Iranian Journal of Insurance Research
(IJIR)

Homepage: https://ijir.irc.ac.ir/?lang=en

Iranian Insurance Research Center (IIRC)

ORIGINAL RESEARCH PAPER

Advanced Analysis of Transportation Data Using Machine Learning: Bill of Lading Insurance Premium Prediction

Zeynab Jafarzadeh[1], Narges Mirehi[2,*], Efat Golpar Raboky [3]

[1] *B.Sc. Student, Department of Computer Science, Faculty of Basic Sciences, University of Qom, Qom, Iran.*

[2] *Assistant Professor, Department of Computer Sciences, Faculty of Sciences, University of Qom, Qom, Iran.*

[3] *Associate Professor, Department of Mathematical Sciences, Faculty of Sciences, University of Qom, Qom, Iran.*

ARTICLE INFO

Article History:
Received: 6 November 2025
Final revision: 6 February 2026
Accepted: 24 February 2026
Early online access: 24 February 2026
Published: 1 July 2026

Keywords:
Anomaly detection
Financial risk management
Insurance premium prediction
CatBoost model
Intelligent logistics

ABSTRACT

BACKGROUND AND OBJECTIVES: Intelligent transportation and logistics management are critical components of modern supply chains, directly influencing cost reduction and risk management. With the rapid increase in transportation data, the need for data-driven approaches to financial decision-making has become increasingly evident. One of the main challenges in this field is the non-linear nature of relationships in premium calculations, which traditional models fail to adequately capture. This research aims to transition from conventional methods to an intelligent insurance pricing system in Qom Province, Iran—recognized as a major transit hub and the logistics center of the country. The significance of this study lies in its ability to uncover non-linear patterns within large-scale datasets that traditional models are unable to analyze effectively. The operational benefits of this system go beyond mere numerical predictions; unlike them, it includes real-time financial risk management, increased tariff transparency, and the intelligent detection of anomalies in the bill of lading issuance process. These features contribute to optimized budgeting and improved decision-making accuracy in this critical transportation bottleneck.

METHODS: This data-driven research focuses on over 600,000 bills of lading records from Qom Province. In the first stage, the data underwent thorough cleaning and outlier management, followed by the identification of key variables that influence insurance premiums such as cargo weight, transportation distance, and product type. Due to the high complexity and non-linear relationships among these variables, traditional linear models were unable to capture the underlying pricing patterns. To address this limitation, advanced algorithms including Symmetric Gradient Boosting (Categorical Boosting) and Extreme Gradient Boosting (XGBoost) were deployed. The Symmetric Gradient Boosting model was chosen as the core algorithm due to its superior ability to handle categorical variables (such as cargo type) and its robust performance in preventing overfitting. In the final stage, the models' performance was assessed using statistical metrics, including the Coefficient of Determination (R2) and Median Absolute Error (MedAE), to evaluate the precision of the predicted values.

FINDINGS: The analysis revealed that intelligent algorithms significantly improve the accuracy of insurance premium calculations. The primary innovation of this research is the identification of hidden non-linear and multi-dimensional patterns within large-scale logistics data. The Symmetric Gradient Boosting model achieved an R2 score of 0.86, accurately reconstructing market fluctuations with high precision. This result suggests that the model can replace traditional methods, which, due to their rigid structure, lack the flexibility to adapt to the complexities of the logistics industry. By leveraging advanced machine learning techniques, this study demonstrated that insurance pricing is governed by non-linear relationships rather than simple linear dependencies. Furthermore, the research showed that qualitative factors such as cargo type and the identity of the transportation companies play a far more significant role in financial risk management than before. Traditional models typically placed heavy emphasis on quantitative variables such as distance, but this study demonstrated that categorical variables could influence pricing much more significantly.

CONCLUSION: A key strategic takeaway from this study is the identification of cargo nature and the identity of transportation companies as the primary factors influencing insurance risk assessment. Unlike traditional methods, which predominantly focus on distance, the findings from this research suggest that qualitative parameters such as the type of goods being transported and the reliability of the transport company play a far more substantial role in managing financial risk. The intelligent architecture developed in this study minimizes prediction errors, making it possible to implement dynamic and fair pricing mechanisms. This not only ensures the protection of the interests of insurance companies but also safeguards the rights of cargo owners, thereby creating a more balanced and equitable system for all stakeholders. By moving beyond outdated pricing models, this intelligent system provides a foundation for a more transparent, efficient, and responsive insurance pricing model in the transportation and logistics sector. Moreover, the proposed system offers real-time capabilities for financial risk management and anomaly detection, enhancing operational efficiency and reducing human error. This is particularly crucial in sectors where financial decision-making is complex and data-heavy, such as logistics and transportation. The system also facilitates the detection of potential discrepancies in bill of lading issuance, thereby reducing fraudulent activities and ensuring compliance with industry standards. These capabilities can have a far-reaching impact on the logistics sector, improving the transparency and accountability of transactions while also supporting better budgeting and forecasting.

*Corresponding Author:
Email: *N.mirehi@qom.ac.ir*
Phone: +98 2532103855
ORCID: 0009-0005-4110-8899

DOI: 10.22056/ijir.2026.03.02

نشریه علمی پژوهشنامه بیمه

سایت نشریه: https://ijir.irc.ac.ir/?lang=fa

مقاله پژوهشی

تحلیل پیشرفتهٔ داده‌های حمل‌ونقل با استفاده از یادگیری ماشین: پیش‌بینی مبلغ بیمهٔ بارنامه

زینب جعفرزاده[۱]، نرگس میره‌ای[۲,*]، عفت گلپر رابوکی[۳]

[۱] دانشجوی کارشناسی، گروه علوم کامپیوتر، دانشکده علوم پایه، دانشگاه قم، قم، ایران.

[۲] استادیار، گروه علوم کامپیوتر، دانشکده علوم پایه، دانشگاه قم، قم، ایران.

[۳] دانشیار، گروه علوم ریاضی، دانشکده علوم پایه، دانشگاه قم، قم، ایران.

اطلاعات مقاله

تاریخ های مقاله:

دریافت: ۱۵ آبان ۱۴۰۴

بازنگری نهایی: ۱۷ بهمن ۱۴۰۴

پذیرش: ۵ اسفند ۱۴۰۴

زودآیند: ۵ اسفند ۱۴۰۴

انتشار: ۱۰ تیر ۱۴۰۵

کلمات کلیدی:

پیش‌بینی مبلغ بیمه

شناسایی ناهنجاری

لجستیک هوشمند

مدل CatBoost

مدیریت ریسک مالی

***نویسنده مسئول:**

ایمیل: N.mirehi@qom.ac.ir

تلفن: ۲۵۳۲۱۰۳۸۵۵ ۹۸+

ORCID: 0009-0005-4110-8899

DOI: 10.22056/ijir.2026.03.02

چکیده:

پیشینه و اهداف: مدیریت هوشمند حمل‌ونقل و لجستیک یکی از ارکان حیاتی در زنجیرهٔ تأمین مدرن است که بر کاهش هزینه‌ها و مدیریت ریسک تأثیر مستقیم دارد. با افزایش حجم داده‌های حمل‌ونقل، ضرورت استفاده از روش‌های داده‌محور برای تصمیم‌گیری‌های مالی بیش‌ازپیش احساس می‌شود. چالش اصلی در این حوزه، وجود روابط غیرخطی در تعیین مبلغ بیمهٔ بارنامه‌هاست که مدل‌های سنتی قادر به کشف آن‌ها نیستند. این پژوهش با هدف گذار از محاسبات سنتی به‌سمت نظام هوشمند قیمت‌گذاری بیمه در استان قم — به‌عنوان شاهراه ترانزیتی و قلب لجستیک ایران — انجام شده است. ضرورت این مطالعه در شناسایی روابط غیرخطی داده‌های حجیم نهفته است که مدل‌های کلاسیک در تحلیل آن‌ها ناتوان‌اند. دستاورد عملیاتی این معماری، فراتر از پیش‌بینی عددی، شامل مدیریت آنی ریسک مالی، شفاف‌سازی تعرفه‌ها و شناسایی هوشمند تخلفات در صدور بارنامه است که موجب بهینه‌سازی بودجه‌بندی و افزایش دقت تصمیم‌گیری در این گلوگاه حیاتی حمل‌ونقل می‌شود.

روش‌شناسی: پژوهش حاضر با رویکردی داده‌محور بر بیش از ۶۰۰ هزار رکورد از بارنامه‌های استان قم به‌عنوان شاهراه ترانزیتی کشور تمرکز دارد. در گام نخست، پس از پالایش داده‌ها و مدیریت مقادیر پرت، متغیرهای اثرگذار همچون وزن، مسافت و نوع کالا شناسایی شدند. به‌دلیل پیچیدگی بالا و ماهیت غیرخطی روابط میان این متغیرها، روش‌های سنتی خطی کارایی لازم را نداشتند و نتوانستند الگوهای حاکم بر قیمت‌گذاری را کشف کنند. ازاین‌رو، از الگوریتم‌های پیشرفتهٔ «تقویت گرادیان متقارن» و «تقویت گرادیان شدید» استفاده شد. مدل متقارن به‌دلیل توانایی در پردازش مستقیم متغیرهای کیفی (مانند نوع محموله) و جلوگیری از خطای انطباق بیش از حد، به‌عنوان الگوریتم محوری برگزیده شد. در نهایت، عملکرد مدل‌ها با شاخص‌های آماری همچون ضریب تعیین و میانهٔ قدرمطلق خطا ارزیابی شد.

