فصلنامــه اقتصاد مالی
دورهٔ ۱۹ / شمارهٔ ۳ (پیاپی ۷۲)/ پائیز ۱٤۰٤
صفحه ۰۱ تا ۲۹

یک رویکرد مقایسه ای یادگیری ماشینی برای پیش‌بینی داده های ذخایر خسارت‌های واقع‌شده ولی گزارش‌نشده بیمه ای در حضور داده های سانسور شده و بریده شده

اکبر پیله ور سلطان احمدی[۱]
کیومرث شهبازی[۲]✉
حمزه دیدار[۳]

تاریخ دریافت: ۱٤۰٤/۰٤/۱۰ تاریخ پذیرش: ۱٤۰٤/۰٦/۱۰

چکیده

این مطالعه با هدف پیش‌بینی ذخایر خسارت‌های واقع‌شده ولی گزارش‌نشده، در رشته‌های مختلف بیمه‌ای، از مدل‌های یادگیری ماشین پیشرفته و تحلیل داده‌های سانسورشده و بریده‌شده استفاده کرده است. داده‌ها شامل اطلاعات تاریخ‌های وقوع و گزارش حادثه در پنج رشته بیمه‌ای مهم، شامل ثالث مالی، بدنه، ثالث جانی و حوادث راننده، آتش‌سوزی و مسئولیت بوده و روش‌ها شامل رگرسیون خطی چندگانه (MLR)، مدل خطی تعمیم‌یافته (GLM)، مدل افزایشی تعمیم‌یافته (GAM)، جنگل تصادفی (RF)، شبکه عصبی (MLP) و حافظه کوتاه‌مدت و بلندمدت (LSTM) در دوره زمانی ۱۴۰۰ تا ۱۴۰۱ در شرکت بیمه ایران می‌باشند. با سانسور کردن و برش داده‌ها در مقاطع مختلف، بر حسب روزهای تعطیل، نوروز، روزهای شلوغ سال و دوره‌های رونق ساخت و ساز، ویژگی‌های اثرگذار داده‌ها، براساس نوع رشته بیمه‌ای مدل‌سازی شده است. نتایج نشان داد که مدل‌های LSTM و RF در پیش‌بینی تاخیرها عملکرد بسیار بهتری نسبت به مدل‌های خطی داشتند؛ به‌طور خاص، مدل RF در رشته‌های بدنه و ثالث مالی با خطا به ترتیب ۱۰/۶۴ و ۱۱/۰۲ و مدل LSTM با خطا به‌ترتیب ۹/۸۳ و ۱۰/۷۲، دقت بالاتری نسبت به سایر مدل‌ها داشته‌اند. این مدل‌ها در شناسایی الگوهای پیچیده موجود در داده‌ها توانمند بوده و نشان دادند که با توجه به تأثیرگذاری عواملی مانند تعطیلات، آخر هفته‌ها و نوع ترکیب داده‌ها می‌توانند الگوهای پیچیده‌تری را در داده‌های بیمه‌ای شناسایی کنند. این نتایج تأکید دارد که مدل‌های LSTM و جنگل تصادفی به‌طور چشمگیری قابلیت بهبود دقت پیش‌بینی را دارا بوده و ابزار مناسبی برای ارزیابی ریسک و تخصیص بهینه ذخایر مالی در صنعت بیمه محسوب می‌شوند.

واژه‌های کلیدی: ذخایر خسارت‌های واقع‌شده ولی گزارش‌نشده، جنگل تصادفی، شبکه عصبی پرسپترون چندلایه، شبکه عصبی با حافظه طولانی کوتاه‌مدت

طبقه بندی JEL: G22, C45, C53, G17

۱ گروه علوم اقتصادی، دانشکده اقتصاد، دانشگاه ارومیه، ارومیه، ایران pilehvar.1394@gmail.com
۲ گروه علوم اقتصادی، دانشکده اقتصاد ، دانشگاه ارومیه، ارومیه، ایران. (نویسنده مسئول)، k.shahbazi@urmia.ac.ir
۳ گروه حسابداری ، دانشکده اقتصاد، دانشگاه ارومیه، ارومیه، ایران h.didar@urmia.ac.ir

نوع مقاله: علمی پژوهشی
Eci@iauctb.ac.ir

۱- مقدمه

پیش‌بینی خسارات واقع‌شده ولی گزارش‌نشده[1]، یکی از جنبه‌های حیاتی در مدیریت ریسک و تضمین پایداری مالی شرکت‌های بیمه است. خسارات واقع‌شده ولی گزارش‌نشده، به خسارت‌هایی اطلاق می‌شود که در طول دوره بیمه‌نامه رخ داده‌اند، اما هنوز به شرکت بیمه گزارش نشده‌اند. اهمیت تخمین خسارات واقع‌شده ولی گزارش‌نشده در توانایی شرکت‌ها برای حفظ پایداری مالی و کنترل خطر نقدینگی الزامی است (سوسا[2]، ۲۰۲۳، ۲). این اهمیت، بخصوص در رشته‌های بیمه‌ای است که تأخیر قابل‌توجهی بین وقوع حادثه و گزارش آن وجود دارد (بالونا و ریچمن[3]، ۲۰۲۱، ۲۴). ارزیابی صحیح میزان پرداخت و جلوگیری از مصرف غیرضروری سرمایه نیز اهمیت زیادی دارد. با افزایش نیاز به گزارش‌های مالی و نگرانی‌های مداوم از بابت کفایت مالی، اکچوئرها بیش از پیش با نیاز به ارائه تخمین‌های قابل اعتماد از هزینه‌ها و ذخایر خسارت مواجه شده‌اند. روش‌های متعددی شامل روش‌های قطعی یا تصادفی براساس داده‌های تجمعی خسارت‌ها که در مثلث‌های توسعه خسارت[4] سازماندهی شده‌اند، برای محاسبه ذخایر خسارت‌های باقی‌مانده وجود دارد. این روش‌ها برای مدیریت ریسک ذخیره در طیف گسترده‌ای از خطوط کسب‌وکار موفق بوده‌اند اما می‌توانند از مشکلاتی نظیر پارامتری کردن بیش از حد، خطر انتقال خطا، عدم استحکام و ناتوانی در تفکیک ارزیابی‌های خسارت‌های واقع‌شده ولی گزارش‌نشده و خسارت‌های گزارش‌شده ولی تسویه‌نشده[5] رنج ببرند (آوانزی و همکاران[6]، ۲۰۲۴، ۹۷۴). همچنین با توجه به نوع داده‌های بیمه‌ای که معمولا دارای چولگی و پراکندگی بالایی هستند، وجود یک سیستم و روش‌مندی مناسب در تجزیه و تحلیل و پیش‌بینی این داده‌ها، اهمیت بسیار چشم‌گیری در کاهش ریسک شرکت‌های بیمه‌ای دارد. پیشرفت‌های اخیر در یادگیری ماشین، از جمله رگرسیون خطی و غیرخطی، جنگل‌های تصادفی، شبکه‌های عصبی و شبکه عصبی با حافظه طولانی کوتاه‌مدت، نقش مهمی در بهبود پیش‌بینی‌های داده‌های بیمه‌ای، بخصوص خسارات واقع‌شده ولی گزارش‌نشده ایفا کرده‌اند. این تکنیک‌ها با استفاده از داده‌های بزرگ و قدرت محاسباتی بالا، الگوها و روابط پیچیده‌ای را در داده‌های خسارات کشف می‌کنند که ممکن است توسط مدل‌های سنتی نادیده گرفته شوند (فونگ، بادسکو، و لین[7]، ۲۰۲۲، ۴۹۹).

از این رو در این مطالعه، با تکنیک‌های سانسور و برش و استفاده از مدل‌های یادگیری ماشین پیشرفته در داده‌های خسارت بیمه‌ای، به این پرسش اساسی پرداخته می‌شود که کدام یک از مدل‌های پیشرفته یادگیری ماشین، شامل رگرسیون خطی و غیرخطی، جنگل تصادفی، شبکه عصبی پرسپترون چندلایه و شبکه عصبی با حافظه طولانی کوتاه‌مدت، در پیش‌بینی دقیق‌تر خسارت‌های واقع‌شده ولی گزارش‌نشده در رشته‌های مختلف

[1] Incurred But Not Reported (IBNR)
[2] Sousa
[3] Balona & Richman
[4] Loss Development Triangles
[5] Reported But Not Settled (RBNS)
[6] Avanzi & et al
[7] Fung, Badescu & Lin

بیمه‌ای عملکرد بهتری دارند و چگونه می‌توان این مدل‌ها را برای شرایط پیچیده و متغیر داده‌های سانسور شده و بریده‌شده بهینه‌سازی کرد؟

در ادامه این مطالعه، مبانی نظری و تحقیقات پیشین مرتبط با موضوع تحقیق بررسی شده و روش پژوهش و مدل‌سازی‌های رگرسیونی و ماشین یادگیری مورد توجه قرار می‌گیرد. همچنین روش‌های ارزیابی عملکرد مدل‌ها و معیارهای سنجش دقت پیش‌بینی‌ها ارائه شده است. بخش نتایج به مقایسه عملکرد مدل‌ها در پیش‌بینی خسارات واقع‌شده ولی گزارش‌نشده در رشته‌های مختلف بیمه‌ای اختصاص دارد. در نهایت بخش پایانی، بحث و نتیجه‌گیری کلی این مطالعه به همراه پیشنهادات برای تحقیقات آینده و کاربردهای عملی در بهینه‌سازی مدل‌ها برای داده‌های بیمه‌ای پیچیده و متغیر ارائه می‌شود.

۲- مبانی نظری و پیشینه مطالعات تجربی

پیش‌بینی آینده از جمله نیازهای اساسی در زندگی روزمره و یکی از حوزه‌های کلیدی در بسیاری از علوم، به‌ویژه در مسائل مالی و اقتصادی، به‌شمار می‌رود (پورزمانی، ۱۳۹۴، ۸۲). در واقع، توانایی پیش‌بینی دقیق نقشی اساسی در تصمیم‌گیری‌های اقتصادی و مدیریتی دارد، زیرا اثربخشی هر تصمیم در نهایت به تحقق نتایج پیش‌بینی‌شده بستگی دارد. از این رو، پیش‌بینی برای بسیاری از سازمان‌ها و نهادها به‌ویژه در صنعت بیمه، به‌عنوان ابزاری برای مقابله با ریسک‌ها ضروری است (آوانزی و همکاران، ۲۰۲۴، ۹۷۳). بازار بیمه که ذاتاً با مفهوم ریسک درآمیخته است، اهمیت خاصی برای پیش‌بینی دقیق خسارت‌های واقع‌شده اما گزارش‌نشده دارد. این پیش‌بینی‌ها بیمه‌گران را قادر می‌سازد تا ذخایر کافی برای پوشش خسارت‌های آتی فراهم آورده و ثبات مالی و مدیریت ریسک خود را بهبود بخشند.

در ادبیات سنتی پیش‌بینی خسارات واقع‌شده ولی گزارش‌نشده، مدل‌هایی مانند زنجیره نردبان[1] و بورنهاتر-فرگوسن[2] استفاده می‌شوند که این مدل‌ها اگرچه در بسیاری از موارد نتایج قابل قبولی ارائه داده‌اند، اما با محدودیت‌هایی نیز مواجه هستند. این مدل‌ها برای ارائه نتایج پیش‌بینی بر فرضیات ساده‌کننده‌ای استوارند که گاهی موجب کاهش دقت آن‌ها می‌شود. علاوه بر این، این روش‌ها اغلب قادر به استفاده از تمامی ویژگی‌های موجود در داده‌های خسارت‌های فردی نیستند و در مواجهه با تغییرات ناگهانی یا نوسانات شدید در داده‌ها ممکن است عملکرد مناسبی نداشته باشند. برخی از محققان، از جمله بالونا و ریچمن (۲۰۲۱)، این مدل‌ها را به دلیل مشکلات پارامتری کردن بیش‌ازحد، خطر انتقال خطا و عدم استحکام، مورد انتقاد قرار داده‌اند. با این حال، این مدل‌ها به دلیل سادگی و قابلیت فهم بالا همچنان مورد استفاده قرار می‌گیرند و در برخی موارد نیز همچنان کارایی مناسبی دارند (بالونا و ریچمن، ۲۰۲۱، ۵). در مواجهه با این محدودیت‌ها، تحقیقات جدید رویکردهای نوینی را ارائه کرده‌اند. مدل‌های کاپولا[3] یکی از رویکردهای نوین هستند که به دلیل توانایی‌شان در مدل‌سازی

