, E., Müller, K. R., & Kloft, M. (2019). Deep semi-supervised anomaly detection. arXiv preprint arXiv:1906.02694.
* Rukhsar, L., Bangyal, W. H., Nisar, K., & Nisar, S. (2022). Prediction of insurance fraud detection using machine learning algorithms. Mehran University Research Journal Of Engineering & Technology, 41(1), 33-40
* Severino, M. K., & Peng, Y. (2021). Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. Machine Learning with Applications, 5, 100074.
* Smiti, A. (2020). A critical overview of outlier detection methods. Computer Science Review, 38, 100306.
* Subudhi, S., & Panigrahi, S., (2018), Detection of automobile insurance fraud using feature selection and data mining techniques, International Journal of Rough Sets and Data Analysis (IJRSDA), 5(3), 1-20.
* Tiwari, P., Mehta, S., Sakhuja, N., Kumar, J., & Singh, A. K. (2021). Credit Card Fraud Detection using Machine Learning: A Study. arXiv preprint arXiv:2108.10005.
* Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in outlier detection techniques: A survey. Ieee Access, 7, 107964-108000.
* Weisberg, H. I., & Derrig, R. A., (1998), Quantitative methods for detecting fraudulent automobile bodily injury claims. Risques, 35(July–September), 75-99.

Journal of Management Accounting
Vol. 15, No.53, Summer 2022

Employing unsupervised learning to detect fraudulent claims in auto insurance (isolation forest)

Farbod Khanizadeh [1]
Farzan Khamesian [*2]
Maryam Esna-Ashari [*3]

Abstract

For insurance companies, fraud detection strategies are of significant importance. Lack of such a plan to prevent insurance fraud and making payments quickly to insured in order to compensate for losses will lead to customer satisfaction and increase companies' portfolio in short term. However in the long run, it will have dire consequences for the insurance industry. In other words, the cost of fraudulent claims would be transferred indirectly to insured in the form of a rise in premiums. The purpose of this study is to provide insurers with a mechanism to detect fraudulent claims. This goal is achieved through an unsupervised algorithm to detect anomalies in the data set. The use of this algorithm, as it is an ensemble learning, increases the accuracy in detecting suspicious cases and reduces false positives. According to the results, the damage to the culprit, the type and use of the vehicle, and the sex of the victim are among the most important indicators in the detection of fraudulent cases.

.

Keywords: Unsupervised learning, Isolation forest, Fraud detection, Auto insurance.

[1] Assistant Professor, Property and Casualty Insurance Research Group, Insurance Research Group, Tehran, Iran. khanizadeh@irc.ac.ir
[2] Assistant Professor, General Insurance Research Group, Insurance Research Group, Tehran, Iran. khamesian@irc.ac.ir
[3] Assistant Professor, Property and Casualty Insurance Research Group, Insurance Research Group, Tehran, Iran. (Corresponding author): esnaashari@irc.ac.ir.