یافته‌ها: تحلیل خروجی‌ها نشان داد که استفاده از الگوریتم‌های هوشمند، تحولی اساسی در دقت محاسبات بیمه‌ای ایجاد کرده است. نوآوری اصلی این پژوهش در شناسایی الگوهای غیرخطی و چندوجهی است که پیش از این در لایه‌های پنهان داده‌های حجیم بارنامه‌ها مخفی مانده بود. مدل «تقویت گرادیان متقارن» با دستیابی به ضریب تعیین ۰٫۸۶، توانست با دقت بسیار بالایی نوسانات واقعی بازار را بازسازی کند. این نتایج نشان‌دهندهٔ ظرفیت بالای مدل برای بهبود و ارتقای روش‌های سنتی است؛ روش‌هایی که به‌دلیل ساختار نسبتاً صلب خود، توانایی محدودی در انطباق با پیچیدگی‌های لجستیکی دارند.

نتیجه‌گیری: دستاورد راهبردی این پژوهش شناسایی ماهیت محموله و هویت شرکت‌های حمل‌ونقل به‌مثابهٔ ستون‌های اصلی تعیین ریسک است. برخلاف تصورات سنتی که فقط بر مسافت تمرکز داشتند، این مدل ثابت کرد که پارامترهای کیفی، نقش بسیار مهمتری در مدیریت ریسک مالی ایفا می‌کنند. نتایج نشان داد که این معماری هوشمند می‌تواند خطای پیش‌بینی را به حداقل برساند و بستری برای قیمت‌گذاری پویا و عادلانه فراهم کند که علاوه‌بر حفظ منافع شرکت‌های بیمه، از تضییع حقوق صاحبان کالا نیز جلوگیری می‌نماید.

توجه: مدت‌زمان بحث و انتقاد برای این مقاله تا ۱ اکتبر ۲۰۲۶ در وب‌سایت IJIR در «نمایش مقاله» باز است.

مقدمه

مدیریت هوشمند حمل‌ونقل و لجستیک به‌مثابهٔ یکی از ارکان اساسی و حیاتی در زنجیرهٔ تأمین مدرن، تأثیرات بسیاری در بهینه‌سازی هزینه‌ها و مدیریت ریسک دارد. این امر به‌ویژه در دنیای پیچیده و به‌سرعت در حال تغییر امروز اهمیت دارد، جایی که رقابت در بازارهای جهانی و فشار بر افزایش کارایی و کاهش هزینه‌ها، شرکت‌ها و سازمان‌ها را بر آن می‌دارد تا در پی روش‌های نوآورانه و داده‌محور باشند (بزی و رشیدی محمودی، ۱۳۹۸). در این راستا، استفاده از فنّاوری‌های پیشرفته برای تحلیل داده‌ها و مدیریت حمل‌ونقل به یک نیاز اجتناب‌ناپذیر تبدیل شده است، زیرا این فنّاوری‌ها می‌توانند به بهبود پیش‌بینی‌های مالی، به‌ویژه در زمینهٔ پیش‌بینی هزینه‌های بیمهٔ بارنامه‌ها، کمک کنند.

رشد سریع مبادلات کالا در سطح جهانی و ظهور داده‌های کلان (Big Data) تغییرات عمده‌ای در صنعت حمل‌ونقل به وجود آورده است. این تحولات به‌ویژه در روند پیش‌بینی‌های مالی و بیمه‌گذاری تأثیرگذار بوده است. مدل‌های سنتی که براساس داده‌های خطی و روابط ساده طراحی شده‌اند، به‌طور فزاینده‌ای ناکارآمد شده‌اند و توانایی تحلیل پیچیدگی‌های موجود در روابط میان متغیرهای مختلف حمل‌ونقل را ندارند. این چالش‌ها شامل پیچیدگی‌های غیرخطی در تعیین هزینه‌های بیمه، تأثیر عوامل متنوعی مانند نوع محموله، مسافت، شرایط جوی و ویژگی‌های حمل‌کننده است که مدل‌های سنتی قادر به شبیه‌سازی آن‌ها نیستند (پرستش و همکاران، ۱۴۰۴). در نتیجه، مدل‌های هوشمند مبتنی بر یادگیری ماشین و الگوریتم‌های پیشرفته به‌ویژه در این حوزه‌ها از اهمیت ویژه‌ای یافته‌اند.

استان قم به‌مثابهٔ شاهراهی ترانزیتی و نقطه‌ای حیاتی در حمل‌ونقل کالاهای کشور، موقعیتی استراتژیکی دارد که آن را به گلوگاه ارتباطی میان بنادر جنوبی و مراکز صنعتی شمال و غرب ایران تبدیل کرده است. حجم بالای بارهای عبوری از این استان، داده‌های حمل‌ونقل آن را به یک «مجموعه دادهٔ مرجع» تبدیل کرده است که می‌توان از آن برای تحلیل دقیق و پیش‌بینی‌های بهینه در سطح ملی استفاده کرد. در این زمینه، مدل‌های هوشمند می‌توانند با تحلیل داده‌های پیچیده، به شرکت‌های بیمه و حمل‌ونقل کمک کنند تا فرایندهای قیمت‌گذاری بیمه را دقیق‌تر و شفاف‌تر انجام دهند و از این طریق خطرات مالی را مدیریت کنند (حسینی و صیفی‌زاده، ۱۴۰۲).

در این پژوهش، هدف این است که با استفاده از مدل‌های هوشمند، مانند الگوریتم‌های تقویت گرادیان متقارن (CatBoost) و یادگیری عمیق (Deep Learning)، الگوهای پیچیده و غیرخطی موجود در داده‌های حمل‌ونقل کشف شود. این مدل‌ها به‌ویژه برای تحلیل داده‌های چندوجهی و پیچیده مناسب‌اند و می‌توانند روابط میان متغیرهای مختلفی همچون نوع محموله، وزن، مسافت و ویژگی‌های حمل‌کننده را هم‌زمان بررسی کنند (Duan et al., 2025). این مدل‌ها با پردازش حجم عظیم داده‌ها، قادرند دقت پیش‌بینی‌ها را به‌طور چشمگیری افزایش دهند و در نتیجه به بهبود عملکرد مالی و مدیریتی در صنعت حمل‌ونقل کمک کنند.

برای انجام این تحقیق، بیش از ۶۰۰ هزار رکورد واقعی از بارنامه‌های استان قم تحلیل شد. هدف این تحقیق، ارائهٔ روشی است که با استفاده از الگوریتم‌های هوشمند، بتوان خطای محاسبات بیمه را به حداقل رساند و بستری مناسب برای قیمت‌گذاری پویا و عادلانه فراهم کرد. این مدل‌ها نه‌تنها به بهبود فرایند پیش‌بینی هزینه‌ها کمک خواهند کرد، بلکه خواهند توانست با شناسایی به‌موقع ناهنجاری‌ها و تخلفات، از وقوع خطرات مالی جلوگیری کنند و به شفافیت و اعتبار بیشتر صنعت حمل‌ونقل و بیمه‌گذاری کمک کنند

بنابراین، پرسش اصلی پژوهش این است که چگونه می‌توان با استفاده از مدل‌های هوشمند و تحلیل داده‌های حجیم، به شفافیت و دقت بیشتر در محاسبهٔ هزینه‌های بیمهٔ بارنامه‌ها دست یافت و از روش‌های سنتی فاصله گرفت؟ این سؤال، چالش اصلی پژوهش حاضر است که می‌تواند به تحولی اساسی در شیوه‌های مدیریت ریسک مالی و قیمت‌گذاری بیمه در صنعت حمل‌ونقل منجر شود.

مبانی نظری پژوهش

پژوهش پیش رو بر چهارچوب مفهومی دو حوزهٔ مجزا اما درهم‌تنیده بنا شده است: علم آمار بیمه و نظریه‌های قیمت‌گذاری ریسک و نظریه‌های یادگیری ماشین برای مدل‌سازی روابط غیرخطی. هدف از این بخش، توجیه نظری متغیرهای منتخب و روش‌شناسی محاسباتی مورد استفاده است.

از دیدگاه آمار بیمه، حق‌بیمه برای هر قرارداد (در اینجا بارنامه)، باید براساس اصل انصاف تعیین شود؛ بدین معنا که حق‌بیمهٔ دریافتی باید متناسب با ریسک مورد انتظار بیمه‌گذار باشد. قیمت‌گذاری حق‌بیمه به‌طور سنتی براساس یک معادلهٔ اساسی تعیین می‌شود:

حق‌بیمه = ریسک مورد انتظار (خسارت) + بار هزینه‌ای (Expenses Load) + حاشیه سود

مدل‌سازی این ریسک در حمل‌ونقل کالا بر محور ناهمگونی ریسک (Risk Heterogeneity) تمرکز دارد. متغیرهای تأثیرگذار بر ریسک بیمه بارنامه‌ها به دو دسته اصلی تقسیم می‌شوند که در مدل‌های آماری ما به‌عنوان ویژگی‌های ورودی (Features) تعریف شده‌اند:

۱. ریسک متغیر (Variable Risk): شامل مؤلفه‌هایی است که مستقیماً احتمال یا شدت وقوع خسارت را تغییر می‌دهند. از دیدگاه نظری، «وزن کالا» و «مسافت پیمایش» متغیرهای اصلی هستند، زیرا با افزایش این دو پارامتر، طبق نظریه‌های احتمالات، شدت خسارت و احتمال وقوع حوادث به‌صورت غیرخطی افزایش می‌یابد. شناسایی این رفتار غیرخطی، نقطهٔ عطف استفاده از یادگیری ماشین در این پژوهش است.