[1] Chain Ladder
[2] Bornhuetter-Ferguson
[3] Copula Models

وابستگی‌های پیچیده میان متغیرها، توجه زیادی را به خود جلب کرده‌اند. فارکاس و لوپز[1] (۲۰۲۲)، معتقدند که استفاده از مدل‌های کاپولا می‌تواند پیچیدگی‌های بین متغیرها را بهتر مدیریت کرده و دقت پیش‌بینی‌ها را افزایش دهد. با این حال، همچنان به دلیل عدم تطبیق کامل این مدل‌ها با داده‌های بزرگ و پیچیده، این رویکرد نیز با محدودیت‌هایی مواجه است (فارکاس و لوپز، ۲۰۲۴، ۱۰۶۸). تحقیقات اخیر به‌ویژه بر ذخیره‌گیری خسارت بر اساس داده‌های فردی تاکید دارند. آنتونیو و پلات[2] (۲۰۱۴) و بادسکو و همکاران[3] (۲۰۱۶)، این نیاز را به دلیل انعطاف‌پذیری کم مدل‌های آماری سنتی بیان کردند و معتقد بودند که استفاده از تکنیک‌های یادگیری ماشینی، که انعطاف‌پذیری بیشتری در برابر تغییرات دارند، ضروری است. در این راستا، ووتریش[4] (۲۰۱۸)، در پژوهش خود نشان داد که می‌توان با استفاده از درخت رگرسیون ذخیره‌سازی[5] خسارت‌های فردی را بهینه‌تر کرد، هرچند این مدل نیز هنوز تنها به تعداد پرداخت‌ها توجه دارد و جنبه‌های دیگری از داده‌های خسارت را در نظر نمی‌گیرد. میراندا و همکاران[6] (۲۰۱۳) نیز نشان دادند که مثلث‌های توسعه خسارت[7] که در بیمه‌های غیرزندگی برای ذخیره‌گیری خسارت استفاده می‌شوند، می‌توانند به صورت یک چارچوب پیوسته در نظر گرفته شوند و به تخمین توزیع خسارت‌ها در بخش‌های مختلف مثلث کمک کنند. این پژوهش به دو رویکرد تحقیقاتی منجر شد: رویکرد اول بر تخمین تابع چگالی زیربنایی تمرکز دارد که مدل‌های مختلفی نظیر مدل‌های لی، مامن، نیلسن و پارک[8] (۲۰۱۵، ۲۰۱۷) و همچنین مدل مامن، مارتینز-میراندا، نیلسن[9] (۲۰۲۱) به منظور در نظر گرفتن اثرات فصلی، زمان عملیاتی و اثرات تقویمی توسعه یافته‌اند. رویکرد دوم، با تمرکز بر مسئله سانسور، توسط هیابو، مامن، مارتینز-میراندا و نیلسن[10] (۲۰۱۶) مطرح شد که سانسور راست و چپ را در تحلیل‌های خود وارد کردند. نتایج این تحقیقات نشان می‌دهد که در شرایط مختلف، تحلیل‌های سانسور شده می‌توانند به‌طور موثری به بهبود پیش‌بینی‌ها کمک کنند. در سال‌های اخیر، استفاده از تکنیک‌های پیشرفته یادگیری ماشین، شامل جنگل‌های تصادفی، شبکه‌های عصبی و شبکه عصبی با حافظه طولانی کوتاه‌مدت، در پیش‌بینی خسارات واقع‌شده ولی گزارش‌نشده توجه بیشتری به خود جلب کرده است. این تکنیک‌ها با استفاده از داده‌های بزرگ و قدرت محاسباتی بالا قادر به کشف الگوهای پیچیده و روابط غیرخطی در داده‌های خسارت هستند که مدل‌های سنتی ممکن است نادیده بگیرند (فونگ، بادسکو و لین، ۲۰۲۱، ۴۹۸). مدل‌های یادگیری ماشین، به دلیل توانایی خود در کاهش ابعاد داده‌ها به صورت داده‌محور، از پیچیدگی‌های ابعادی فراتر می‌روند و قادرند ساختارهای زیرین داده‌ها را شناسایی کنند. برای نمونه، وُترش (۲۰۱۸) نشان داد که استفاده از شبکه عصبی پیش‌خور می‌تواند عوامل توسعه را به‌طور مستقیم تخمین

[1] Farkas & Lopez
[2] Antonio & Plat
[3] Badescu & et al
[4] Wuthrich
[5] Reservoir Regression Tree
[6] Miranda & et al
[7] Loss Development Triangles
[8] Lee, Mammen, Nielsen & Park,
[9] Mammen, Martínez-Miranda, Nielsen
[10] Hiabu, Mammen, Martínez-Miranda & Nielsen,

بزند و مشکلات ناشی از ورودهای صفر در مثلث‌های توسعه خسارت را برطرف کند. کالسترو-وانگاس، بادسکو و لین[1] (۲۰۲۳) از تابع ریسک تخمینی برای پیش‌بینی استفاده کردند و از مدل ریسک متناسب کاکس بهره بردند. آن‌ها با استفاده از این مدل موفق به ایجاد یک چارچوب پیش‌بینی شدند که با مدل‌های ذخیره‌گیری استاندارد قابل مقایسه است. این چارچوب نیز از نظر برخی محققان به‌خصوص به دلیل استفاده از وزن‌دهی معکوس احتمال و تبدیل تابع ریسک به عوامل توسعه مورد توجه قرار گرفته است. این پیشرفت‌های جدید در رشته‌های مختلف بیمه، از جمله بیمه مسئولیت شخص ثالث، بیمه خودرو، بیمه آتش‌سوزی و بیمه عمر کاربردهای فراوانی دارند. در هر کدام از این رشته‌ها، چالش‌ها و ویژگی‌های خاص داده‌ای وجود دارند که نیاز به مدل‌های متناسب برای افزایش دقت پیش‌بینی دارند. با توجه به چالش‌های موجود در فرآیند گزارش‌دهی و تسویه خسارات، برآورد دقیق خسارات واقع‌شده ولی گزارش‌نشده نیازمند استفاده از تکنیک‌های آماری و اکچوئری پیچیده و قضاوت دقیق است (آندرسون[2]، ۲۰۲۳، ۵).

مطالعات داخلی و خارجی متعددی به بررسی روش‌های مختلف برای برآورد ذخایر خسارت پرداخته‌اند. جانفشان و بی‌تا (۱۳۸۵)، روش نردبان زنجیره‌ای و رگرسیون لگاریتم خطی را مقایسه کرده و نشان دادند که روش دوم دقت بیشتری دارد. شهریار و همکاران (۱۳۹۴) روش بورن هاتر-فرگسن را برای بهبود دقت برآورد ذخایر معرفی کردند. شکوری و همکاران (۱۴۰۲) نیز دریافتند که مدلبندی وابستگی تقویمی بین خسارت‌های معوق در مثلث‌های تأخیر با استفاده از توزیع چندمتغیره به پیش‌بینی دقیق‌تر ذخایر مربوط به خسارت‌های معوق نسبت به استفاده از عامل اثر سال تقویمی در مدل میانگین توزیع خسارت‌های معوق منجر می‌شود. در مطالعات خارجی، آوانزی و همکاران (۲۰۲۴) نشان دادند که ترکیب مدل‌های مختلف می‌تواند عملکرد پیش‌بینی را بهبود بخشد. هیابو و همکاران[3] (۲۰۲۴) با استفاده از تحلیل بقای مبتنی بر یادگیری ماشین دقت پیش‌بینی‌ها را افزایش دادند. چانگ و همکاران[4] (۲۰۲۴) با استفاده از اسپلاین‌های یکنواخت به بهبود دقت پیش‌بینی ذخایر خسارات واقع‌شده ولی گزارش‌نشده پرداختند. مائیت[5] (۲۰۲۳) نشان داد که تفاوت قابل توجهی بین تخمین ذخایر با استفاده از روش‌های مختلف محاسبه فاکتور توسعه وجود ندارد.

تحقیق حاضر با تمرکز بر پیش‌بینی خسارات واقع‌شده اما گزارش‌نشده در رشته‌های مختلف بیمه‌ای و با استفاده از داده‌های جزئی، از چندین روش پیشرفته یادگیری ماشین از جمله رگرسیون خطی، جنگل تصادفی، شبکه عصبی و LSTM بهره می‌گیرد. این رویکرد نه‌تنها به تحلیل دقیق‌تر و جزئی‌تر از داده‌ها پرداخته، بلکه تأثیر عواملی مانند روزهای پرترافیک مسافرت، تعطیلات، فصول سرد، مناطق خشک و فصول ساخت‌وساز را نیز در نظر می‌گیرد. نوآوری این تحقیق در به‌کارگیری مدل‌های پیشرفته برای تحلیل داده‌های جزئی و ترکیب آن‌ها با

[1] Calcetero-Vanegas, Badescu & Lin
[2] Andersson
[3] Hiabu & et al
[4] Chang
[5] Maait

داده‌های زمانی (مانند روزهای خاص) و همچنین اعمال روش‌های سانسور و برش داده‌های پراکنده و مقیاس بزرگ است که بهبود چشمگیری در دقت و کارایی پیش‌بینی‌ها ایجاد می‌کند.

۳- روش پژوهش

این پژوهش از نوع کاربردی بوده و با هدف شناسایی عوامل مؤثر در پیش‌بینی خسارت‌های واقع‌شده اما گزارش‌نشده و بهره‌گیری از الگوریتم‌های هوشمند برای پیش‌بینی دقیق آن‌ها انجام شده است. همچنین، با توجه به هدف آن، این پژوهش یک نوع آینده‌پژوهی نیز محسوب می‌شود. در راستای اجرای آینده‌پژوهی، از روش تحلیل ساختاری برای شناسایی و ارزیابی متغیرهای مؤثر و تأثیرگذار در پیش‌بینی خسارات واقع‌شده ولی گزارش‌نشده استفاده شد. در مرحله نخست، متغیرهای اصلی از جمله تاریخ حادثه، تاریخ گزارش، تأخیر در گزارش‌دهی، روز هفته، روز ماه، ماه، هفته سال، سال، تعطیلات نوروز، تعطیلات تابستانی و دوره‌های شلوغ مسافرتی شناسایی و استخراج شدند. در این میان، تعطیلات نوروز، تعطیلات تابستانی و دوره‌های شلوغ مسافرتی به طور خاص برای رشته‌های ثالث مالی، ثالث جانی و بدنه اهمیت ویژه دارند، زیرا این دوره‌ها می‌توانند بر شدت و فراوانی خسارت‌های این بیمه‌ها تأثیر مستقیم داشته باشند. همچنین، فصول سرد، مناطق خشک و فصول ساخت‌وساز به‌عنوان عوامل مهم در خسارات مربوط به رشته‌های آتش‌سوزی و مسئولیت شناسایی شدند و می‌توانند در تحلیل و پیش‌بینی دقیق‌تر این خسارات تأثیرگذار باشند. در مرحله دوم، به‌منظور کاهش هم‌خطی و بهبود دقت مدل‌ها، ویژگی‌هایی که همبستگی بالایی با یکدیگر داشتند حذف شدند. داده‌های گم‌شده نیز به‌منظور جلوگیری از کاهش دقت مدل‌ها حذف گردیدند. در ادامه، به‌دلیل وجود داده‌های سانسور شده و برش داده‌ها، مراحل پیش‌پردازش داده‌ها با دقت بیشتری انجام شد. داده‌های سانسور شده به داده‌هایی اشاره دارند که به دلیل ثبت ناقص یا عدم دسترسی کامل، برخی اطلاعات را از دست داده‌اند. در این پژوهش، برای مدیریت این داده‌ها، از روش‌های تحلیل بقای تخصصی جهت برآورد داده‌های ناقص استفاده شد. در خصوص برش داده‌ها، این پژوهش به تحلیل دوره‌های زمانی خاص از جمله تعطیلات نوروز، تابستان، آخر هفته‌ها و دوره‌های پرتردد مسافرتی برای رشته‌های ثالث مالی، ثالث جانی و بدنه و همچنین دوره‌های فصلی و جغرافیایی خاص همچون فصول سرد، مناطق خشک و فصول ساخت‌وساز برای رشته‌های آتش‌سوزی و مسئولیت پرداخته است. این برش زمانی و مکانی به تحلیل دقیق‌تر الگوهای خسارت در دوره‌ها و شرایط خاص کمک کرده و امکان شناسایی روندهای خاص در داده‌ها را فراهم کرده است. در مرحله سوم، متغیرهای کلیدی مؤثر در پیش‌بینی خسارات واقع‌شده ولی گزارش‌نشده، شامل روز هفته، روز ماه، ماه، هفته سال، سال و عوامل خاصی چون تعطیلات نوروز، تعطیلات تابستانی و دوره‌های شلوغ مسافرتی برای بیمه ثالث مالی، جانی و بدنه و همچنین فصول سرد، مناطق خشک و فصول ساخت‌وساز برای بیمه آتش‌سوزی و مسئولیت، به‌عنوان ورودی مدل‌های یادگیری ماشین به‌کار گرفته شدند. این متغیرها با استفاده از تکنیک‌های یادگیری ماشین و الگوریتم‌های پیشرفته، به پیش‌بینی دقیق‌تر و پایدارتر ذخایر خسارت‌ها کمک می‌کنند. پس از تعیین نهایی عوامل مؤثر در مرحله تحلیل ساختاری، مدل‌های یادگیری ماشین برای ارزیابی مدل پژوهش به کار گرفته شدند. جامعه آماری این پژوهش شامل اطلاعات تاریخ‌های وقوع و اعلام حادثه مربوط به شرکت‌های بیمه در ایران

است و نمونه آماری شامل داده‌های تاریخ وقوع و اعلام حادثه بین سال‌های ۱۴۰۰ تا ۱۴۰۱ در شرکت بیمه ایران و در رشته‌های مختلف بیمه‌ای همچون بیمه ثالث مالی، بیمه بدنه، بیمه ثالث جانی و حوادث راننده، بیمه آتش‌سوزی و بیمه مسئولیت است. این داده‌ها از سیستم‌های اطلاعاتی شرکت بیمه ایران به‌صورت کامل و بدون فیلترسازی اولیه استخراج شده‌اند و ویژگی‌های جامعی را در این رشته‌های مختلف بیمه‌ای پوشش می‌دهند.

در این پژوهش، شش مدل یادگیری ماشین شامل رگرسیون خطی چندگانه[۱]، مدل خطی تعمیم‌یافته[۲]، مدل افزودنی تعمیم‌یافته[۳]، جنگل تصادفی[۴]، شبکه عصبی پرسپترون چندلایه[۵] و شبکه عصبی حافظه کوتاه‌مدت بلند[۶] به‌کار گرفته شدند تا دقت پیش‌بینی تأخیر گزارش‌دهی خسارت‌ها ارزیابی شود. هر یک از این مدل‌ها ابتدا با داده‌های آموزشی آموزش دیده و سپس با استفاده از داده‌های آزمایشی ارزیابی شدند. داده‌ها به دو مجموعه آموزش (۸۰٪) و تست (۲۰٪) تقسیم شده‌اند تا عملکرد مدل مورد ارزیابی قرار گیرد. برای مقایسه عملکرد مدل‌ها، از معیار ریشه میانگین مربعات خطا (RMSE) و ضریب تعیین (R^2) استفاده شد تا دقت و قابلیت مدل‌ها در پیش‌بینی تأخیر گزارش‌دهی مشخص گردد. ویژگی‌های استخراج‌شده از داده‌ها شامل تاریخ حادثه، تاریخ گزارش، تأخیر در گزارش‌دهی، روز هفته، روز ماه، ماه، هفته سال، سال، تعطیلات نوروز، تعطیلات تابستانی و دوره‌های شلوغ مسافرتی بوده‌اند. این ویژگی‌ها با هدف تحلیل دقیق‌تر تأثیر عوامل مختلف بر تأخیر گزارش‌دهی خسارات و بهبود دقت مدل‌های پیش‌بینی به کار گرفته شدند. در فرآیند آماده‌سازی داده‌ها، هم‌خطی بین متغیرهای دارای همبستگی بالا حذف شده و داده‌های گم‌شده نیز برای جلوگیری از کاهش دقت مدل‌ها حذف گردیدند. همچنین فرضیه اصلی این پژوهش چنین است: مدل‌های یادگیری ماشین، به ویژه مدل‌هایی مانند شبکه عصبی و حافظه کوتاه‌مدت بلند و مدل جنگل تصادفی، دقت بالاتری نسبت به مدل‌های سنتی در پیش‌بینی خسارات واقع‌شده اما گزارش‌نشده در شرکت بیمه ایران دارند.

در این تحقیق با استفاده از ماشین یادگیری که شاخه‌ای از هوش مصنوعی است، به بررسی فرضیه تحقیق پرداخته‌ایم. هوش مصنوعی[۷] شاخه ای از علوم رایانه است که هدف اصلی آن تولید ماشین‌های هوشمند با توانایی انجام وظایف هوش انسانی است. که درحقیقت نوعی شبیه سازی هوش انسانی برای کامپیوتر است. در این فرآیند در واقع ماشین به گونه ای برنامه نویسی شده که همانند انسان فکر کند و توانایی تقلید از رفتار انسان را داشته باشد . این تعریف می تواند به تمامی ماشین‌هایی اطلاق شود که بگونه‌ای همانند ذهن انسان عمل می‌کنند و می‌توانند کارهایی مانند حل مسئله و یادگیری داشته باشند (کیایی و فرشادفر، ۱۴۰۳، ۳۸۶). یادگیری ماشینی، شاخه‌ای از هوش مصنوعی است که با ابداع الگوریتم‌های مختلف، به تدریج عملکرد خود بر روی یک مسئله خاص

[1] Multiple Linear Regression (MLR)
[2] Generalized Linear Model (GLM)
[3] Generalized Additive Model (GAM)
[4] Random Forest
[5] Multi-Layer Perceptron (MLP) neural network
[6] Long Short-Term Memory (LSTM) neural network
[7] Artificial Intelligence

را بهبود می‌بخشد. یادگیری ماشینی برای یافتن الگوها و کندوکاو تغییرات کوچک، بر پایه بررسی و مقایسه داده‌هایی از مقادیر کوچک تا حجم‌های عظیم داده استوار است (صیادی نژاد و همکاران، ۱۴۰۲، ۲۱۹).

در این تحقیق، شش مدل یادگیری ماشین[1] شامل رگرسیون خطی چندگانه (MLR)، مدل خطی تعمیم‌یافته (GLM)، مدل افزودنی تعمیم‌یافته (GAM)، جنگل تصادفی (Random Forest)، شبکه عصبی پرسپترون چندلایه (MLP)و شبکه عصبی حافظه کوتاه‌مدت بلند (LSTM) مورد استفاده قرار گرفته است که در ادامه به بررسی روش‌شناسی آنها خواهیم پرداخت.

۳-۱- مدل رگرسیون خطی چندگانه (MLR)

مدل‌سازی تأخیر در گزارش ادعاهای بیمه‌ای به روش رگرسیون خطی چندگانه به ما اجازه می‌دهد تا رابطه خطی بین متغیرهای مستقل (مانند روز هفته، ماه، تعطیلات و دوره‌های پرتردد) و متغیر وابسته (تأخیر در گزارش ادعا) را مدل‌سازی کنیم (اسمیت و دو[2]، ۲۰۲۰، ۴۵۹). مدل رگرسیون خطی چندگانه به صورت معادله (۱) تعریف می‌شود:

$$\mathrm{Delay}_i = \beta_0 + \beta_1\,\mathrm{DayOfWeek}_i + \beta_2\,\mathrm{DayOfMonth}_i + \beta_3\,\mathrm{Month}_i + \beta_4\mathrm{WeekOfYear}_i + \beta_5\mathrm{Year}_i + \beta_6\mathrm{IsWeekend}_i + \beta_7\mathrm{IsNowruz}_i + \beta_8\mathrm{IsSummer}_i + \beta_9\mathrm{IsBusyTravel}_i + \beta_{10}\mathrm{WeekdayMonth}_i + \beta_{11}\mathrm{MonthYear}_i + \epsilon_i \quad (۱)$$

که در آن Delay_i: تأخیر در گزارش برای نمونه i است. $\mathrm{DayOfWeek}_i$، $\mathrm{DayOfMonth}_i$، Month_i، $\mathrm{WeekOfYear}_i$، Year_i، ویژگی‌های زمانی استخراج‌شده از تاریخ وقوع حادثه هستند. $\mathrm{IsWeekend}_i$، نشان‌دهنده تعطیلات آخر هفته است. $\mathrm{IsNowruz}_i$، متغیر تعطیلات نوروز است. $\mathrm{IsSummer}_i$، نشان‌دهنده تعطیلات تابستانی است. $\mathrm{IsBusyTravel}_i$، ترکیبی از تعطیلات و دوره‌های پرتردد است. $\mathrm{WeekdayMonth}_i$، ویژگی ترکیبی از روز هفته و ماه؛ $\mathrm{MonthYear}_i$، ویژگی ترکیبی از ماه و سال و ϵ_i: خطای تصادفی مدل می‌باشد.

برای ارزیابی دقت مدل رگرسیون خطی چندگانه، از معیارهای جذر میانگین مربعات خطا (RMSE) و ضریب تعیین(R^2) که در معادلات (۲) و (۳) مشاهده می‌شود، استفاده شده است:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n}\left(\widehat{\mathrm{Delay}}_i - \mathrm{Delay}_i\right)^2} \quad (۲)$$

که در آن n، تعداد نمونه‌ها در مجموعه آزمایشی و $\widehat{\mathrm{Delay}}_i$: مقدار پیش‌بینی شده تأخیر برای نمونه i است.

برای محاسبه ضریب تعیین نیز از معادله ۳ استفاده کرده‌ایم:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\widehat{\mathrm{Delay}}_i - \mathrm{Delay}_i\right)^2}{\sum_{i=1}^{n}\left(\bar{\mathrm{Delay}}_i - \mathrm{Delay}_i\right)^2} - 1 \quad (۳)$$

[1] Machine Learning
[2] Smith & Doe

که در آن $\bar{\text{De}}\text{lay}_i$ میانگین تأخیرها در کل مجموعه داده است (هستی و همکاران[1]، ۲۰۰۹، ۱۸۴).

۳-۲- مدل خطی تعمیم‌یافته (GLM)

مدل خطی تعمیم‌یافته (GLM) و رگرسیون چندگانه هر دو برای مدل‌سازی رابطه متغیر وابسته و متغیرهای مستقل به کار می‌روند، اما تفاوت‌های اساسی دارند. رگرسیون چندگانه تنها توزیع نرمال متغیر وابسته را می‌پذیرد، در حالی که GLM امکان استفاده از توزیع‌های مختلف مانند پواسون و دوجمله‌ای را دارد و آن را برای داده‌های متنوع مناسب می‌سازد (مونتگومری و همکاران[2]، ۲۰۱۲، ۴۵). GLM با استفاده از توابع پیوند، روابط غیرخطی را نیز مدل‌سازی می‌کند، در حالی که رگرسیون چندگانه تنها از روابط خطی استفاده می‌کند (دابسون و بارنت[3]، ۲۰۱۸، ۹۵). به دلیل انعطاف‌پذیری و توانایی کار با داده‌های باینری و شمارشی، GLM برای تحلیل داده‌های پیچیده کاربرد بیشتری دارد، هرچند محاسبات آن پیچیده‌تر از رگرسیون چندگانه است (نلدر و ودر برن[4]، ۱۹۷۲، ۱۳۰). بنابراین، رگرسیون چندگانه برای داده‌های پیوسته و نرمال مناسب است، اما GLM برای داده‌های متنوع و روابط غیرخطی گزینه بهتری است (هستی و همکاران، ۲۰۰۹، ۱۸۵).

۳-۳- مدل جنگل تصادفی

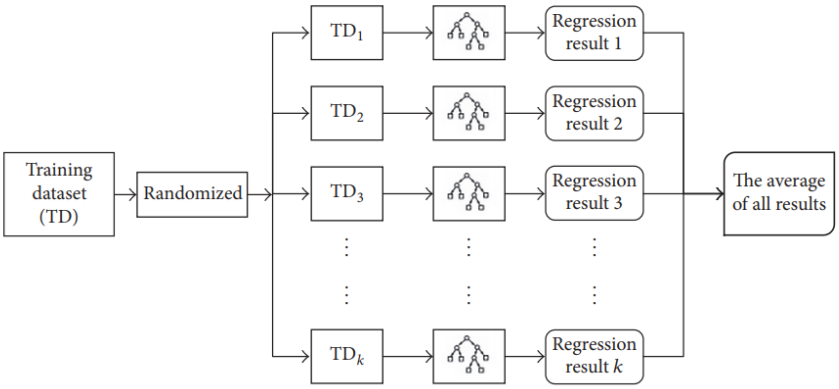
مدل جنگل تصادفی یک روش یادگیری ماشین است که به ما امکان می‌دهد تا روابط پیچیده و غیرخطی بین متغیرهای مستقل (مانند روز هفته، ماه، تعطیلات و دوره‌های پرتردد) و متغیر وابسته (تأخیر در گزارش ادعا) را مدل‌سازی کنیم. این روش با ترکیب تعداد زیادی از درختان تصمیم‌گیری و استفاده از نمونه‌گیری تصادفی از داده‌ها و ویژگی‌ها، دقت پیش‌بینی را افزایش داده و واریانس خطا را کاهش می‌دهد (غلامیان و داوودی، ۱۳۹۷، ۳۰۵ و هستی و همکاران، ۲۰۰۹، ۵۸۷). مدل جنگل تصادفی از مجموعه‌ای از درختان تصمیم‌گیری تشکیل شده است. هر درخت با استفاده از یک نمونه تصادفی از داده‌ها و زیرمجموعه‌ای از ویژگی‌ها آموزش داده می‌شود. سپس پیش‌بینی نهایی مدل با میانگین‌گیری از پیش‌بینی‌های هر درخت محاسبه می‌شود. فرمول نهایی مدل به صورت معادله (۴) بیان می‌شود:

$$y = T_j(x) \sum_{j=1}^{n} \frac{1}{n} \qquad (۴)$$

که در آن y، پیش‌بینی نهایی یا برآورد مدل برای متغیر وابسته؛ n، تعداد درختان در مدل جنگل تصادفی و $T_j(x)$، پیش‌بینی درخت j برای نمونه x است.

[1] Hastie et al
[2] Montgomery et al
[3] Dobson & Barnett
[4] Nelder & Wedderburn

نمودار ۱، ساختار مدل جنگل تصادفی را به طور خلاصه به تصویر می‌کشد. در این مدل، داده‌های آموزشی به‌صورت تصادفی به چندین زیرمجموعه TD_1 تا TD_k تقسیم می‌شوند. سپس هر زیرمجموعه برای ایجاد یک درخت تصمیم‌گیری مجزا (درخت رگرسیون) استفاده می‌شود. نتایج هر درخت به‌صورت یک پیش‌بینی مستقل به دست می‌آید و در نهایت، میانگین تمامی نتایج محاسبه می‌شود تا پیش‌بینی نهایی مدل ایجاد گردد. این فرآیند به مدل اجازه می‌دهد تا دقت پیش‌بینی را افزایش داده و اثر نوسانات را کاهش دهد.



نمودار ۱- نمودار ساختار مدل جنگل تصادفی (Random Forest)

منبع: بریمن، ۲۰۰۱

۳-۴- مدل افزایشی تعمیم‌یافته[1]

مدل افزایشی تعمیم‌یافته (GAM) یک روش انعطاف‌پذیر برای مدل‌سازی رابطه بین یک متغیر وابسته و چند متغیر مستقل است که می‌تواند روابط غیرخطی پیچیده را بدون نیاز به تعریف دقیق تابع پیوند خطی مدل‌سازی کند. مدل افزایشی تعمیم‌یافته، این امکان را فراهم می‌کند تا تأثیر متغیرهای مستقل به صورت افزایشی و به صورت توابعی انعطاف‌پذیر از این متغیرها مدل‌سازی شود (هستی و تیبشیرانی[2]، ۱۹۹۰، ۱).

مدل مدل افزایشی تعمیم‌یافته به صورت معادله (۵) است:

$$Delay_i = \beta_0 + f_1(DayOfWeek_i) + f_2(DayOfMonth_i) + f_3(Month_i) + f_4(WeekOfYear_i) + f_5(Year_i) + f_6(IsWeekend_i) + f_7(IsNowruz_i) + f_8(IsSummer_i) + f_9(IsBusyTravel_i) + f_{10}(WeekdayMonth_i) + f_{11}(MonthYear_i) + \epsilon_i \quad (۵)$$

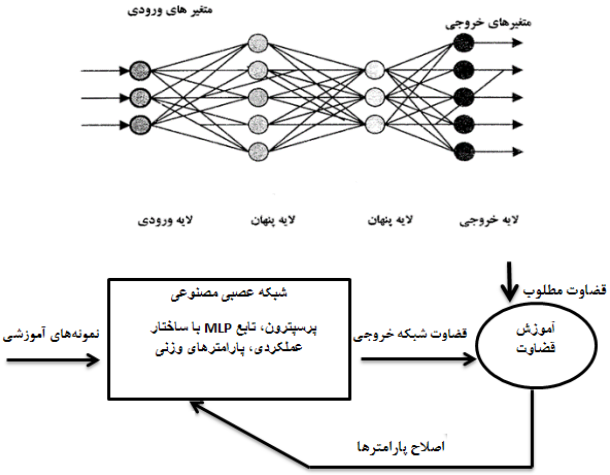
[1] Generalized Additive Model - GAM
[2] Hastie & Tibshirani

که در آن f_i، توابع افزایشی هستند که به طور انعطاف‌پذیر تأثیر هر متغیر مستقل بر تأخیر گزارش را مدل‌سازی می‌کنند. درGAM، هر تابع افزایشی f_i می‌تواند به‌طور جداگانه به عنوان یک تابع غیرخطی از هر متغیر مستقل بهینه‌سازی شود. (وود[1]، ۲۰۱۷، ۱۲۳).