۲. بار هزینه‌ای (Expense Load): شامل مؤلفه‌های اجرایی و مالی است که اگرچه ذاتاً ریسک تصادفی ندارند، اما بخشی جدایی‌ناپذیر از ساختار تعرفه‌گذاری هستند. متغیرهایی چون «کارمزد شرکت حمل‌ونقل» و «مالیات بر ارزش افزوده» در این دسته قرار می‌گیرند. این پارامترها در مدل به‌عنوان فاکتورهای مالی اثرگذار

در قیمت نهایی لحاظ شده‌اند.

مدل‌های سنتی (مانند رگرسیون خطی) با فرض وجود رابطهٔ مستقیم و هموار میان متغیرها طراحی شده‌اند. اما در کلان‌داده‌های لجستیکی، روابط غالباً **بسیار غیرخطی و مرکب**اند. مثلاً، اثر افزایش وزن بر مبلغ بیمه در مسافت‌های کوتاه با اثر آن در مسافت‌های طولانی یکسان نیست. مدل‌های خطی در شناسایی این «نقاط شکست» و «تعاملات متقاطع» ناتوان‌اند، که همین امر به خطای بالا در تخمین مبالغ بیمه منجر می‌شود. برای غلبه بر نارسایی‌های یادشده، این پژوهش بر نظریهٔ «یادگیری جمعی» تمرکز دارد که هدف آن ترکیب چندین مدل ضعیف برای دستیابی به یک پیش‌بینی‌گر قدرتمند است.

- **الف) نظریه تقویت گرادیان متقارن (CatBoost):** این مدل که هستهٔ اصلی پژوهش حاضر را تشکیل می‌دهد، بر پایهٔ نظریهٔ «درختان تصمیم متقارن» بنا شده است. برخلاف مدل‌های دیگر، این رویکرد نوآورانه می‌تواند **متغیرهای کیفی** (مانند نام شهر یا نوع کالا) را بدون تبدیل‌های عددیِ آسیب‌زننده، به طور مستقیم پردازش کند. این ویژگی از دیدگاه نظری، مانع از دست رفتن اطلاعات در فرایند کدگذاری می‌شود و دقت مدل را در مواجهه با داده‌های دسته‌بندی‌شده به‌شدت افزایش می‌دهد.
- **ب) نظریهٔ تقویت گرادیان شدید (XGBoost):** این مدل از طریق «تخمین مرتبهٔ دوم تابع هزینه» و استفاده از مکانیسم‌های جریمه‌گر، مانع از انطباق بیش از حد بر داده‌های آموزشی می‌شود. این ویژگی نظری، تضمین‌کنندهٔ «تعمیم‌پذیری» مدل است؛ یعنی مدل نه‌تنها بر داده‌های گذشته، بلکه بر بارنامه‌های جدید نیز با پایداری بالا عمل می‌کند.

در نهایت، نظریهٔ «مدیریت مبتنی بر شواهد» بیان می‌دارد که تحلیل اهمیت ویژگی‌ها صرفاً یک خروجی فنی نیست، بلکه ابزاری برای شناسایی **پیشران‌های اصلی ریسک** است. از طریق این تحلیل، مدیران لجستیک می‌توانند دریابند که کدام متغیرها (مثلاً نوع محموله در مقابل وزن) بیشترین سهم را در نوسانات قیمتی دارند و بر همان اساس، سیاست‌های کنترلی خود را تنظیم کنند.

مروری بر پیشینهٔ پژوهش

در دهه‌های اخیر، تحلیل داده‌های حمل‌ونقل و شناسایی الگوهای رفتاری در سیستم‌های لجستیکی به یکی از محورهای اصلی پژوهش در حوزهٔ مدیریت زنجیرهٔ تأمین تبدیل شده است. پژوهشگران متعددی در سطح بین‌المللی و داخلی به بررسی راهکارهای هوشمند برای پیش‌بینی، بهینه‌سازی و تشخیص ناهنجاری‌ها در داده‌های حمل‌ونقل پرداخته‌اند.

در میان پژوهش‌های داخلی، بزی و رشیدی محمودی (۱۳۹۸) از الگوریتم‌های یادگیری ماشین برای شناسایی ناهنجاری در داده‌های ترافیکی شهری استفاده کردند. هرچند نتایج پژوهش نشان‌دهندهٔ اثربخشی روش‌های هوشمند بود، اما داده‌های مورد استفاده محدود به حوزهٔ درون‌شهری بودند و ابعاد اقتصادی و کرایه‌ای حمل‌ونقل در آن لحاظ نشد. ژانگ و همکاران (Zhang et al., 2021) از مدلی مبتنی بر شبکهٔ فضایی برای تحلیل مسیر حرکتی وسایل نقلیهٔ باری استفاده کردند. نتایج آن‌ها نشان داد که الگوهای غیرعادی حرکتی می‌توانند به‌طور مستقیم به بروز تأخیر در تحویل یا افزایش هزینه‌ها منجر شوند. بااین‌حال، تمرکز این پژوهش عمدتاً بر تحلیل مکانی بود و عواملی چون وزن یا نوع کالا توجه محدودی دریافت کردند. در همین سال، داراوژو و اعتماد (De Araujo & Etemad, 2021) با بهره‌گیری از شبکه‌های عصبی عمیق، زمان تحویل مرسولات را در بستر شهر هوشمند پیش‌بینی کردند. اگرچه مدل آن‌ها دقت بالایی داشت، اما تمرکز اصلی بر الگوریتم بود و از ابزارهای هوش تجاری برای مصورسازی و پشتیبانی از تصمیم‌گیری مدیریتی استفاده نشد. گنگ و همکاران (Gang et al., 2023) مدلی برای تشخیص ناهنجاری در اسناد حمل‌ونقل شبکه‌ای ارائه کردند که با استفاده از رویکرد استنتاج بیزین چندگانه توانست دقت بالایی در شناسایی داده‌های پرت به‌دست آورد. بااین‌حال، پژوهش آن‌ها عمدتاً بر داده‌های شبیه‌سازی‌شده متکی بود و فاقد تحلیل بصری تعاملی برای تصمیم‌گیری مدیریتی محسوب می‌شد.

در سطح پژوهش‌های داخلی، حسینی و صیفی‌زاده (۱۴۰۲) الگوریتم‌های پیش‌بینی قیمت حمل را با تمرکز بر ویژگی‌های وزنی و مسافتی بررسی کردند. اگرچه مدل‌های پیشنهادی عملکرد مناسبی در پیش‌بینی کرایه داشتند، اما تحلیل تعاملی و مصورسازی داده‌ها برای پشتیبانی از تصمیم‌گیری مدیریتی در این پژوهش غایب بود. در پژوهش‌های جدیدتر، روکاس و همکاران (Rokoss et al., 2024) از الگوریتم‌های یادگیری ماشین برای پیش‌بینی زمان تحویل بار در محیط‌های تولیدی استفاده کردند. نتایج آن‌ها نشان داد که متغیرهایی مانند مسافت و فصل تأثیر معناداری بر زمان تحویل دارند، اما به‌دلیل محدود بودن داده‌ها، قابلیت تعمیم نتایج کاهش یافته بود. همچنین، آگراوال و همکاران (Agrawal et al., 2024) بر اهمیت رویکردهای داده‌محور در تصمیم‌گیری‌های لجستیکی و بهینه‌سازی زنجیرهٔ تأمین تأکید کردند و نشان دادند که ترکیب الگوریتم‌های یادگیری ماشین با ابزارهای هوش تجاری می‌تواند به تصمیم‌سازی هوشمندتر منجر شود. در نهایت، دوآن و همکاران (Duan et al., 2025) با استفاده از مدل‌های ترکیبی یادگیری عمیق و بهینه‌سازی، به تحلیل داده‌های ترافیکی پرداختند و بر اهمیت استفاده از داده‌های چندمنظوره شامل اطلاعات مکانی، آب‌وهوایی و اقتصادی تأکید کردند. در همین راستا، پرستش و همکاران (۱۴۰۴) در پژوهش خود با بهره‌گیری از مدل‌های رگرسیون عمیق ترتیبی مبتنی بر شبکه‌های LSTM، توانستند روابط زمانی و الگوهای پیچیده در داده‌های بیمه‌ای را مدل‌سازی کنند و دقت پیش‌بینی ریسک و تعیین نرخ بهینه بیمه‌نامه‌ها را به‌صورت معناداری بهبود دهند.