۳-۵- شبکه عصبی چند لایه پرسپترون (MLP)[2]

در این تحقیق، از شبکه عصبی چندلایه پرسپترون با الگوریتم یادگیری پس‌انتشار استفاده شده است. شبکه عصبی چندلایه پرسپترون، شبیه به سیستم پردازش شبکه‌های عصبی بیولوژیکی، شامل گره‌هایی با توابع فعال‌سازی است که در لایه‌های مختلف قرار دارند. این شبکه شامل یک لایه ورودی، یک یا چند لایه پنهان، و یک لایه خروجی است که هر گره خروجی لایه قبلی را با ضرایب وزنی پردازش کرده و به لایه بعدی منتقل می‌کند. اولین مرحله در این روش تعیین معماری شبکه شامل تعداد لایه‌های پنهان، تعداد نرون‌ها در هر لایه، انتخاب توابع فعال‌سازی، نرمال‌سازی داده‌ها، و تعیین داده‌های آموزش و تست است. در مرحله دوم، طراحی الگوریتم صورت می‌گیرد. در شبکه‌های عصبی، یادگیری معمولاً به صورت نظارت‌شده انجام می‌شود، و الگوریتم پس‌انتشار خطا به‌عنوان رایج‌ترین روش در آموزش شبکه‌های چندلایه استفاده می‌شود. این الگوریتم خطاهای پیش‌بینی را کاهش داده و به شبکه آموزش می‌دهد تا دقت پیش‌بینی‌ها بهبود یابد (شهبازی و پیله ور ، ۱۳۹۴، ۲۱۳).

در نمودار شماره ۲، ساختار و یادگیری روش شبکه عصبی چندلایه بصورت خلاصه شده توضیح داده شده است.



نمودار ۲- شبکه عصبی استاندارد پیشخور و یادگیری آن

منبع: شهبازی و پیله ور، ۱۳۹۴

[1] Wood
[2] Multi-Layer Perceptron

۳-۶- شبکه عصبی با حافظه طولانی کوتاه‌مدت (LSTM)

شبکه عصبی با حافظه طولانی کوتاه‌مدت، یک مدل شبکه عصبی است که ابتدا در سال ۱۹۹۷ پیشنهاد شد. ویژگی کلیدی این مدل این است که می‌توانند اطلاعاتی را ذخیره کنند که برای پردازش سلولی آینده استفاده شود. حافظه‌ی این شبکه دارای دو بردار کلیدی است: ۱. حالت کوتاه مدت: خروجی را در مرحله‌ی زمانی فعلی حفظ می‌کند. ۲. حالت بلند مدت: مواردی را که برای طولانی مدت در نظر گرفته شده‌اند را در حین عبور از شبکه ذخیره می‌کند، می‌خواند و یا فراموش می‌کند (بابانژاد باقری و همکاران، ۱۴۰۲، ۲۹۷).

این شبکه‌ها برای پردازش سری‌های زمانی با تأخیرهای نامشخص مناسب بوده و از الگوریتم پس انتشار برای آموزش مدل استفاده می‌کنند. ساختار شبکه عصبی با حافظه طولانی کوتاه‌مدت از نوع LSTM در نمودار ۳ نشان داده شده است.

در یک شبکه عصبی با حافظه طولانی کوتاه‌مدت، سه دروازه وجود دارد:

۱) دروازه ورودی: تشخیص می‌دهد که از کدام مقدار ورودی باید برای بهبود حافظه استفاده شود. تابع سیگموئید تصمیم می‌گیرد که کدام مقادیر را از ۰ و ۱ عبور دهد. تابع tanh به مقادیر عبور کرده، بر اساس اهمیت آنها، وزنی در بازه ۱- تا ۱ می‌دهد.

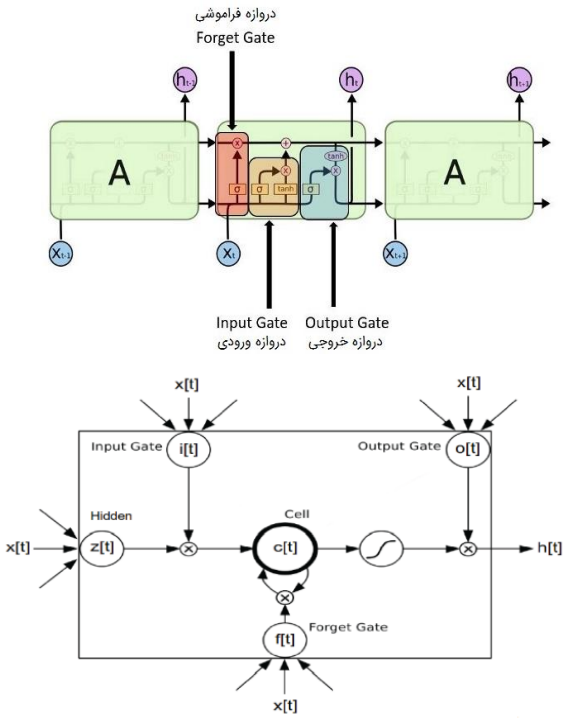
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$\tilde{C}_t = \tanh\,(W_C \cdot [h_{t-1}, x_t] + b_C) \qquad (۶)$$

۲) دروازه فراموشی: تشخیص می‌دهد چه جزئیاتی را باید از بلوک دور انداخت. این موضوع توسط تابع سیگموئید تصمیم‌گیری می‌شود. تابع سیگموئید، به حالت قبلی (h_{t-1}) و ورودی محتوا (X_t) نگاه می‌کند و برای هر عدد در وضعیت سلولC_{t-1} ، عددی بین ۰ (این را حذف کنید) و ۱ (این را نگه دارید) به عنوان خروجی برمی گرداند.

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big) \qquad (۷)$$

۳) دروازه خروجی: از ورودی و حافظه بلوک برای تصمیم‌گیری در مورد خروجی استفاده می‌شود. تابع سیگموئید تصمیم می‌گیرد که کدام مقادیر را از ۰ و ۱ عبور دهد. تابع tanh به مقادیر عبور کرده، بر اساس اهمیت آنها، وزنی در بازه ۱- تا ۱ می‌دهد و با خروجی تابع سیگموئید ضرب می‌شود.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$
$$h_t = o_t * \tanh\,(C_t) \qquad (۸)$$



نمودار ۳- ساختار LSTM با گیت فراموشی

منبع: گودفلو و همکاران[1]، ۲۰۱۶

این ساختار، شبکه‌های LSTM را برای مدل‌سازی داده‌های پیچیده و سری‌های زمانی مؤثر می‌سازد (هستی و تیبشیرانی، ۲۰۱۷، ۸۴).

۴- یافته ها

۴-۱- توصیف داده‌ها

در این بخش، نتایج توصیفی برای تأخیر در گزارش‌دهی خسارت‌های بیمه‌ای در پنج رشته مختلف بیمه‌ای شامل بیمه بدنه اتومبیل، آتش‌سوزی، جانی، شخص ثالث مالی و مسئولیت بررسی شده است. جدول ۱، شاخص‌های مهمی همچون تعداد داده‌ها، میانگین، میانه، انحراف معیار، کمترین و بیشترین تأخیر را برای هر رشته بیمه‌ای نشان می‌دهد.

[1] Goodfellow & et al

جدول ۱- آمار توصیفی برای تأخیر در رشته‌های مختلف بیمه‌ای شرکت بیمه ایران

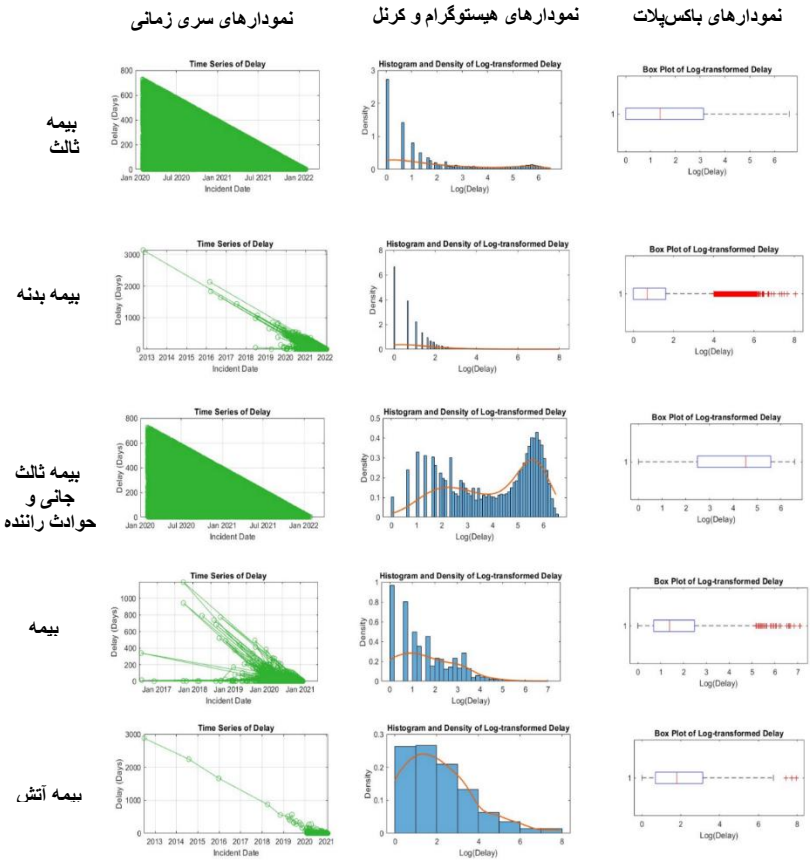
رشته بیمه‌ای	تعداد داده‌ها	میانگین تأخیر	میانه تأخیر	انحراف معیار	کمترین تأخیر	بیشترین تأخیر
بیمه بدنه اتوموبیل	۱۵۶۵۵۰	۷۸/۷	۲۳	۱۴/۳	۱	۳۶۵
بیمه آتش‌سوزی	۱۴۷۶۶	۶۴/۴	۳۸	۴۱/۵۴	۲۰	۳۶۵
بیمه جانی	۱۵۳۸۸۰	۱۰۷/۸	۶۴	۷۵/۱	۱۶	۶۷۷
بیمه شخص ثالث مالی	۵۳۹۴۵۷	۴۷/۳	۵۴۳	۳۰/۵	۱	۶۷۷
بیمه مسئولیت	۱۱۱۴	۹۳/۲	۶۰	۴۶/۹	۴	۳۶۵

منبع: یافته‌های پژوهشگر

با توجه به جدول ۱، تعداد داده‌ها در هر رشته متفاوت است؛ به‌عنوان مثال، بیمه شخص ثالث مالی با ۵۳۹۴۵۷ مشاهده، بیشترین فراوانی داده را داراست که این امر می‌تواند به دلیل پوشش گسترده‌تر این نوع بیمه باشد. در مقابل، بیمه مسئولیت تنها ۱۱۱۴ مشاهده دارد که می‌تواند به محدودیت‌های خاص این بیمه مرتبط باشد. تحلیل میانگین و میانه تأخیرها نشان‌دهنده تفاوت‌های قابل‌توجه میان رشته‌هاست. میانگین تأخیر در رشته‌هایی مانند بیمه جانی و بیمه آتش‌سوزی به دلیل ماهیت خسارت‌ها و پیچیدگی در گزارش‌دهی، بیشتر است. در تمامی رشته‌ها، میانه تأخیر کمتر از میانگین است که نشان‌دهنده توزیع غیرمتقارن و پراکندگی بالا در تأخیرها می‌باشد. بررسی انحراف معیار به عنوان شاخص پراکندگی، حاکی از آن است که بیمه آتش‌سوزی و بیمه مسئولیت با انحراف معیارهای بالا، پراکندگی بیشتری در تأخیرهای گزارش‌دهی دارند. این موضوع می‌تواند نشان‌دهنده نوسانات زیاد در تأخیر گزارش‌دهی در این رشته‌ها باشد. در نهایت، بررسی بیشترین تأخیرها نشان می‌دهد که بیمه‌های شخص ثالث مالی و جانی با بیشترین تأخیرها روبه‌رو هستند که می‌تواند ناشی از پیچیدگی در گزارش‌دهی و تایید خسارت‌ها در این رشته‌ها باشد. این تحلیل توصیفی از تأخیرها در رشته‌های مختلف می‌تواند به شرکت‌های بیمه در شناسایی الگوهای گزارش‌دهی، برنامه‌ریزی بهتر برای جبران خسارت‌ها و بهبود فرآیندهای مدیریتی کمک کند.

در نمودار ۴، تأخیر در گزارش‌دهی بیمه در پنج رشته بیمه‌ای مورد بررسی قرار گرفته است. این نمودارهای سری زمانی توزیع تأخیر گزارش‌دهی در طول زمان را نشان می‌دهند. برای بیمه‌های شخص ثالث مالی و شخص ثالث جانی و حوادث راننده، کاهش چشمگیری در تأخیرهای گزارش‌دهی در بازه زمانی دیده می‌شود که می‌تواند بهبود فرآیندهای ثبت خسارت و سرعت گزارش‌دهی را نشان دهد. در بیمه‌های آتش‌سوزی و بدنه، روند تأخیر به صورت نوسانی مشاهده می‌شود. نمودارهای هیستوگرام و چگالی نشان می‌دهند که بیشتر تأخیرها در محدوده‌های پایین قرار دارند، اما برخی تأخیرهای بالاتر نیز وجود دارد. این تأخیرهای بالا به خصوص در بیمه بدنه و بیمه آتش‌سوزی مشاهده می‌شود که می‌تواند ناشی از پیچیدگی‌ها یا شرایط خاص در پردازش این نوع بیمه‌ها باشد. در نهایت نمودارهای باکس‌پلات نشان‌دهنده میانه، چارک‌ها و نقاط دورافتاده در داده‌های تأخیر هستند. در بیمه بدنه و بیمه آتش‌سوزی، تعداد زیادی داده‌های دورافتاده وجود دارد که بیانگر تأخیرهای غیرمعمول و بیشتر از حد نرمال است. این امر می‌تواند به علت پیچیدگی‌های ثبت و پردازش خسارات در این دو رشته باشد. به طور کلی،

نتایج این نمودارها نشان‌دهنده آن است که بیشتر تأخیرها در تمامی رشته‌ها در محدوده‌های پایین قرار دارند، اما در رشته‌هایی مانند بیمه بدنه و بیمه آتش‌سوزی، توزیع تأخیرهای بالاتر و نقاط دورافتاده بیشتر مشاهده می‌شود. این یافته‌ها می‌توانند به شناسایی نقاط ضعف در فرآیندهای ثبت و رسیدگی به خسارت کمک کنند و بهبود این فرآیندها در رشته‌های پرریسک را هدف قرار دهند.



نمودار ۴- تحلیل توزیع و سری زمانی تأخیر گزارش‌دهی خسارت در رشته‌های مختلف بیمه‌ای

منبع: یافته‌های پژوهشگر

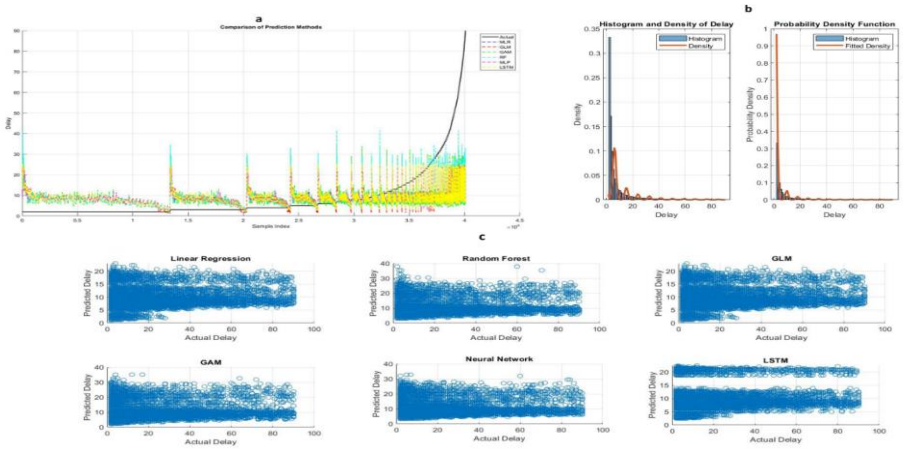
۴-۲- نتایج

در این پژوهش، ابتدا تاریخ‌های شمسی به میلادی تبدیل شدند تا تحلیل‌های زمانی ساده‌تر و با استانداردهای بین‌المللی هم‌خوان شوند. سپس، رکوردهای نامعتبر که تاریخ وقوع حادثه آنها پس از تاریخ گزارش بود، حذف

شدند. برای تمرکز بر تأخیرهای منطقی، تنها تأخیرهای ۱ تا ۹۰ روز نگه داشته شدند و داده‌های خارج از این محدوده حذف گردیدند. از تاریخ وقوع حادثه ویژگی‌هایی نظیر روز هفته، ماه، و نشانه‌های تعطیلات نوروزی و تابستانی استخراج شد تا الگوهای زمانی بهتر شناسایی شوند. داده‌های گمشده نیز حذف و مجموعه به دو بخش آموزشی (۸۰٪) و آزمایشی (۲۰٪) تقسیم شد تا مدل‌ها آموزش دیده و سپس ارزیابی شوند. این مراحل پیش‌پردازش، دقت و انسجام نتایج مدل‌سازی را تضمین می‌کنند.

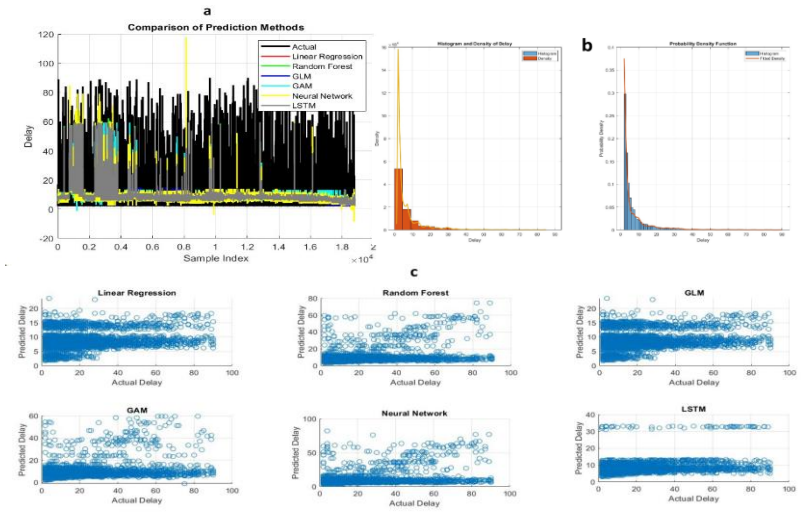
نمودار ۵، نتایج برآورد و پیش‌بینی‌های مدل‌های مختلف در رشته شخص ثالث مالی را نشان می‌دهند. در نمودار مقایسه پیش‌بینی مدل‌ها، مقادیر پیش‌بینی شده توسط هر مدل در مقایسه با مقادیر واقعی به نمایش درآمده‌اند. مشاهده می‌شود که مدل‌های مختلف تقریباً الگوهای مشابهی را دنبال کرده‌اند، اما در برخی نقاط از مقدار واقعی فاصله دارند. از میان مدل‌ها، شبکه‌های عصبی (به ویژه LSTM) و جنگل تصادفی (RF) در بعضی از بخش‌ها تطابق بهتری با مقدار واقعی دارند. اما در نقاطی، این مدل‌ها نیز انحراف‌هایی از مقدار واقعی دارند که نشان‌دهنده عدم قطعیت یا نویز در داده‌ها است. در هیستوگرام و تابع چگالی احتمالی تأخیر، توزیع داده‌های تأخیر نمایش داده شده است. بر اساس این نمودارها، اکثر داده‌ها در محدوده‌های تأخیر پایین متمرکز هستند، و تنها تعداد کمی از مشاهدات دارای تأخیرهای بالا هستند. این موضوع می‌تواند ناشی از وقوع برخی حوادث نادر با تأخیرهای زیاد باشد که می‌تواند مدل‌ها را تحت تاثیر قرار دهد و بر دقت پیش‌بینی‌ها اثر بگذارد. نمودارهای پراکندگی تأخیر واقعی در مقابل تأخیر پیش‌بینی شده نیز تطابق مقادیر پیش‌بینی شده هر مدل با مقادیر واقعی را نشان می‌دهند. مدل‌هایی که نقاط پراکندگی کمتری اطراف خط قطری دارند، دقت بالاتری دارند. به عنوان مثال، مدل LSTM و جنگل تصادفی (RF) به نظر می‌رسد که نسبت به سایر مدل‌ها تطابق بهتری دارند، هرچند هنوز برخی نقاط دارای انحراف‌های بالا هستند. مدل‌های رگرسیون خطی و GAM نیز تطابق معقولی دارند، اما انحراف‌های بیشتری در پیش‌بینی‌های آن‌ها مشاهده می‌شود. به طور کلی، مدل‌های مختلف در پیش‌بینی تأخیر در رشته شخص ثالث مالی عملکرد متفاوتی دارند. مدل‌های پیچیده‌تر مانند LSTM و جنگل تصادفی (RF) در مقایسه با مدل‌های ساده‌تر، دقت بالاتری از خود نشان داده‌اند.

همچنین، توزیع داده‌ها و نادر بودن تأخیرهای بالا می‌تواند بر دقت مدل‌ها اثر بگذارد، و نشان‌دهنده این است که داده‌های آموزش و ارزیابی ممکن است به تنظیمات بیشتری برای کاهش اثر این انحراف‌ها نیاز داشته باشند.



نمودار۵- نتایج برآورد و پیش‌بینی‌های مدل‌های مختلف در رشته شخص ثالث مالی[1]

منبع: یافته‌های پژوهشگر



نمودار۶- نتایج برآورد و پیش‌بینی‌های مدل‌های مختلف در رشته بدنه[2]

منبع: یافته‌های پژوهشگر

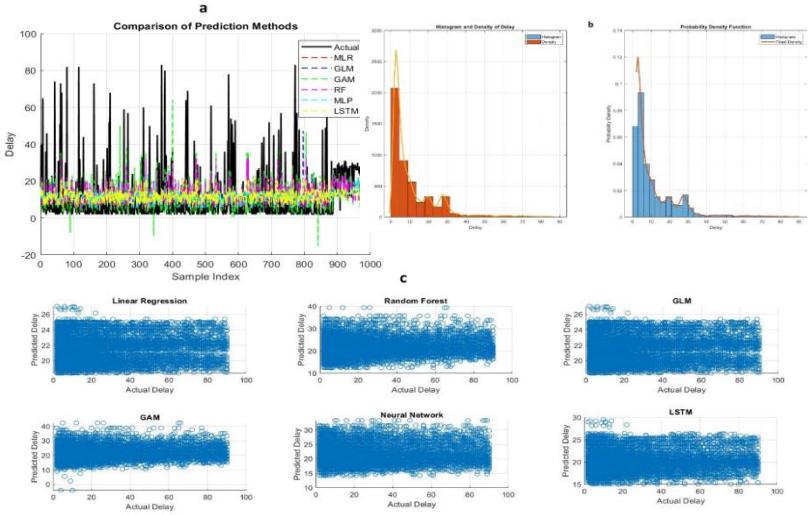
[1] نمودار a مقایسه روش‌های مختلف پیش‌بینی تأخیر، نمودار b نمودار هیستوگرام و تابع چگالی احتمال تأخیر و نمودار c مقایسه تأخیر واقعی و پیش‌بینی‌شده توسط مدل‌های مختلف است.

[2] نمودار a مقایسه روش‌های مختلف پیش‌بینی تأخیر، نمودار b نمودار هیستوگرام و تابع چگالی احتمال تأخیر و نمودار c مقایسه تأخیر واقعی و پیش‌بینی‌شده توسط مدل‌های مختلف است.

نمودار ۶، نتایج برآورد و پیش‌بینی‌های مدل‌های مختلف در رشته بدنه اتومبیل را نشان می‌دهد. در نمودار مقایسه روش‌های پیش‌بینی، مشاهده می‌شود که مقادیر پیش‌بینی شده توسط مدل‌ها در مقایسه با مقادیر واقعی الگوهای مختلفی را نشان می‌دهند. نمودارها بیانگر این موضوع هستند که جنگل تصادفی (RF) و شبکه‌های عصبی MLP و LSTM در مقایسه با مدل‌های ساده‌تر همچون رگرسیون خطی چندگانه و GLM تطابق بهتری با مقادیر واقعی دارند. با این حال، انحراف‌هایی نیز در برخی از نقاط پیش‌بینی شده مشاهده می‌شود که ممکن است به دلیل وجود نویز یا عدم دقت کافی در برخی از داده‌ها باشد. نمودارهای پراکندگی که پیش‌بینی هر مدل را در برابر تأخیر واقعی نمایش می‌دهند، عملکرد هر مدل در برآورد دقیق تأخیر را به تصویر می‌کشند. نقاطی که به خط قطری نزدیک‌تر هستند نشان‌دهنده دقت بالاتر پیش‌بینی می‌باشند. به عنوان مثال، در مدل‌های LSTM و جنگل تصادفی، پراکندگی نقاط به صورت فشرده‌تر و نزدیک‌تر به خط قطری مشاهده می‌شود که نشان‌دهنده تطابق بهتر است. مدل‌های ساده‌تر همچون رگرسیون خطی چندگانه (MLR) و GLM نقاط بیشتری با انحراف‌های بالا دارند. همچنین، هیستوگرام و تابع چگالی احتمالی تأخیر بیانگر توزیع داده‌های تأخیر در این رشته است. طبق این نمودارها، اکثر داده‌ها در بازه‌های پایین تأخیر متمرکز هستند، و تأخیرهای بالا به ندرت مشاهده می‌شود. این نکته می‌تواند دلیلی باشد بر اینکه مدل‌ها در پیش‌بینی تأخیرهای بالا دچار چالش می‌شوند، زیرا تعداد این نقاط در داده‌های آموزشی کمتر است. به طور کلی، مدل‌های پیچیده‌تر مانند LSTM و جنگل تصادفی عملکرد بهتری در پیش‌بینی تأخیر دارند و نسبت به مدل‌های ساده‌تر دقت بالاتری از خود نشان می‌دهند. اما همچنان به دلیل وجود نویز و داده‌های با تأخیر بالا که به ندرت رخ می‌دهند، عملکرد مدل‌ها در تمامی نقاط یکسان نبوده و برای بهبود پیش‌بینی‌ها می‌توان از روش‌های بیشتری برای پردازش و بهینه‌سازی داده‌ها استفاده کرد.