روش‌شناسی پژوهش

پژوهش حاضر با رویکردی مستخرج از فرایند استاندارد صنعتی برای داده‌کاوی (CRISP-DM) و بر پایهٔ یادگیری ماشین نظارت‌شده

انجام شده است. مراحل دقیق اجرایی به شرح زیر است:

گردآوری و درک داده‌ها (Data Acquisition)

داده‌های این پژوهش شامل ۶۰۰۰۰۰ **رکورد** واقعی از سامانهٔ جامع بارنامه‌های استان قم در سال ۱۴۰۳ است. این مجموعه داده به‌دلیل موقعیت جغرافیایی قم به‌عنوان شاهراه مواصلاتی، طیف وسیعی از انواع کالاها، مقاصد جغرافیایی (۳۱ استان کشور) و شرکت‌های حمل‌ونقل متنوع را پوشش می‌دهد. هر رکورد شامل ۲۰ ویژگی اولیه بود که ابعاد مختلف یک عملیات لجستیکی (فیزیکی، زمانی و مالی) را توصیف می‌کرد.

پیش‌پردازش و مهندسی ویژگی‌های پیشرفته

به‌منظور ارتقای نرخ سیگنال به نویز و پاسخ به چالش‌های آماری، فرایندهای زیر در محیط پایتون پیاده‌سازی شد:

- **حذف نشت داده (Data Leakage)**: در این بخش، شناسایی و حذف متغیرهایی مانند «مالیات بر ارزش افزوده» و «کرایهٔ کل» انجام شد. این متغیرها خود تابعی از مبلغ بیمه هستند و حضور آن‌ها در زمان آموزش باعث ایجاد دقت کاذب می‌شود. حذف این موارد، مدل را برای پیش‌بینی واقعی «پیش از صدور بارنامه» آماده کرد.
- **مدیریت مقادیر گمشده:** برای اطمینان از کیفیت داده‌ها و جلوگیری از اثرگذاری مقادیر گمشده یا نامعتبر بر مدل پیش‌بینی، ابتدا رکوردهایی که مقدار بیمهٔ آن‌ها ثبت نشده بود، حذف شدند. سپس در ستون‌های عددی، مقادیر گمشده با میانهٔ همان ستون جایگزین شد تا تأثیر مقادیر پرت و نبود داده کاهش یابد. در ستون‌های متنی نیز مقادیر گمشده با عبارت «نامشخص» پر شدند. علاوه‌براین، مقادیر بی‌نهایت و منفی بی‌نهایت پیش از جایگزینی، به مقادیر گمشده تبدیل شدند تا فرایند پاک‌سازی کامل و سازگار باشد
- **استراتژی پالایش ۱۰-۹۰:** برای مدیریت داده‌های پرت، از روش حذف براساس چارک‌های آماری استفاده شد. فقط رکوردهایی که مبلغ بیمهٔ آن‌ها در بازهٔ ۱۰ تا ۹۰ درصد توزیع قرار داشت، ابقا شدند تا مدل از الگوهای متعارف بازار تبعیت کند و تحت تأثیر خطاهای ثبتی مبالغ بسیار ناچیز یا نجومی قرار نگیرد.
- **تبدیل لگاریتمی متغیر هدف:** با توجه به چولگی مثبت (Right-Skewed) در توزیع مبالغ بیمه، از تابع log1p استفاده شد. این تبدیل ریاضی باعث تقارن در توزیع خطا شد و حساسیت مدل را نسبت به تفاوت‌های مقیاسی در مبالغ مختلف بهبود بخشید.

$$\log1p(x) = \ln(1 + x)$$

- **خلق ویژگی‌های جدید (Feature Creation)**: ویژگی «نسبت وزن به مسافت» و «لگاریتم وزن» تولید شدند تا روابط غیرخطی میان فیزیک بار و نرخ ریسک، با وضوح بیشتری برای الگوریتم قابل شناسایی باشد.

معماری مدل پیشنهادی و تنظیم ابرپارامترها

در مراحل اولیهٔ پژوهش، مدل‌های خطی سنتی مانند رگرسیون خطی و مدل‌های بیمسنجی نیز بررسی شدند. نتایج نشان داد که این مدل‌ها به‌دلیل ماهیت غیرخطی و پیچیدهٔ داده‌ها، به‌ویژه تعامل میان متغیرهایی مانند نوع کالا، وزن و مسافت، قادر به بازسازی دقیق الگوهای واقعی بازار نبودند و خطای سیستماتیک بالایی داشتند. ازاین‌رو، برای دستیابی به پیش‌بینی دقیق‌تر و پایدارتر، استفاده از مدل‌های تقویت گرادیان درختی مانند XGBoost و CatBoost ضروری و منطقی تشخیص داده شد.

۱. مدل XGBoost

XGBoost یکی از مدل‌های **تقویت گرادیان درختی** است. به زبان ساده، این مدل به‌جای ساخت یک درخت تصمیم واحد، **یک مجموعهٔ درختی از درختان کوچک و ساده می‌سازد** که هر درخت جدید تلاش می‌کند **اشتباه در پیش‌بینی‌های درخت قبلی را اصلاح کند.**

فرایند کار:

۱. ابتدا یک مدل ساده ساخته می‌شود و خطای پیش‌بینی محاسبه می‌شود؛

۲. یک درخت جدید می‌سازد که تمرکز آن بر **نمونه‌هایی است که مدل قبلی بیشترین خطا را داشته است؛**

۳. این روند تکرار می‌شود تا مدل نهایی، جمع پیش‌بینی‌های همهٔ درخت‌ها باشد و خطا حداقل شود.

ویژگی‌ها:

- مستلزم تبدیل متغیرهای دسته‌ای به عددی است؛
- توانایی یادگیری روابط غیرخطی و تعامل متغیرها را دارد؛
- حساس به داده‌های پرت است و ممکن است بیش‌برازش رخ دهد اگر تعداد درخت یا عمق زیاد باشد.

۲. مدل CatBoost

CatBoost نیز نوعی **تقویت گرادیان درختی** است، اما با نوآوری‌های خاص:

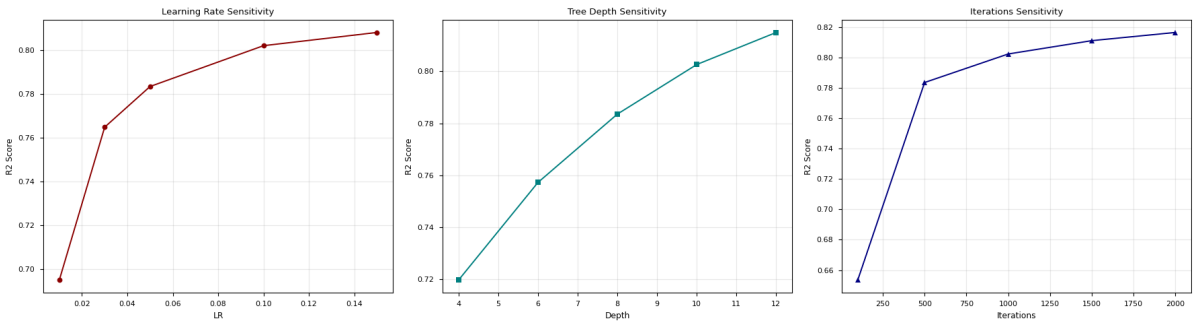
۱. پردازش مستقیم متغیرهای دسته‌ای: برخلاف XGBoost، نیازی به تبدیل متغیرهای کیفی به عددی ندارد. به‌این‌ترتیب، اطلاعات اصلی متغیرهای دسته‌ای حفظ می‌شود؛

۲. درختان متقارن: درخت‌های CatBoost همیشه شکل متقارن دارند، یعنی همهٔ مسیرها از ریشه تا برگ طول یکسان دارند. این کار باعث می‌شود پیش‌بینی پایدارتر و کمتر دچار بیش‌برازش شود؛

۳. Ordered Boosting: الگوریتمی که تلاش می‌کند پیش‌بینی هر نمونه را **فقط براساس نمونه‌های قبلی و بدون دسترسی به پاسخ آن نمونه** انجام دهد، تا مشکل انحراف داده‌ها (leakage) کاهش یابد.

ویژگی‌ها:

- پایداری بالا در داده‌های پرت و متغیرهای نادر؛
- قابلیت شناسایی تعامل پیچیده میان متغیرهای عددی و دسته‌ای



شکل ۱. تحلیل حساسیت عملکرد مدل CatBoost نسبت به تغییرات ابرپارامترها. این نمودار تغییرات شاخص دقت (R^2) در مدل CatBoost را به‌ترتیب در برابر تغییرات تعداد تکرار (سمت راست)، عمق درخت (وسط) و نرخ یادگیری (سمت چپ) نمایش می‌دهد.

Fig. 1. Sensitivity analysis of the CatBoost model performance relative to hyperparameter tuning. The plots display the variations in the R-squared (R^2) score for the CatBoost model against changes in iterations (right), tree depth (center), and learning rate (left).

• مناسب داده‌های حجیم و متنوع در حوزه‌های عملیاتی.

در این پژوهش، از مدل **تقویت گرادیان متقارن (CatBoost)** به‌عنوان مدل اصلی استفاده شد. علت انتخاب این مدل، توانایی منحصربه‌فرد آن در پردازش مستقیم متغیرهای کیفی (مانند نام کالا و شهر) بدون نیاز به تبدیل‌های عددی مخرب است. تنظیمات مدل براساس روش جست‌وجوی شبکه‌ای به شرح زیر تثبیت شد:

• **تعداد تکرار (۲۰۰۰ درخت)**: برای تضمین یادگیری الگوهای پیچیده در داده‌های حجیم؛

• **نرخ یادگیری (۰.۱)**: برای دستیابی به همگرایی نرم و جلوگیری از پرش مدل از روی نقاط بهینه (Overshooting)؛

• **عمق درخت (۱۰)**: این میزان عمق شناسایی تعاملات متقاطع تا سطح ۱۰ متغیر را (مثلاً اثر هم‌زمان نوع کالا، مقصد و وزن) برای مدل امکان‌پذیر می‌کند.