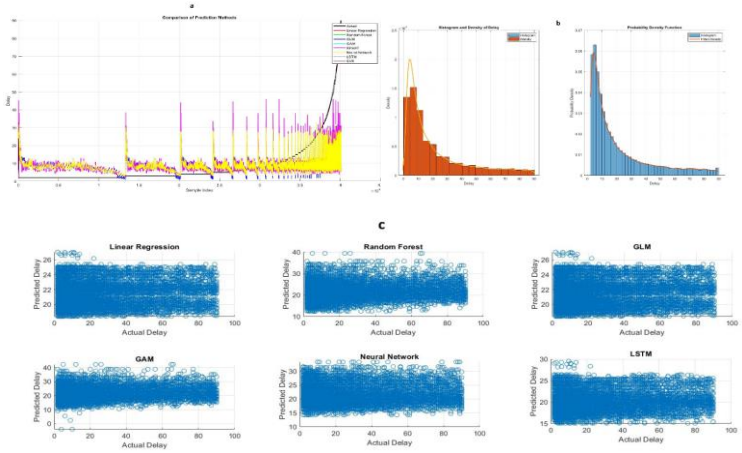
نمودار ۷، نتایج پیش‌بینی تأخیر در رشته آتش‌سوزی را با استفاده از شش مدل مختلف به نمایش می‌گذارند. نمودار (a) که به مقایسه روش‌های پیش‌بینی پرداخته است، نشان می‌دهد که مدل‌های مختلف در پیش‌بینی تأخیر مقادیر متفاوتی را نشان می‌دهند. جنگل تصادفی و مدل‌های مبتنی بر شبکه عصبی MLP و LSTM نسبت به مدل‌های ساده‌تر مانند رگرسیون خطی و GLM، تطابق بهتری با داده‌های واقعی دارند، اما همچنان مقداری انحراف از داده‌های واقعی وجود دارد. این اختلاف در برخی از نقاط نشان‌دهنده پیچیدگی داده‌ها و چالش مدل‌ها در تطبیق با تمامی جزئیات داده‌ها است. نمودارهای پراکندگی در ردیف دوم، رابطه بین تأخیر واقعی و پیش‌بینی شده برای هر مدل را نشان می‌دهند. نزدیکی نقاط به خط قطری نمایانگر دقت بهتر پیش‌بینی است. به‌طور کلی، مدل‌های شبکه عصبی و جنگل تصادفی در این نمودارها تطابق بهتری با خط قطری دارند و نشان‌دهنده توانایی آن‌ها در پیش‌بینی دقیق‌تر تأخیر در داده‌های آتش‌سوزی هستند. مدل‌های ساده‌تر مانند MLR و GLM پراکندگی بیشتری در اطراف خط قطری دارند، که نشان‌دهنده عملکرد ضعیف‌تر در برآورد دقیق تأخیر است. هیستوگرام و نمودار چگالی احتمالی تأخیر نیز نشان‌دهنده توزیع داده‌های تأخیر در این رشته هستند. بیشتر تأخیرها در بازه‌های کوتاه قرار دارند و تعداد کمی از داده‌ها دارای تأخیرهای طولانی‌تر هستند. این موضوع بیانگر این است که پیش‌بینی تأخیرهای بالاتر ممکن است برای مدل‌ها چالش‌برانگیزتر باشد. در مجموع، مدل‌های پیچیده‌تر مانند جنگل

تصادفی و شبکه‌های عصبی عملکرد بهتری در پیش‌بینی تأخیر حوادث آتش‌سوزی داشته‌اند و به عنوان ابزارهای پیش‌بینی دقیق‌تر قابل اعتمادتر هستند.



نمودار۷- نتایج برآورد و پیش‌بینی‌های مدل‌های مختلف در رشته آتش‌سوزی [۱]

منبع: یافته‌های پژوهشگر



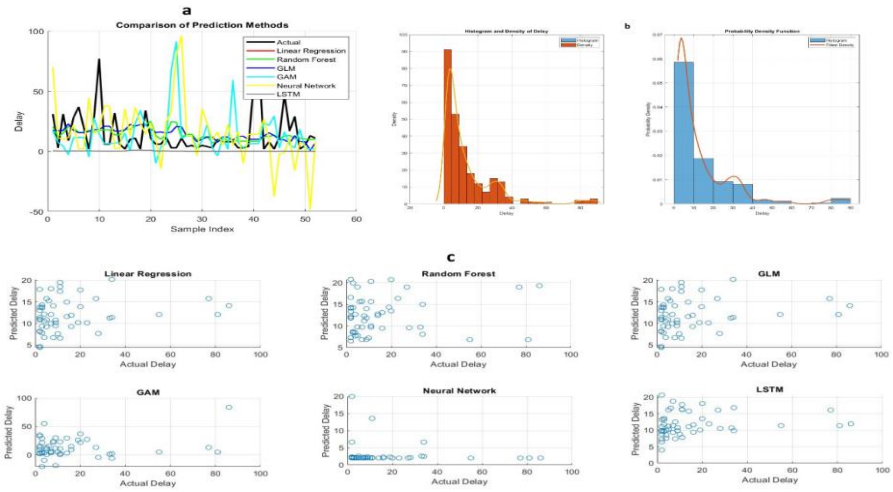
نمودار۸- نتایج برآورد و پیش‌بینی‌های مدل‌های مختلف در رشته ثالث جانی حوادث راننده [۲]

منبع: یافته‌های پژوهشگر

[۱] نمودار a مقایسه روش‌های مختلف پیش‌بینی تأخیر، نمودار b نمودار هیستوگرام و تابع چگالی احتمال تأخیر و نمودار c مقایسه تأخیر واقعی و پیش‌بینی‌شده توسط مدل‌های مختلف است.

[۲] نمودار a مقایسه روش‌های مختلف پیش‌بینی تأخیر، نمودار b نمودار هیستوگرام و تابع چگالی احتمال تأخیر و نمودار c مقایسه تأخیر واقعی و پیش‌بینی‌شده توسط مدل‌های مختلف است.

نمودار ۸، نتایج پیش‌بینی تأخیر در رشته ثالث جانی و حوادث راننده را با استفاده از چند مدل مختلف به نمایش می‌گذارند. در نمودار مقایسه روش‌های پیش‌بینی، مشاهده می‌شود که مدل‌های پیچیده‌تر مانند جنگل تصادفی و شبکه‌های عصبی در تطبیق با داده‌های واقعی عملکرد بهتری دارند. مدل‌های خطی ساده‌تر مانند MLR و GLM در مقایسه با مدل‌های غیرخطی دقت کمتری دارند و پراکندگی بیشتری نسبت به داده‌های واقعی از خود نشان می‌دهند. این امر به‌ویژه در نقاط دارای تأخیر بالا مشهود است که مدل‌های ساده‌تر از الگوی دقیق داده‌ها منحرف می‌شوند. نمودارهای پراکندگی برای هر مدل نیز به درک بهتر عملکرد آن‌ها کمک می‌کند. نقاط پراکنده در اطراف خط قطری نشان می‌دهد که برخی مدل‌ها، به ویژه مدل‌های غیرخطی مانند شبکه‌های عصبی و جنگل تصادفی، در پیش‌بینی دقیق‌تر تأخیر در بیمه ثالث جانی موفق‌تر عمل کرده‌اند. مدل GAM نیز بهبودهایی در پیش‌بینی نسبت به مدل‌های خطی از خود نشان می‌دهد، اما همچنان از نظر دقت از مدل‌های شبکه عصبی و جنگل تصادفی عقب‌تر است. در نمودارهای هیستوگرام و چگالی احتمال تأخیر، الگوی توزیع داده‌های تأخیر نمایش داده شده است. بیشتر داده‌های تأخیر در بازه‌های کوتاه قرار دارند، اما در این رشته، توزیع به‌گونه‌ای است که برخی تأخیرهای طولانی‌تر نیز به چشم می‌خورند. این نشان می‌دهد که مدل‌ها در پیش‌بینی تأخیرهای بالاتر دچار چالش بیشتری می‌شوند، به‌ویژه مدل‌های ساده‌تر که پراکندگی بیشتری در نتایج آن‌ها مشاهده می‌شود. در مجموع، مدل‌های پیچیده‌تر مانند جنگل تصادفی و شبکه‌های عصبی عملکرد بهتری در پیش‌بینی تأخیر حوادث جانی دارند و می‌توانند به عنوان ابزارهای مؤثر در تحلیل تأخیرهای بیمه ثالث جانی مورد استفاده قرار گیرند.



نمودار۹- نتایج برآورد و پیش‌بینی‌های مدل‌های مختلف در رشته مسئولیت [۱]

منبع: یافته‌های پژوهشگر

[۱] نمودار a مقایسه روش‌های مختلف پیش‌بینی تأخیر، نمودار b نمودار هیستوگرام و تابع چگالی احتمال تأخیر و نمودار c مقایسه تأخیر واقعی و پیش‌بینی‌شده توسط مدل‌های مختلف است.

نمودار ۹ نیز نشان‌دهنده عملکرد مدل‌های مختلف در پیش‌بینی تأخیر در رشته بیمه مسئولیت هستند. در نمودار مقایسه روش‌های پیش‌بینی، مشاهده می‌شود که تأخیرهای واقعی در برخی نمونه‌ها نوسانات شدیدتری دارند که مدل‌های مختلف به درجات مختلف قادر به برآورد آن‌ها هستند. مدل‌های جنگل تصادفی و شبکه‌های عصبی در بهینه‌سازی دقت در مقایسه با داده‌های واقعی عملکرد بهتری دارند و به دلیل پیچیدگی بیشتر قادر به شناسایی الگوهای غیرخطی در داده‌ها هستند. در مقابل، مدل‌های ساده‌تر نظیر MLR و GLM با دقت پایین‌تری همراه هستند و تفاوت‌های بیشتری با مقادیر واقعی نشان می‌دهند. نمودارهای پراکندگی برای هر مدل نیز به وضوح نشان می‌دهند که مدل‌های غیرخطی مانند شبکه‌های عصبی و جنگل تصادفی نسبت به مدل‌های خطی قابلیت بیشتری در به‌دست‌آوردن تطابق با داده‌های واقعی دارند. با این حال، برخی از نقاط دارای تأخیر بالا، به ویژه در مدل‌های خطی مانند MLR و GLM، از الگوهای واقعی منحرف می‌شوند و دقت کمتری را نشان می‌دهند. نمودارهای هیستوگرام و تابع چگالی احتمال توزیع تأخیرها را نشان می‌دهند که بیشتر داده‌های تأخیر در بازه‌های کوتاه قرار دارند، ولی توزیع داده‌ها به گونه‌ای است که در برخی موارد، تأخیرهای طولانی‌تر نیز مشاهده می‌شوند. این ویژگی توزیع تأخیرها نشان‌دهنده آن است که مدل‌ها باید قادر به مدیریت این نوسانات غیرعادی باشند و این امر به ویژه برای مدل‌های غیرخطی اهمیت دارد. در مجموع، مدل‌های غیرخطی مانند جنگل تصادفی و شبکه‌های عصبی عملکرد بهتری در پیش‌بینی تأخیرها در بیمه مسئولیت دارند و می‌توانند به عنوان ابزارهای مؤثری برای بهبود دقت پیش‌بینی در این حوزه به کار روند.

جدول ۲ نیز، نتایج مقایسه‌ای مدل‌های مختلف را به تفکیک رشته‌ها، همراه با مقادیر RMSE و R^2 ، نشان می‌دهد.

جدول ۲- بررسی RMSE و R^2 در مدل‌های مختلف براساس رشته‌های بیمه‌ای

رشته	روش	RMSE	R^2
ثالث مالی	رگرسیون خطی چندگانه(MLR)	۲۰/۲۲	۰/۱۴۵۲۵
	مدل خطی تعمیم‌یافته(GLM)	۱۸/۵۴۵	۰/۱۴۸۸۲
	مدل افزایشی تعمیم‌یافته(GAM)	۱۴/۰۷	۰/۳۳۲۷۴
	جنگل تصادفی(RF)	۱۱/۰۲	۰/۶۸۵۸
	شبکه عصبی پرسپترون چندلایه(MLP)	۱۲/۱۲۵	۰/۵۵۱۳
	شبکه عصبی حافظه کوتاه‌مدت بلند(LSTM)	۱۰/۰۷۲	۰/۷۱۳۳
بدنه	رگرسیون خطی چندگانه(MLR)	۴۰/۳۷	۰/۰۵
	جنگل تصادفی(RF)	۱۰/۶۴	۰/۵۴
	مدل خطی تعمیم‌یافته(GLM)	۳۸/۴۶	۰/۰۵
	مدل افزایشی تعمیم‌یافته(GAM)	۲۶/۵۵	۰/۱۲
	شبکه عصبی پرسپترون چندلایه(MLP)	۹/۸۷	۰/۶۴
	شبکه عصبی حافظه کوتاه‌مدت بلند(LSTM)	۹/۸۳	۰/۶۴
	رگرسیون خطی چندگانه(MLR)	۶۶/۵۳	۰/۰۲

رشته	روش	RMSE	R^2
ثالث جانی و حوادث راننده	جنگل تصادفی(RF)	۲۷/۳۰	۰/۱۳
	مدل خطی تعمیم‌یافته(GLM)	۳۱/۸۱	۰/۱۵
	مدل افزایشی تعمیم‌یافته(GAM)	۲۱/۳۳	۰/۲۲
	شبکه عصبی پرسپترون چندلایه(MLP)	۱۷/۱۵	۰/۴۷
	شبکه عصبی حافظه کوتاه‌مدت بلند(LSTM)	۲۲/۷۶	۰/۳۸
آتش سوزی	رگرسیون خطی چندگانه(MLR)	۷۷/۳۴	۰/۰۲
	جنگل تصادفی(RF)	۱۹/۲۶	۰/۴۸
	مدل خطی تعمیم‌یافته(GLM)	۳۱/۵۱	۰/۲۰
	مدل افزایشی تعمیم‌یافته(GAM)	۲۵/۹۲	۰/۳۱
	شبکه عصبی پرسپترون چندلایه(MLP)	۱۵/۱۵	۰/۵۵
	شبکه عصبی حافظه کوتاه‌مدت بلند(LSTM)	۱۳/۶۲	۰/۵۳
مسئولیت	رگرسیون خطی چندگانه(MLR)	۸۸/۱۰	۰/۰۱
	جنگل تصادفی(RF)	۱۸/۴۸	۰/۵۱
	مدل خطی تعمیم‌یافته(GLM)	۳۹/۴۵	۰/۲۲
	مدل افزایشی تعمیم‌یافته(GAM)	۵۵/۶۴	۰/۰۳
	شبکه عصبی پرسپترون چندلایه(MLP)	۱۸/۸۴	۰/۶۳
	شبکه عصبی حافظه کوتاه‌مدت بلند(LSTM)	۲۳/۰۴	۰/۳۸

منبع: یافته‌های پژوهشگر

جدول ۲ مقایسه عملکرد مدل‌های مختلف یادگیری ماشین در پیش‌بینی تأخیر گزارش‌دهی خسارات بیمه‌ای، بینش ارزشمندی در خصوص قابلیت و کارایی هر مدل در تحلیل داده‌های مرتبط با رشته‌های مختلف بیمه‌ای ارائه می‌دهد. نتایج این جدول بیانگر اهمیت استفاده از مدل‌های پیچیده و پیشرفته برای مواجهه با داده‌های بیمه‌ای با ویژگی‌های غیرخطی و ساختارهای پیچیده است. در رشته ثالث مالی، مدل جنگل تصادفی (RF) با مقدار RMSE برابر با ۱۱/۰۲ و ضریب تعیین R^2 برابر با ۰/۶۸۵۸ توانسته است به‌عنوان یکی از برترین مدل‌ها ظاهر شود. این مدل با بهره‌گیری از توانایی شناسایی الگوهای غیرخطی در داده‌ها، خطاهای پیش‌بینی را به حداقل رسانده و تطابق بسیار خوبی با داده‌های واقعی داشته است. علاوه بر آن، شبکه عصبی حافظه کوتاه‌مدت بلند (LSTM)با RMSE برابر با ۱۰/۰۷۲ و R^2 برابر با ۰/۷۱۳۳، بهترین عملکرد را در این رشته داشته و نشان داده است که توانایی فوق‌العاده‌ای در شناسایی روابط زمانی و پیچیده دارد. این مدل، به دلیل بهره‌گیری از معماری مبتنی بر حافظه و توانایی در پردازش داده‌های سری زمانی، عملکردی فراتر از سایر مدل‌ها ارائه داده است. از سوی دیگر، مدل‌های خطی نظیر رگرسیون خطی چندگانه (MLR) و مدل خطی تعمیم‌یافته (GLM)، به دلیل محدودیت در شناسایی روابط پیچیده، دقت بسیار پایینی داشته و خطای زیادی در پیش‌بینی‌ها نشان داده‌اند.

در رشته بدنه نیز الگوی مشابهی مشاهده می‌شود. مدل‌های شبکه عصبی پرسپترون چندلایه (MLP) و LSTM با مقادیر RMSE برابر با ۹/۸۷ و ۹/۸۳ و مقادیر R^2 برابر با ۰/۶۴ توانستند دقت بالایی در پیش‌بینی‌ها ارائه دهند. این دو مدل نشان دادند که قابلیت شناسایی الگوهای پیچیده و غیرخطی در داده‌های مرتبط با بیمه بدنه را دارند و می‌توانند به‌عنوان ابزارهای کارآمد در تحلیل داده‌های این حوزه استفاده شوند. مدل جنگل تصادفی نیز با RMSEبرابر با ۱۰/۶۴ و R^2 برابر با ۰/۵۴، عملکردی قابل‌قبول از خود نشان داد، اما همچنان نسبت به شبکه‌های عصبی ضعیف‌تر بود. مدل‌های خطی در این رشته نیز، همانند رشته ثالث مالی، عملکرد ضعیفی داشتند و نتایج ضعیف‌تری در پیش‌بینی‌ها ارائه دادند. این نتایج نشان می‌دهد که داده‌های بیمه بدنه، به دلیل ماهیت پیچیده و تنوع بالا، نیازمند استفاده از مدل‌های پیچیده‌تر هستند. در رشته ثالث جانی و حوادث راننده، شبکه عصبی MLP با RMSE برابر با ۱۷/۱۵ و R^2 برابر با ۰/۴۷ بهترین عملکرد را داشت. این مدل توانست با بهره‌گیری از ساختار غیرخطی خود، داده‌های این رشته را به‌خوبی تحلیل کند و پیش‌بینی‌های دقیقی ارائه دهد. مدل LSTM نیز با RMSEبرابر با ۲۲/۷۶ و R^2 برابر با ۰/۳۸ عملکرد قابل‌قبولی داشت، اما همچنان از مدل MLP ضعیف‌تر بود. جنگل تصادفی نیز توانست با RMSE برابر با ۲۷/۳۰ عملکردی نسبتاً مناسب ارائه دهد. مدل‌های خطی، به‌ویژه MLR، با RMSEبرابر با ۶۶/۵۳ و R^2 برابر با ۰/۰۲، نشان‌دهنده ضعف عمده خود در تحلیل داده‌های این رشته بودند. این نتایج تأکید می‌کنند که داده‌های این حوزه نیز نیازمند استفاده از مدل‌های پیشرفته‌تر برای دستیابی به دقت مطلوب هستند. در رشته آتش‌سوزی، شبکه عصبی MLP با RMSE برابر با ۱۵/۱۵ و R^2 برابر با ۰/۵۵ و مدل LSTMبا RMSE برابر با ۱۳/۶۲ و R^2 برابر با ۰/۵۳ بهترین نتایج را ارائه دادند. این دو مدل با دقت بالا و خطای کم، نشان دادند که در مواجهه با داده‌های پیچیده و متغیر این رشته، توانایی شناسایی الگوهای پنهان و پیش‌بینی دقیق را دارند. جنگل تصادفی نیز عملکرد مناسبی داشت و با RMSE برابر با ۱۹/۲۶ و R^2 برابر با ۰/۴۸، توانست به‌عنوان یکی از مدل‌های کارآمد در این رشته شناخته شود. مدل‌های خطی مانند MLR و GLM، همان‌طور که انتظار می‌رفت، عملکرد ضعیفی ارائه دادند و نتوانستند با مدل‌های پیچیده رقابت کنند. در رشته مسئولیت، شبکه عصبی MLP با RMSE برابر با ۱۸/۸۴ و R^2 برابر با ۰/۶۳ بهترین عملکرد را ارائه داد. این مدل با توانایی در شناسایی الگوهای غیرخطی، خطای پیش‌بینی را به حداقل رسانده و به‌عنوان ابزار کارآمدی در تحلیل داده‌های این رشته ظاهر شد. جنگل تصادفی نیز با RMSE برابر با ۱۸/۴۸ و R^2 برابر با ۰/۵۱ عملکرد رضایت‌بخشی داشت و نشان داد که می‌تواند در پیش‌بینی تأخیرها مؤثر باشد. مدل LSTM نیز با RMSE برابر با ۲۳/۰۴ و R^2 برابر با ۰/۳۸، اگرچه عملکردی ضعیف‌تر از سایر مدل‌های پیچیده داشت، اما همچنان بهتر از مدل‌های خطی بود. مدل‌های خطی در این رشته نیز، مشابه سایر رشته‌ها، عملکرد نامناسبی از خود نشان دادند.

به‌طور کلی، نتایج این جدول نشان می‌دهد که مدل‌های مبتنی بر شبکه عصبی و مدل‌های پیچیده‌تر مانند جنگل تصادفی، به دلیل توانایی بالا در شناسایی روابط غیرخطی و پیچیده، عملکرد بهتری در پیش‌بینی تأخیرها در رشته‌های مختلف بیمه‌ای داشتند. این مدل‌ها توانستند خطای پیش‌بینی را به‌طور چشمگیری کاهش دهند و به تصمیم‌گیری دقیق‌تر و مؤثرتر کمک کنند. در مقابل، مدل‌های خطی به دلیل محدودیت در تحلیل پیچیدگی داده‌ها، نتایج ضعیف‌تری ارائه دادند و نتوانستند به‌اندازه مدل‌های پیچیده مؤثر باشند. این نتایج تأکید می‌کند که

استفاده از مدل‌های یادگیری ماشین پیشرفته، نه تنها دقت پیش‌بینی‌ها را افزایش می‌دهد، بلکه ابزار قدرتمندی برای تحلیل داده‌های پیچیده در صنعت بیمه فراهم می‌آورد و راه را برای بهبود فرآیندهای مدیریتی و برنامه‌ریزی هموار می‌سازد.

۵- نتیجه‌گیری و پیشنهادها

این پژوهش با هدف پیش‌بینی تأخیر در گزارش‌دهی خسارات واقع‌شده اما گزارش‌نشده در پنج رشته بیمه‌ای شامل بدنه، آتش‌سوزی، ثالث مالی، ثالث جانی و مسئولیت انجام شد. داده‌ها با استفاده از روش‌های پیش‌پردازش استاندارد تحلیل شدند که این فرآیند شامل حذف رکوردهای نامعتبر، تمرکز بر تأخیرهای منطقی، حذف داده‌های گم‌شده و تبدیل تاریخ‌ها به استاندارد میلادی بود. همچنین، ویژگی‌هایی نظیر تاریخ حادثه، تأخیر در گزارش‌دهی، تعطیلات رسمی و دوره‌های خاص زمانی استخراج شد تا دقت پیش‌بینی بهبود یابد. در این مطالعه، شش مدل یادگیری ماشین شامل رگرسیون خطی چندگانه، مدل خطی تعمیم‌یافته، مدل افزایشی تعمیم‌یافته، جنگل تصادفی، شبکه عصبی پرسپترون چندلایه و شبکه عصبی حافظه کوتاه‌مدت بلند به‌کار گرفته شدند و عملکرد آن‌ها با استفاده از داده‌های آموزشی و آزمایشی ارزیابی شد. معیارهای ریشه میانگین مربعات خطا و ضریب تعیین برای مقایسه دقت مدل‌ها به‌کار رفت.

نتایج پژوهش نشان داد که مدل‌های پیچیده‌تر مانند LSTM و جنگل تصادفی عملکرد بهتری در پیش‌بینی تأخیرها داشتند. در رشته ثالث مالی، جنگل تصادفی با شناسایی روابط غیرخطی میان متغیرها و شبکه عصبی LSTMبا توانایی مدل‌سازی روابط زمانی، دقت بالایی در پیش‌بینی‌ها ارائه دادند. مدل‌های خطی در این رشته، به دلیل محدودیت در شناسایی الگوهای پیچیده، نتایج قابل قبولی نداشتند. در رشته بدنه، شبکه‌های عصبی MLP و LSTM توانستند با شناسایی الگوهای پنهان و تحلیل پیچیدگی داده‌ها، بالاترین دقت را ارائه دهند. جنگل تصادفی نیز عملکردی مناسب داشت، اما مدل‌های خطی در این رشته نیز ضعف قابل توجهی نشان دادند. در رشته ثالث جانی و حوادث راننده، شبکه عصبی MLP بهترین عملکرد را ارائه داد و توانست روابط غیرخطی را به‌خوبی شناسایی کند. مدل LSTM نیز با وجود دقت کمتر نسبت به MLP، توانست روابط زمانی موجود در داده‌ها را مدل‌سازی کند. در این رشته، مدل‌های خطی نتوانستند به‌خوبی با پیچیدگی داده‌ها سازگار شوند. در رشته آتش‌سوزی، مدل‌های MLP و LSTM به دلیل توانایی در شناسایی الگوهای پیچیده، پیش‌بینی‌های دقیقی ارائه دادند و جنگل تصادفی نیز عملکرد رضایت‌بخشی داشت، درحالی‌که مدل‌های خطی، مطابق انتظار، نتایج ضعیفی ارائه کردند. در رشته مسئولیت نیز، شبکه عصبی MLP دقیق‌ترین مدل بود و توانست خطاهای پیش‌بینی را به حداقل برساند. جنگل تصادفی نیز نتایج خوبی ارائه کرد، درحالی‌که مدل‌های خطی بار دیگر عملکرد ضعیفی نشان دادند. به‌طور کلی، این پژوهش نشان داد که مدل‌های پیچیده‌تر مانند شبکه‌های عصبی و جنگل تصادفی در پیش‌بینی تأخیر گزارش‌دهی خسارات بیمه‌ای عملکرد بهتری از خود نشان دادند که این برتری به دلایل متعددی قابل توضیح است. داده‌های بیمه‌ای معمولاً دارای روابط پیچیده و غیرخطی میان متغیرها هستند، به‌طوری که تأخیر در گزارش‌دهی ممکن است تحت تأثیر ترکیبی از عوامل مانند تاریخ حادثه، شدت خسارت، نوع بیمه‌نامه و

دوره‌های خاص زمانی نظیر تعطیلات رسمی باشد. این روابط به‌صورت ساده و خطی قابل مدل‌سازی نیستند، و اینجا است که مدل‌های پیشرفته‌ای مانند شبکه عصبی و جنگل تصادفی با توانایی خود در شناسایی الگوهای پنهان و غیرخطی، برتری پیدا می‌کنند. شبکه عصبی حافظه کوتاه‌مدت بلند به دلیل معماری خاص خود، قادر است روابط زمانی طولانی‌مدت و الگوهای متوالی در داده‌ها را شناسایی کند، که این ویژگی در تحلیل داده‌های سری زمانی بیمه‌ای مانند تأخیرهای گزارش‌دهی بسیار مهم است. این مدل توانایی درک تغییرات فصلی یا وابستگی‌های زمانی پیچیده را دارد که در داده‌های بیمه‌ای نقش بسزایی ایفا می‌کنند. جنگل تصادفی نیز با استفاده از مجموعه‌ای از درخت‌های تصمیم، می‌تواند به‌طور مؤثر تعاملات میان متغیرها را کشف کرده و تأثیر متغیرهای مختلف را بر نتایج پیش‌بینی مشخص کند. این ویژگی باعث می‌شود که جنگل تصادفی حتی در حضور داده‌های پرت یا نویز نیز عملکرد پایداری داشته باشد.

یکی دیگر از دلایل برتری این مدل‌ها، توانایی آن‌ها در تحلیل داده‌های چندبعدی و بزرگ است. داده‌های بیمه‌ای معمولاً شامل تعداد زیادی ویژگی مرتبط با زمان، نوع خسارت و رفتار بیمه‌گذاران هستند. مدل‌های خطی به دلیل محدودیت در شناسایی همبستگی‌های پیچیده میان این ویژگی‌ها، دقت پیش‌بینی پایینی دارند. در مقابل، شبکه‌های عصبی مانند LSTM و MLP با استفاده از لایه‌های متعدد و ساختارهای غیرخطی، قادر به تحلیل عمیق‌تر داده‌ها و شناسایی روابط چندبعدی هستند. جنگل تصادفی نیز با ترکیب پیش‌بینی‌های چندین درخت تصمیم، می‌تواند تأثیر متغیرهای مختلف را بهتر در نظر گرفته و نتایج بهینه‌تری ارائه دهد. علاوه بر این، مدل‌های پیچیده‌تر انعطاف‌پذیری بیشتری در مواجهه با داده‌های نویزی و مقادیر پرت دارند. جنگل تصادفی به‌طور خاص با استفاده از الگوریتم‌های ترکیبی می‌تواند اثر داده‌های غیرعادی را کاهش دهد و پیش‌بینی‌های پایدار و دقیقی ارائه دهد. شبکه‌های عصبی نیز با بهره‌گیری از فرآیند یادگیری تدریجی و تنظیم وزن‌ها، قادر به کاهش تأثیر نویز و تقویت روابط اصلی در داده‌ها هستند. این ویژگی‌ها به این مدل‌ها امکان می‌دهد تا در شرایطی که داده‌های ناقص یا پرت وجود دارند، همچنان عملکرد قابل‌اعتمادی داشته باشند. همچنین، مدل‌های پیشرفته‌تر از الگوریتم‌های بهینه‌سازی قدرتمندتری استفاده می‌کنند که به آن‌ها اجازه می‌دهد روابط پیچیده و چندلایه‌ای را بهتر یاد بگیرند. در حالی که مدل‌های خطی تنها می‌توانند روابط مستقیم میان متغیرها را مدل‌سازی کنند، مدل‌های پیچیده‌تر قادر به شناسایی الگوهای غیرمستقیم و تعاملات پیچیده میان متغیرها هستند. این توانایی برای داده‌های بیمه‌ای که به‌طور معمول شامل وابستگی‌های پیچیده هستند، بسیار مهم است. در نهایت، مدل‌های پیچیده‌تر مانند LSTMو جنگل تصادفی توانایی تعمیم‌دهی به داده‌های جدید و ناشناخته را دارند. این مدل‌ها با شناسایی الگوهای کلی در داده‌های آموزشی، می‌توانند پیش‌بینی‌های دقیقی برای داده‌های آزمایشی ارائه دهند. در مقابل، مدل‌های خطی معمولاً بیش از حد وابسته به داده‌های آموزشی هستند و نمی‌توانند به‌خوبی داده‌های جدید را تحلیل کنند. این ویژگی‌های منحصربه‌فرد مدل‌های پیچیده‌تر باعث شده است که در پیش‌بینی تأخیرهای گزارش‌دهی خسارات بیمه‌ای، عملکرد بهتری نسبت به مدل‌های خطی داشته باشند. این برتری نه‌تنها دقت پیش‌بینی‌ها را افزایش داده، بلکه به بهبود تصمیم‌گیری‌های مدیریتی و عملیاتی در صنعت بیمه نیز کمک شایانی می‌کند.