تحلیل پایداری و حساسیت مدل (sisylanA ytivitisneS)

در این مرحله، برای اطمینان از اینکه مدل پیشنهادی CatBoost در برابر تغییرات ابرپارامترها (Hyperparameters) رفتار پایداری دارد و بهترین نقطهٔ عملکردی انتخاب شده است، یک تحلیل حساسیت سه‌گانه انجام شد. نتایج این تحلیل در **شکل ۱** ارائه شده است.

• **تعداد تکرار (Iterations)**: در **شکل ۱ (نمودار سمت راست)**، روند یادگیری مدل نشان می‌دهد که پس از ۵۰۰ تکرار، بخش بزرگی از الگوها شناسایی و از تکرار ۱۵۰۰ به بعد، منحنی دقت کاملاً مسطح می‌شود. این همگرایی (Convergence) تضمین می‌کند که مدل دچار بیش‌برازش نشده و به پاسخی پایدار و تعمیم‌پذیر دست یافته است.

• **نرخ یادگیری (Learning Rate)**: مطابق با **شکل ۱ (نمودار سمت چپ)**، مشاهده می‌شود که با افزایش نرخ یادگیری از ۰.۰۱ به ۰.۱، ضریب تعیین (R^2) با شیب تندی صعود می‌کند و سپس به ثبات می‌رسد. این رفتار نشان‌دهندهٔ آن است که مدل در مقادیر پایین‌تر دچار «کندآموزی» بوده و در محدودهٔ ۰.۱، بهینه‌ترین تعادل را میان سرعت همگرایی و دقت پیش‌بینی پیدا کرده است.

• **عمق درخت (Tree Depth)**: براساس **شکل ۱** (نمودار وسط)، افزایش عمق درخت از ۴ به ۱۰ به بهبود مداوم دقت منجر شده است. از منظر تحلیل بیمه‌ای، عمق بیشتر شناسایی تعاملات متقاطع (Cross-interactions) بین متغیرها (مثلاً تأثیر هم‌زمان وزن بالا در یک مسیر جغرافیایی خاص برای یک نوع کالای ویژه) را برای مدل امکان‌پذیر می‌کند. دستیابی به دقت بیش از ۰.۸۰ در عمق ۱۰ نشان‌دهندهٔ پیچیدگی ذاتی روابط در داده‌های بارنامه است

معیارهای ارزیابی و اعتبارسنجی

به‌منظور سنجش دقیق توانمندی مدل‌های اجراشده، داده‌ها به دو بخش آموزش (۸۰٪) و تست (۲۰٪) تقسیم شدند. با توجه به استفاده از تبدیل لگاریتمی در مرحلهٔ پیش‌پردازش برای نرمال‌سازی توزیع مبالغ، تمامی پیش‌بینی‌های خروجی مدل با استفاده از تابع نمایشی (expm1) به مقیاس واقعی (ریال) بازگردانده شدند. تا ارزیابی‌ها بر مبنای ارزش‌های پولی واقعی انجام شود.

$$\mathrm{expm1}(x) = e^{x} - 1$$

عملکرد مدل با استفاده از معیارهای مکمل زیر واکاوی شد:

• **ضریب تعیین (R^2)**: این شاخص برای سنجش توانایی مدل در تبیین واریانس داده‌ها و میزان انطباق کلی پیش‌بینی‌ها با واقعیت بازار به کار گرفته شد.

• **ریشهٔ میانگین مربعات خطا (RMSE)**: این معیار به‌عنوان تابع هدف در فرایند آموزش مدل استفاده شد. به‌دلیل به توان رسیدن خطاها در این شاخص، مدل نسبت به خطاهای بزرگ (Outliers) حساس می‌شود و تلاش می‌کند انحرافات فاحش را در پیش‌بینی مبالغ بیمه به حداقل برساند.

• **میانهٔ قدر مطلق خطا (MedAE)**: با توجه به وجود داده‌های پرت احتمالی در بارنامه‌ها، از این شاخص برای درک ملموس

جدول ۱. مقایسه تطبیقی عملکرد مدل‌های پیش‌بینی براساس شاخص‌های دقت کلی و پایداری توزیع خطا.
Table 1. Comparative performance of prediction models based on overall accuracy metrics and error distribution stability indicators.

مدل	ضریب تعیین (R^2)	میانهٔ قدر مطلق خطا (ریال)[1]	صدک ۹۰ام قدر مطلق خطا (ریال)[2]	بیشینه خطا (ریال)[3]
XGBoost	0.84	53,446	410,313	2,926,858
CatBoost	0.86	51,839	395,317	2,637,000

و کاربردی میزان خطا در دنیای واقعی استفاده شد. MedAE نشان‌دهندهٔ خطای متوسط در ۵۰٪ از جامعهٔ آماری است و نسبت به مقادیر استثنایی مقاوم (Robust) است.

- **صدک نودم خطا (90th Percentile Error)**: این معیار به‌منظور تضمین پایداری عملیاتی مدل در ۹۰٪ موارد استفاده شد تا قابلیت اعتماد مدل در پایانه‌های حمل‌ونقل از دیدگاه نظارتی اثبات شود.
- **بیشینهٔ خطا (MaxE)**: این معیار، بزرگ‌ترین خطای پیش‌بینی مشاهده‌شده را نشان می‌دهد و معمولاً مربوط به کمتر از ۱۰٪ داده‌هاست که شامل محموله‌های ویژه با ارزش غیرمتعارف هستند. MaxE بیانگر توانایی مدل در مقابله با نوسانات شدید و داده‌های پرت است.

نتایج و بحث

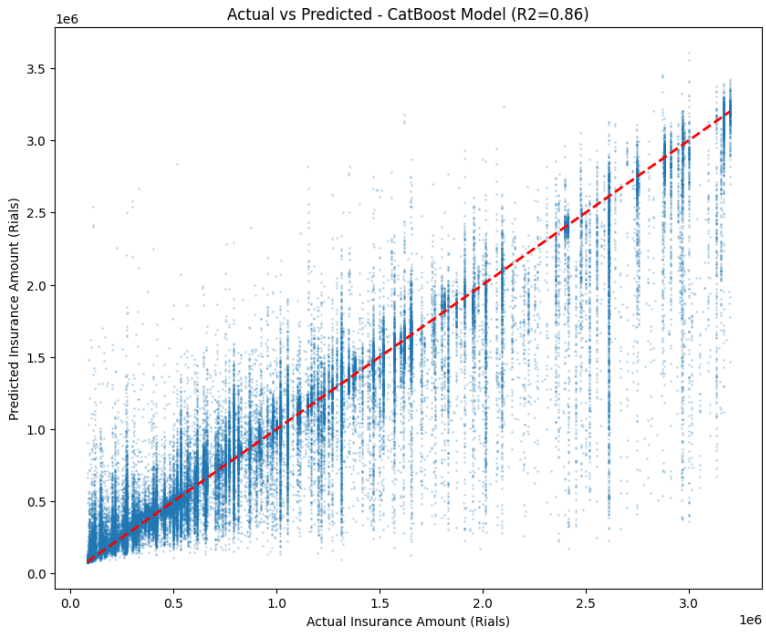
در این بخش، نتایج حاصل از اجرای مدل‌های یادگیری ماشین بر مجموعه دادهٔ جامع بارنامه‌های استان قم ارائه می‌شود. فرایند ارزیابی در دو سطح کمّی و کیفی انجام شده است؛ ابتدا با استفاده از معیارهای آماری استاندارد، کارایی مدل پیشنهادی در مقایسه با دیگر الگوریتم‌های پیشرو سنجیده شده و سپس از طریق تحلیل‌های نموداری، پایداری و رفتار مدل در مواجهه با نوسانات بازار بیمه واکاوی شده است. هدف از این تحلیل‌ها، صحه‌گذاری بر توانمندی مدل در جایگزینی با روش‌های سنتی و دستی قیمت‌گذاری است.