این نتایج با پژوهش‌های پیشین همخوانی دارد و یافته‌های این مطالعه با مطالعات آوانزی و همکاران (۲۰۲۴) و هیابو و همکاران (۲۰۲۴) که نقش مدل‌های ترکیبی در بهبود پیش‌بینی‌های خسارات واقع‌شده ولی گزارش‌نشده را نشان داده‌اند، سازگار است. از سوی دیگر، برخی تفاوت‌ها با تحقیقاتی نظیر مائیت (۲۰۲۳) و بودری و رابرت (۲۰۱۷) که به مدل‌های سنتی و غیرپارامتری تأکید داشته‌اند، مشاهده شد. این اختلاف‌ها ممکن است به دلیل پیچیدگی بیشتر داده‌ها و نیاز به مدل‌های پیچیده‌تر برای شناسایی الگوهای پنهان در داده‌ها باشد و بر اهمیت انتخاب مدل‌های مناسب برای شرایط خاص تأکید می‌کند. این پژوهش نشان داد که روش‌های یادگیری ماشین و مدل‌های پیچیده، به‌ویژه در داده‌های با ساختار پیچیده مانند داده‌های بیمه‌ای، می‌توانند عملکرد بهتری در پیش‌بینی خسارات واقع‌شده ولی گزارش‌نشده داشته باشند. این مدل‌ها با قابلیت شناسایی روابط غیرخطی و الگوهای زمانی پیچیده، دقت بالاتری در شرایط پیچیده و ناپایدار داده‌ها ارائه می‌دهند. همچنین، این مدل‌ها به شرکت‌های بیمه در مدیریت ریسک‌های مرتبط با تأخیر در گزارش‌دهی خسارت و تنظیم دقیق‌تر ذخایر مالی کمک می‌کنند و با افزایش دقت پیش‌بینی‌ها، می‌توانند به تصمیم‌گیری‌های بهینه‌تری منجر شوند. پیشنهاد می‌شود در پژوهش‌های آتی، داده‌های متنوع‌تر وگسترده‌تری از منابع مختلف گردآوری شود تا تعمیم‌پذیری مدل‌ها و دقت نتایج افزایش یابد. ترکیب مدل‌های مختلف یادگیری ماشین و استفاده از روش‌های هیبریدی نیز می‌تواند به بهبود دقت پیش‌بینی‌ها کمک کند و از محدودیت‌های موجود در هر مدل کاسته و به شرکت‌های بیمه امکان تصمیم‌گیری‌های دقیق‌تری را بدهد. در نهایت، این پژوهش نقشی مهم در ارتقای سطح پیش‌بینی، برنامه‌ریزی ذخایر مالی و بهبود مدیریت ریسک در صنعت بیمه دارد و به شرکت‌های بیمه کمک می‌کند تا با تکیه بر مدل‌های یادگیری ماشین، به تخصیص منابع مالی و ارزیابی دقیق‌تری از خسارت‌ها دست یابند. این پژوهش نشان داد که استفاده از مدل‌های پیشرفته یادگیری ماشین، به‌ویژه مدل‌های LSTM و جنگل تصادفی، می‌تواند دقت پیش‌بینی تأخیرهای گزارش‌دهی را در بیمه بهبود بخشد. شرکت‌های بیمه با به‌کارگیری این مدل‌ها قادر خواهند بود مدیریت ریسک و تنظیم ذخایر مالی خود را به‌طور بهینه‌تری انجام دهند. همچنین، بهبود سیستم‌های گزارش‌دهی و تسریع فرآیند انتقال اطلاعات می‌تواند دقت این پیش‌بینی‌ها را بیشتر ارتقا دهد. توجه به این نکات به ایجاد چارچوب‌های کارآمدتر در تنظیم مقررات ذخیره‌گیری و کنترل ریسک‌های مالی منجر خواهد شد و بستر مناسبی برای تصمیم‌گیری‌های استراتژیک فراهم می‌کند.

۶- منابع

بابانژاد باقری، سیده مریم، پورآقاجان، عباسعلی و عباسیان، محمد مهدی (۱۴۰۲)،.پیش بینی ارزش شرکت مبتنی بر روش‌های یادگیری عمیق، اقتصاد مالی، ۶۴(۱۷)، ۲۹۱-۳۱۸.

پورزمانی، زهرا (۱۳۹۴)، کاربرد الگوریتم ژنتیک خطی و غیرخطی در بهبود قدرت پیش‌بینی، مهندسی مالی و مدیریت اوراق بهادار، ۲۴، صص ۹۴-۸۱.

جانفشان، بیتا (۱۳۸۵)، معرفی دو روش برآورد ذخیره خسارت‌های واقع‌شده اما گزارش‌نشده، فصلنامه صنعت بیمه، ۲۱(۲)، صص ۵۰-۳۳.

شکری، افروز، ایزدی، محی‌الدین و خالدی، بهاالدین (۱۴۰۲)، مدل‌بندی خسارت‌های معوق در مثلث‌های تأخیر وابسته با در نظر گرفتن وابستگی تقویمی، پژوهشنامه بیمه، ۱۲ (۴)، صص ۲۹۸-۲۸۳.

شهبازی، کیومرث و پیله‌ور سلطان احمدی، اکبر (۱۳۹۴)، معرفی یک سیستم پیش‌بینی مناسب برای برآورد تقاضای درمان در بیمارستان امام رضا (ع) ارومیه، فصلنامه مطالعات اقتصادی کاربردی ایران، ۱۶ (۴)، صص ۲۳۲-۲۰۵.

شهریار، بهنام، امدادی، فاطمه و صیادزاده، علی (۱۳۹۴)، اندازه‌گیری ذخایر خسارت معوق به‌عنوان مهم‌ترین ذخیره فنی شرکت‌های بیمه: رویکرد توانگری .II مجموعه مقالات بیست و دومین همایش ملی و هشتمین همایش بین‌المللی بیمه و توسعه، ۲۲.

صیادی نژاد، سکینه، اسماعیل زاده مقری،علی و رستمی، محمدرضا (۱۴۰۱)، ارائه مدل پیش‌بینی بازدهی بیت‌کوین با استفاده از روش هیبریدی یادگیری عمیق - الگوریتم تجزیه سیگنال (CEEMD- DL)، اقتصاد مالی، ۶۲(۱۷)، ۲۱۷-۲۳۸.

غلامیان، الهام و داوودی، سید محمدرضا (۱۳۹۷)، پیش‌بینی روند قیمت در بازار سهام با استفاده از الگوریتم جنگل تصادفی، مهندسی مالی و مدیریت اوراق بهادار، ۹(۳۵)، صص ۳۲۲-۳۰۱.

کیایی، سید جواد و فرشادفر، زهرا (۱۴۰۳)، کاربرد مقایسه ای الگوریتم ذرات و الگوریتم ژنتیک در پیش بینی روند بلندمدت و کوتاه مدت بازده سهام، فصلنامـه اقتصاد مالی، ۳(۶۸)، صص ۳۸۳-۴۰۰.

Agresti, A. (2015), Foundations of Linear and Generalized Linear Models. Wiley.

Andersson, D. U. (2023), Optimizing method selection for IBNR-reserve calculation using machine learning. (Master's thesis, Stockholm University, Department of Mathematics, Mathematical Statistics). Stockholm, Sweden: Stockholm University.

Antonio, K., & Plat, R. (2014), Micro-level stochastic loss reserving for general insurance. Scandinavian Actuarial Journal, 649–669.

Badescu, A. L., Lin, X. S., & Tang, Q. (2016), A marked Cox model for the number of IBNR claims: Theory. Insurance: Mathematics and Economics, 69, 29–37.

Breiman, L. (2001), Random forests. Machine Learning, 45(1), 5–32.

Calcetero-Vanegas, S., Badescu, A. L., & Lin, X. S. (2023), Claim reserving via inverse probability weighting: A micro-level chain-ladder method. arXiv preprint, arXiv:2307.10808.

Chang, W. (2024), Improving the accuracy of IBNR reserve predictions using uniform splines. Scandinavian Actuarial Journal.

Dobson, A. J., & Barnett, A. G. (2018), An Introduction to Generalized Linear Models (4th ed.). Chapman and Hall/CRC.

Farkas, S., & Lopez, O. (2024), Semiparametric copula models applied to the decomposition of claim amounts. Scandinavian Actuarial Journal, (10), 1065–1092.

Goodfellow, I., Bengio, Y., & Courville, A. (2016), Deep Learning. MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009), The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

Hiabu, M., Hofman, E. D., & Pittarello, G. (2024), A machine learning approach based on survival analysis for IBNR frequencies in non-life reserving. arXiv preprint.

Hiabu, M., Mammen, E., Martínez-Miranda, M. D., & Nielsen, J. P. (2016), In-sample forecasting with local linear survival densities. Biometrika, 103(4), 843–859.

Lee, Y. K., Mammen, E., Nielsen, J. P., & Park, B. U. (2015), Asymptotics for in-sample density forecasting. The Annals of Statistics, 43(2), 620–651.

Lee, Y. K., Mammen, E., Nielsen, J. P., & Park, B. U. (2017), Operational time and in-sample density forecasting. The Annals of Statistics, 45(3), 1312–1341.

Maait, M. A. M. (2023), Estimating claims reserves in insurance industries: Evidence from the Egyptian Market. Scientific Journal for Financial and Commercial Studies and Research, Faculty of Commerce, Damietta University, 4(1), 967–984.

Mammen, E., Martínez-Miranda, M. D., Nielsen, J. P., & Vogt, M. (2021), Calendar effect and in-sample forecasting. Insurance: Mathematics and Economics, 96, 31–52.

Miranda, M. D. M., Nielsen, J. P., Sperlich, S., & Verrall, R. (2013), Continuous run-off triangles in loss reserving. ASTIN Bulletin, 43(3), 321–349.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012), Introduction to Linear Regression Analysis (5th ed.). Wiley.

Nelder, J. A., & Wedderburn, R. W. (1972), Generalized Linear Models. Journal of the Royal Statistical Society: Series A (General), 135(3), 370–384.

Smith, J., & Doe, R. (2020), Machine Learning in Insurance Claims. Journal of Insurance Studies, 45(3), 456–468.

Wood, S. N. (2017), Generalized Additive Models: An Introduction with R (2nd ed.). CRC Press.

Wuthrich, M. (2018), Neural networks applied to Chain–Ladder reserving. European Actuarial Journal, 8(2), 407–436.

Financial Economics
Vol. (19) Issue (72) October 2025

Abstract

A Comparative Machine Learning Approach for Predicting Incurred but not Reported (IBNR) Insurance Reserve Data in the Presence of Censored and Truncated Data

Akbar Pilehvar Soltanahmadi[1]
Kiumars Shahbazi[2]
Hamzeh Didar[3]

Received: 01/ July /2025 Accepted: 01/ September /2025

Abstract

This study aims to predict incurred but not reported (IBNR) reserves in various insurance lines by employing advanced machine learning models and analyzing censored and trimmed data. The dataset includes information on incident and report dates for five major insurance lines: third-party financial, vehicle, third-party bodily injury and driver accidents, fire, and liability. The methods applied in this study are Multiple Linear Regression (MLR), Generalized Linear Model (GLM), Generalized Additive Model (GAM), Random Forest (RF), Multilayer Perceptron (MLP), and Long Short-Term Memory (LSTM) networks, using data from Iran Insurance Company for the period of 2021-2022. The data were censored and trimmed based on specific periods, such as holidays, Nowruz, peak travel seasons, and construction periods, to model impactful features according to the insurance line type. Results indicate that LSTM and RF models outperform linear models in predicting delays; specifically, RF achieved errors of 10.64 and 11.02 in vehicle and third-party financial lines, while LSTM attained errors of 9.83 and 10.72, respectively. These models effectively identified complex patterns in the data, revealing that considering factors such as holidays, weekends, and data structure can help capture intricate insurance data patterns. The findings underscore that LSTM and Random Forest models significantly enhance prediction accuracy, serving as valuable tools for risk assessment and optimal reserve allocation in the insurance industry.

Keywords: Incurred But Not Reported Reserves, Random Forest, Multi-Layer Perceptron, Long Short-Term Memory

[1] Department of Economic Sciences, Faculty of Economics, Urmia University, Urmia, Iran. pilehvar.1394@gmail.com
[2]* Department of Economic Sciences, Faculty of Economics, Urmia University, Urmia, Iran. (Corresponding author), k.shahbazi@urmia.ac.ir
[3] Department of Accounting, Faculty of Economics, Urmia University, Urmia, Iran. h.didar@urmia.ac.ir