جدول ۱ نمایی جامع از **اعتبارسنجی آماری** و **کارایی عملیاتی** دو الگوریتم برتر یادگیری ماشین را در پیش‌بینی مبالغ بیمه نشان می‌دهد. در نگاه اول، هر دو روش **XGBoost** و **CatBoost** عملکرد بسیار مطلوبی از خود نشان داده و توانسته‌اند با دقت بالایی الگوهای پیچیده حمل‌ونقل استان قم را شناسایی کنند؛ بااین‌حال، همان‌طور که از اعداد مشخص است، مدل **CatBoost** در شاخص **R^2** و تمامی معیارهای خطا، برتری نسبی خود را حفظ می‌کند و به‌عنوان مدل بهینه معرفی می‌شود. در ادامه، تحلیل جامع این نتایج آورده شده است:

نتایج به‌دست‌آمده نشان می‌دهد که هر دو مدل در تبیین رفتار داده‌ها بسیار موفق عمل کرده‌اند، اما مدل **CatBoost** با دستیابی به ضریب تعیین **R^2 = 0.86**، توانمندی بالاتری در استخراج روابط غیرخطی میان متغیرهایی همچون نوع کالا، وزن و مسافت دارد. کسب این امتیاز به این معناست که ۸۶ درصد از نوسانات مبالغ بیمه توسط این مدل پوشش داده شده است که در داده‌های واقعی و پرنویز لجستیک، یک استاندارد طلایی و نشان‌دهندهٔ دقت بسیار بالای پیش‌بینی است. در تحلیل لایه‌های خطا، شاخص **میانهٔ قدر مطلق خطا (MedAE)** ثابت می‌کند که هر دو مدل برای اکثریت جامعه (بارهای روزمره) بسیار دقیق‌اند؛ به‌طوری که در ۵۰ درصد کل بارنامه‌ها، انحراف مدل CatBoost تنها **۵۱,۸۳۹ ریال** است. این مقدار انحراف در مقیاس هزینه‌های حمل‌ونقل عملاً ناچیز است و نشان می‌دهد که مدل در کاربرد عملیاتی، بسیار منصفانه و دقیق قیمت‌گذاری می‌کند. همچنین، شاخص **صدک ۹۰ام** همچون تیر خلاص در پاسخ به انتقادات احتمالی پیرامون بزرگی خطاها عمل می‌کند؛ این شاخص نشان می‌دهد که **۹۰ درصد پیش‌بینی‌های مدل** خطایی کمتر از **۳۹۵,۳۱۷ ریال** دارند. این بدان معناست که مدل در ۹ مورد از هر ۱۰ مورد، پایداری فوق‌العاده‌ای دارد. این خطاهای حداکثری (MaxE) که در بخش کوچکی از داده‌ها مشاهده می‌شود، عمدتاً مربوط به محموله‌های استراتژیک و خاص با ارزش اظهارشده غیرمتعارف است که از الگوی عمومی بازار تبعیت نمی‌کنند. در نهایت، کاهش چشمگیر در شاخص **بیشینهٔ خطا (MaxE)** توسط CatBoost نسبت به رقیب خود، نشان‌دهندهٔ قدرت این الگوریتم در مهار نوسانات شدید و مدیریت داده‌های استثنایی است. در مجموع، این جدول تأیید می‌کند که مدل پیشنهادی نه‌تنها از لحاظ نظری (ضریب تعیین) دقیق‌تر است، بلکه از نظر پایداری در ۹۰ درصد عملیات واقعی نیز، قابل اعتمادترین گزینه برای هوشمندسازی پایانه‌های حمل‌ونقل محسوب می‌شود.

شکل ۲ که به نمودار رگرسیون انطباق شناخته می‌شود، میزان دقت پیش‌بینی‌های مدل CatBoost را در برابر مقادیر واقعی می‌سنجد و نشان می‌دهد که نقاط آبی‌رنگ به‌عنوان نمایندهٔ بارنامه‌های انفرادی، با تمرکز بسیار بالایی پیرامون خط‌چین قرمز که وضعیت پیش‌بینی ایده‌آل است، قرار گرفته‌اند. نزدیکی حداکثری این نقاط به خط نیم‌ساز، گویای دقت بالای مدل در تخمین مبالغ بیمه است، به‌طوری که تحلیل ضریب تعیین با مقدار $R^2 = 0.86$ تأیید می‌کند که مدل CatBoost موفق شده است ۸۶ درصد از تغییرات و نوسانات کل مبلغ بیمه را فقط براساس متغیرهای ورودی شامل وزن، مسافت، نوع کالا و هویت شرکت حمل‌ونقل توضیح دهد که این عدد

[1] Median Absolute Error (MedAE)
[2] 90th Percentile of Absolute Error
[3] Maximum Error (MaxE)



شکل ۲. نمودار پراکندگی انطباق مقادیر واقعی و پیش‌بینی‌شده
Fig. 2. scatter plot of actual vs. predicted values

در محیط‌های واقعی لجستیک با نویز بالا، نشان‌دهندهٔ مدلی بسیار قدرتمند تلقی می‌شود. در بررسی رفتار مدل در بازه‌های مختلف، مشاهده می‌شود که در مبالغ پایین و متوسط، تراکم نقاط بر روی خط قرمز فوق‌العاده بالاست که دقت مدل در قیمت‌گذاری بارهای متعارف و روزمره را به اثبات می‌رساند، درحالی‌که در مقادیر بالاتر از ۲ میلیون ریال، اگرچه پراکندگی نقاط به‌دلیل وجود محموله‌های خاص و شرایط ریسک متفاوت اندکی افزایش می‌یابد، اما سوگیری کلی مدل کاملاً حفظ شده است. در نهایت، این نمودار ثابت می‌کند که الگوریتم پیشنهادی از تعمیم‌پذیری بالایی برخوردار است و می‌تواند روی داده‌های جدیدی که در فرایند آموزش مشاهده نکرده، با اطمینان بالا پاسخ دهد.

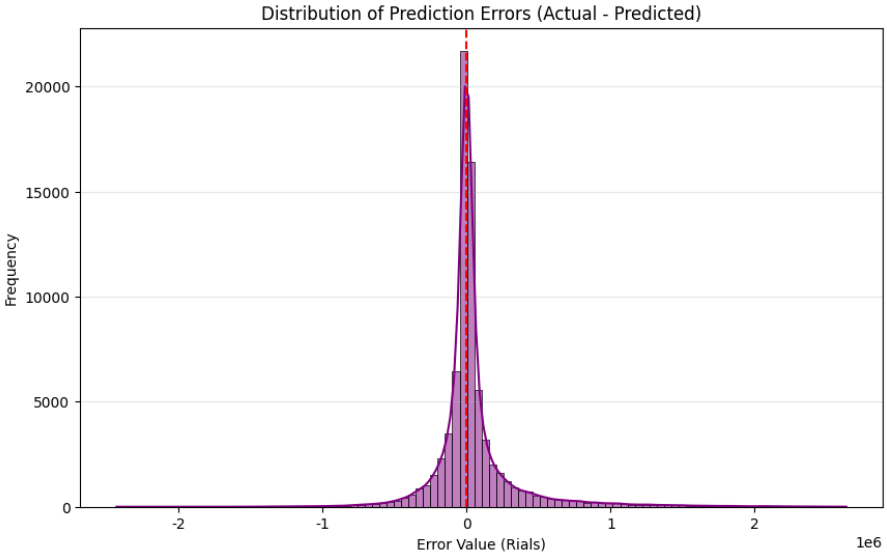
شکل ۳ توزیع خطاهای پیش‌بینی (Residuals) مدل CatBoost بر روی داده‌های تست، یک **توزیع زنگوله‌ای و متقارن** حول نقطهٔ صفر را نشان می‌دهد. تمرکز بالای فراوانی خطاها در بازهٔ نزدیک به صفر (قلهٔ نمودار) گواهی بر دقت بالای مدل در تخمین مبالغ بیمه بارنامه‌هاست. همچنین، متقارن بودن توزیع باقی‌مانده‌ها بیانگر این است که مدل دچار سوگیری (Bias) نیست و مبالغ را به طور مداوم کمتر یا بیشتر از حد واقعی پیش‌بینی نمی‌کند. کشیدگی ناچیز در دنباله‌های نمودار نیز نشان‌دهندهٔ پایداری مدل در مواجهه با داده‌های پرت (Outliers) پس از اعمال استراتژی پالایش ۱۰-۹۰ است.

علاوه‌بر دقت پیش‌بینی، تحلیل اهمیت ویژگی‌ها (Feature Importance) مشخص می‌کند که متغیرهای نام کالا (Product name) و نام شرکت حمل‌ونقل (Company name) بیشترین سهم را در تصمیم‌گیری‌های مدل ایفا می‌کنند. تأثیر بالای نام کالا نشان‌دهندهٔ ارتباط مستقیم نوع محموله با میزان ریسک و ارزش بیمه‌ای آن است. همچنین حضور ویژگی‌های مهندسی‌شده‌ای مانند تبدیل لگاریتمی متغیر وزن (Log_weight) در رتبه‌های برتر، صحت رویکرد پیش‌پردازش داده‌ها در این پژوهش را تأیید می‌کند؛ چراکه نشان می‌دهد مدل به‌خوبی توانسته است رابطهٔ غیرخطی بین وزن محموله و مبلغ بیمه را درک و در پیش‌بینی نهایی لحاظ کند (شکل ۴).

شکل ۵ نشان‌دهندهٔ تغییرات معیار خطای RMSE در طول ۲۰۰۰ تکرار برای مجموعه‌های آموزش و تست است. همگرایی سریع و یکنواخت هر دو منحنی در ۵۰۰ تکرار اول، بیانگر توانایی بالای مدل CatBoost در شناسایی الگوهای پیچیده و غیرخطی موجود در داده‌های بارنامه است. نکتهٔ حائز اهمیت، فاصلهٔ (Gap) بسیار ناچیز میان خطای آموزش و تست تا انتهای فرایند یادگیری است؛ این ویژگی پایداری مدل را تأیید می‌کند و نشان می‌دهد که ساختار پیشنهادی با وجود پیچیدگی داده‌ها، دچار سوگیری (Bias) یا بیش‌برازش (Overfitting) نشده و از قابلیت تعمیم‌پذیری بالایی برخوردار است.

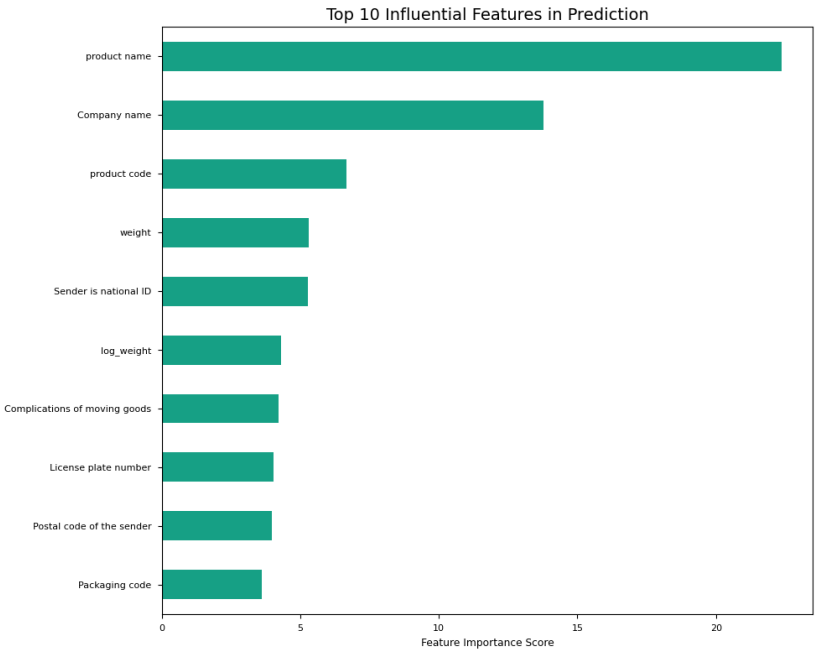
شکل ۶ نمودار توزیع تجمعی خطا (CDF) را نشان می‌دهد که ابزاری قدرتمند برای ارزیابی قابلیت اطمینان مدل در کاربردهای عملیاتی است. طبق این نمودار، حدود ۹۰٪ از پیش‌بینی‌های انجام‌شده با مدل پیشنهادی، دارای خطای مطلق کمتر از ۴۰۰,۰۰۰ ریال هستند. شیب تند نمودار در ابتدای محور، گواهی بر این واقعیت است که اکثر پیش‌بینی‌ها با کمترین انحراف نسبت به مبالغ واقعی بارنامه انجام شده‌اند.

علاوه‌بر دقت پیش‌بینی، یکی از اصلی‌ترین دستاوردهای عملیاتی این مدل، توانمندی آن در **ردیابی هوشمند نویزهای مشکوک و داده‌های پرت (Outliers)** در فرایند صدور بارنامه است.



شکل ۳. توزیع فراوانی خطای پیش‌بینی در مدل پیشنهادی (CatBoost)

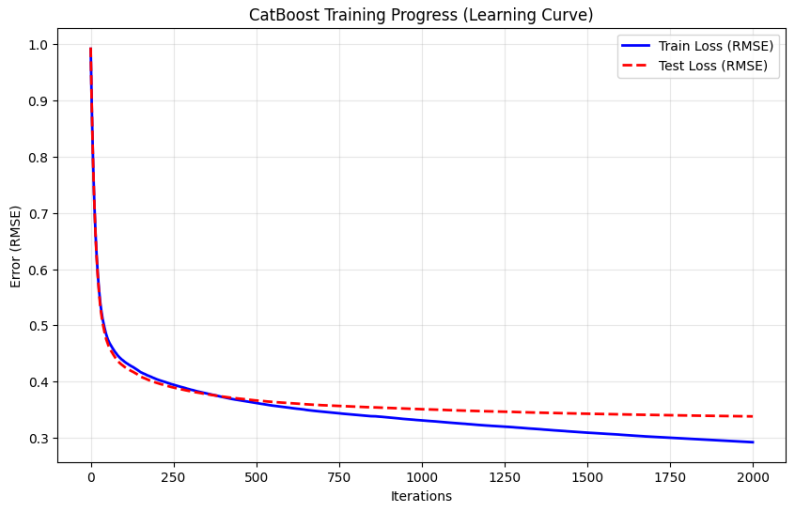
Fig. 3. Frequency distribution of prediction errors (residuals) for the proposed CatBoost model



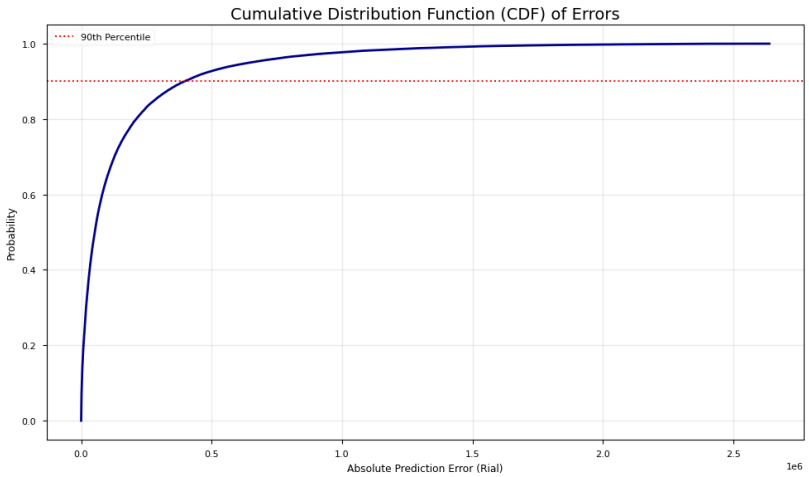
شکل ۴ .تحلیل اهمیت ویژگی‌ها در مدل CatBoost؛ شناسایی پیشران‌های اصلی تعیین مبلغ بیمه.

Fig. 4. Feature importance analysis identifying the primary drivers of insurance premium determination.

ازآنجاکه مدل CatBoost با استفاده از داده‌های حجیم و واقعی، الگوی حاکم بر بازار را فرا گرفته است، هرگونه انحراف فاحش میان «مبلغ پیش‌بینی‌شده» و «مبلغ ثبت‌شده در سیستم» می‌تواند یک **هشدار نظارتی (Flag)** تلقی شود. این قابلیت به سازمان راهداری و شرکت‌های بیمه امکان می‌دهد تا به‌جای بازرسی‌های تصادفی، بر موارد مشکوک تمرکز کنند. مثلاً، مواردی که در آن‌ها مبلغ بیمه بسیار کمتر از تخمین مدل است، می‌تواند نشان‌دهندهٔ کم‌اظهاری در ارزش کالا یا خطای ثبت سیستمی باشد. این رویکرد، مدل را از ابزاری محاسباتی به **سامانهٔ پایش آنی** تبدیل می‌کند که به‌صورت خودکار، الگوهای غیرمتعارف و تخلفات احتمالی را در گلوگاه‌های ترانزیتی استان قم شناسایی و از تضییع حقوق ذی‌نفعان جلوگیری می‌کند.



شکل ۵. منحنی یادگیری مدل CatBoost؛ مقایسهٔ روند کاهش خطای RMSE در مجموعه‌های آموزش (Train) و تست (Test)
Fig. 5. Learning curve of the CatBoost model showing the convergence of RMSE for training and testing sets



شکل ۶. تابع توزیع تجمعی (CDF) خطاهای مطلق پیش‌بینی. خط‌چین قرمز نشان‌دهندهٔ صدک ۹۰ام (90th Percentile) است.
Fig. 6. The CDF plot illustrates that the vast majority of prediction errors are concentrated near zero

جمع‌بندی و پیشنهادها

در این پژوهش، به‌منظور دستیابی به دقیق‌ترین الگوی پیش‌بینی مبلغ بیمه در صنعت حمل‌ونقل استان قم، یک خط‌لوله (Pipeline) محاسباتی قدرتمند طراحی و اجرا شد. نتایج اولیه با استفاده از مدل‌های خطی نشان داد که این الگوریتم به‌دلیل ماهیت ساده و فرض وجود روابط خطی، توانایی شناسایی الگوهای حاکم بر بازار بیمه را ندارد. در داده‌های حمل‌ونقل، متغیرهایی همچون وزن و مسافت به‌صورت هم‌افزا و غیرخطی بر ریسک و مبلغ بیمه اثر می‌گذارند. عدم انطباق پیش‌بینی‌های این مدل با واقعیت، تأییدی بر این مدعاست که قیمت‌گذاری بیمه فرایندی پیچیده و چندمتغیره است که فقط از طریق مدل‌های غیرپارامتریک درختی قابل ردیابی است. مدل **CatBoost** به‌عنوان هستهٔ اصلی این پژوهش، با دستیابی به **ضریب تعیین $R^2 = 0.86$** عملکردی فراتر از مدل‌های سنتی و حتی مدل‌های رقیب چون XGBoost نشان داد. برتری فنی این مدل در این پروژه مدیون دو نوآوری اصلی است:

۱. **مدیریت هوشمند متغیرهای دسته‌ای (Categorical Encoding)**: بخش بزرگی از دقت مدل مدیون توانایی CatBoost در پردازش مستقیم متغیرهایی همچون «**نام کالا**» و «**نام شرکت حمل‌ونقل**» است. این مدل بدون نیاز به تبدیل‌های عددی مخرب، توانسته است وزنِ ریسک هر نوع کالا و اعتبار هر شرکت را در محاسبات نهایی لحاظ کند.

۲. **تأثیر مهندسی ویژگی (Feature Engineering)**: معرفی متغیر **Log_weight** (لگاریتم طبیعی وزن) به‌عنوان یکی از پیشران‌های اصلی، موفقیت رویکرد ریاضی این پژوهش را ثابت کرد. تبدیل لگاریتمی

با هموارسازی چولگی (Skewness) داده‌های وزنی، به مدل کمک کرد تا تفاوت ریسک در محموله‌های سبک و سنگین را با دقت تفکیک کند. برای اطمینان از قابلیت اعتماد مدل در محیط واقعی، توزیع خطاها (Residuals) واکاوی شد. تمرکز بالای خطاها حول نقطهٔ صفر و توزیع متقارن آن‌ها نشان می‌دهد که مدل دچار سوگیری (Bias) سیستماتیک نیست. همچنین، شاخص **صدک ۹۰ام** تأیید می‌کند که ۹۰٪ از پیش‌بینی‌های مدل، خطایی کاملاً در بازهٔ پذیرش عملیاتی دارند. این پایداری، مدل پیشنهادی را به ابزاری قدرتمند برای مدیران تبدیل می‌کند تا پیش از صدور بارنامه، تخمینی دقیق از هزینه‌های بیمه‌ای داشته باشند و از اختلافات مالی احتمالی جلوگیری کنند.

تحلیل این نتایج فراتر از محاسبات آماری محض است و در واقع دریچه‌ای به‌سوی مدیریت هوشمند و عدالت‌محور در صنعت لجستیک استان قم می‌گشاید. دستیابی به دقت ۸۶ درصدی در تبیین نوسانات حق‌بیمه نشان می‌دهد که می‌توان با جایگزینی سیستم‌های سنتی و سلیقه‌ای با الگوریتم‌های داده‌محوری چون CatBoost، علاوه‌بر حذف خطای انسانی، شفافیت مالی را میان رانندگان، شرکت‌های حمل‌ونقل و نهادهای بیمه‌گر برقرار کرد. اهمیت عملیاتی این دستاورد در شاخص‌هایی همچون میانهٔ خطای ناچیز (۵ هزار تومان) نهفته است که به شرکت‌ها امکان می‌دهد پیش از صدور قطعی بارنامه، برآوردی دقیق از هزینه‌ها داشته باشند و ریسک‌های مالی ناشی از قیمت‌گذاری نادرست را به حداقل برسانند. همچنین، شناسایی متغیرهای کلیدی مانند لگاریتم وزن و نوع کالا به‌عنوان پیشران‌های اصلی، به سیاست‌گذاران این حوزه کمک می‌کند تا فرمول‌های تعرفه‌گذاری خود را براساس واقعیت‌های فیزیکی محموله اصلاح کرده، از هدررفت منابع در زنجیرهٔ تأمین جلوگیری کنند. در نهایت، این مدل نه‌تنها ابزاری برای پیش‌بینی، بلکه مرجعی منصف برای کاهش اختلافات حقوقی و بهینه‌سازی جریان نقدینگی در پایانه‌های باربری محسوب می‌شود

مشارکت نویسندگان

زینب جعفرزاده: تفسیر داده‌ها، تحلیل آماری و پیش‌نویس پژوهش. نرگس میره‌ای: روش‌شناسی و نتایج پژوهش. عفت گلپر رابوکی: نظارت و سرپرستی، تدوین و بازنگری پژوهش.

تشکر و قدردانی

تشکر و قدردانی نویسندگان از داوران محترم این پژوهش که با پیشنهادهای ارزشمند خود باعث بهبود و ارتقای محتوای شدند.

تعارض منافع

نویسندگان اعلام می‌دارند که در خصوص انتشار این مقاله تضاد منافع وجود ندارد. علاوه‌براین، موضوعات اخلاقی، از جمله سرقت ادبی، رضایت آگاهانه، سوءرفتار، جعل داده‌ها، انتشار و ارسال مجدد و مکرر و همچنین، سیاست مجله در قبال استفاده از هوش مصنوعی ازسوی نویسندگان رعایت شده است.

یادداشت ناشر

ناشر نشریهٔ *پژوهشنامه* بیمه با توجه به مرزهای حقوقی در نقشه‌های منتشرشده بی‌طرف باقی می‌ماند.

منابع

بزی، ب.، و رشیدی محمودی، م.، ع. (۱۳۹۸). *تشخیص هوشمند ناهنجاری در حمل‌ونقل شهری* [مقاله کنفرانسی]. سومین کنفرانس ملی پژوهش‌های کاربردی در علوم برق، کامپیوتر و مهندسی پزشکی. https://civilica.com/doc/۹۹۰۳۲۶

پرستش، م.، بهشتی‌فرد، ض.، و راد، ع. (۱۴۰۴). پیش‌بینی خسارت با استفاده از تکنیک‌های رگرسیون عمیق متوالی در روش‌های یادگیری عمیق. *پژوهشنامه بیمه*، ۱۴(۴)، ۲۸۱-۲۹۴. https://doi.org/۱۰/۲۲۰۵۶.ijir/۲۰۲۵/۰۴/۰۲

حسینی، ر.، و صیفی‌زاده، الف. (۱۴۰۲). *پیش‌بینی قیمت حمل در سیستم‌های مدیریت حمل کالا با استفاده از مدل‌های هوشمند* [مقاله کنفرانسی]. هفدهمین کنفرانس بین‌المللی فنّاوری اطلاعات. https://civilica.com/doc/1588819

Agrawal, N., Cohen, M. A., Deshpande, R., & Deshpande, V. (2024). How machine learning will transform supply chain management. *Harvard Business Review, 103*(3–4), 128–137. https://www.ew360.mx/clientes/Ado/especiales/050424/HBR4.pdf

Bazi, B., & Rashidi Mahmoudi, M. (2019). *Intelligent anomaly detection in urban transportation* [Conference presentation]. Third National Conference on Applied Research in Electrical, Computer and Medical Engineering. https://civilica.com/doc/990326 **[In Persian].**

De Araujo, A. C., & Etemad, A. (2021). End-to-end prediction of parcel delivery time with deep learning for smart-city applications. *IEEE Internet*

of Things Journal, 8(23), 17043–17056. https://doi.org/10.1109/JIOT.2021.3077007

Duan, M., Sun, S., & Liu, M. (2025). A multimodal deep fusion framework for highway traffic anomaly detection. *Scientific Reports, 15*(1), 33573. https://doi.org/10.1038/s41598-025-18671-x

Gang, H. R., Dong, N., & Zhao, L. (2023). An anomaly detection method for network freight documents based on improved multiple BRB. In *International Conference on Signal and Information Processing, Networking and Computers* (pp. 409–416). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-97-2124-5_49

Hosseini, R., & Seyfizadeh, A. (2023). *Predicting freight prices in freight management systems using intelligent models* [Conference presentation]. 17th International Conference on Information Technology, Computer and Telecommunications, Qods City, Tehran. https://civilica.com/doc/1588819

[In Persian].

Parastesh, M., Beheshtifard, Z., & Rad, A. (2025). Predicting damage using sequential deep regression techniques in deep learning methods. *Iranian Journal of Insurance Research*, *14*(4), 281–294. https://doi.org/10.22056/ijir.2025.04.02 **[In Persian].**

Rokoss, A., Syberg, M., Tomidei, L., Hülsing, C., Deuse, J., & Schmidt, M. (2024). Case study on delivery time determination using a machine learning approach in small batch production companies. *Journal of Intelligent Manufacturing, 35*(8), 3937–3958. https://doi.org/10.1007/s10845-023-02290-2

Zhang, Z., Qi, G., Ceder, A., Guan, W., Guo, R., & Wei, Z. (2021). Grid-based anomaly detection of freight vehicle trajectory considering local temporal window. *Journal of Advanced Transportation, 2021*(1), 8103333. https://doi.org/10.1155/2021/8103333

AUTHOR(S) BIOSKETCHES **معرفی نویسندگان**

زینب جعفرزاده (Zeynab Jafarzadeh)، کارشناسی، گروه علوم کامپیوتر، دانشکده علوم پایه، دانشگاه قم، قم، ایران.

- Email: *Zeinab.82jafarzadeh82@gmail.com*
- ORCID: 0009-0005-1675-2412

نرگس میره‌ای (Narges Mirehi)، استادیار، گروه علوم کامپیوتر، دانشکده علوم پایه، دانشگاه قم، قم، ایران.

- Email: *N.mirehi@qom.ac.ir*
- ORCID: 0009-0005-4110-8899

عفت گلپر رابوکی (Efat Golpar Raboky)، دانشیار، گروه علوم ریاضی، دانشکده علوم پایه، دانشگاه قم، قم، ایران.

- Email: *G.raboky@qom.ac.ir*
- ORCID: 0000-0002-1699-0403

HOW TO CITE THIS ARTICLE

Jafarzadeh, Z., Mirehi, N., & Golpar Raboky, E. (2026). Advanced analysis of transportation data using machine learning: Bill of lading insurance premium prediction. *Iranian Journal of Insurance Research*, *15*(3), 227-238.

DOI: 10.22056/ijir.2026.03.02

URL: https://ijir.irc.ac.ir/article_160374.html