Predicting Auto Insurance Risk Using Gradient Boosting
Analyzing Socio-Economic and Crash Data in New York City

AJ Strauman-Scott[a]

[a]*City University Of New York (CUNY), Department of Data Science, New York City, United States of America, 11212*

Abstract

PUT AN ABSTRACT HERE!!

Keywords: Gradient Boosting, XGBoost, SHAP explainability, hyperparameter optimization, auto insurance risk, American Community Survey (ACS), NYC Open Data, predictive modeling

1. 1. Introduction

Accurate insurance risk modeling is critical for setting fair premiums, mitigating losses, and ensuring financial stability within the insurance industry (Henckaerts et al., 2021, Clemente et al., 2023). Predicting claim frequency and severity not only supports pricing but also enables insurers to manage portfolio-level risk and optimize resource allocation (Mohamed et al., 2025).

New York City (NYC) presents a complex urban environment where traffic risks are shaped by socio-economic factors, dense infrastructure, and scaling dynamics typical of large metropolitan areas (Cabrera-Arnau et al., 2020, Bettencourt et al., 2007). The availability of open datasets—such as NYC's Motor Vehicle Collision (MVC) data and socio-economic indicators from the American Community Survey (ACS)—offers a unique opportunity to develop proxy models for insurance claim risk. These data sources provide detailed insights into crash frequency, injury severity, commuting behaviors, and neighborhood-level demographics (Adeniyi, 2024, Brubacher et al., 2016).

Traditional actuarial methods, such as Generalized Linear Models (GLMs), have long been the foundation of risk pricing and underwriting due to their interpretability and regulatory acceptance (Henckaerts et al., 2021). However, GLMs are limited in their ability to capture non-linear relationships and interactions among complex predictors like socio-economic factors, urban infrastructure, and driving behavior (Clemente et al., 2023). These limitations are particularly pronounced in urban contexts, where crash risk is shaped by

*Corresponding author
Email address: `true` (AJ Strauman-Scott)

heterogeneous population dynamics and localized factors (Cabrera-Arnau et al., 2020, Brubacher et al., 2016).

There is a growing need for data-driven approaches that can flexibly incorporate diverse predictors—such as open crash data and socio-economic variables—while addressing the complex temporal and spatial patterns of accidents highlighted in recent reviews (Grigorev et al., 2024, Behboudi et al., 2024). Recent studies and systematic reviews confirm that machine learning (ML) methods, particularly ensemble models like Gradient Boosting Machines (GBMs), XGBoost, and LightGBM, outperform traditional GLMs for predicting both claim frequency and severity (Clemente et al., 2023, Mohamed et al., 2025, Behboudi et al., 2024). These models are capable of handling mixed data types (categorical and continuous) and capturing complex feature interactions that linear models often miss.

To address the interpretability challenge of "black box" ML models, SHAP (SHapley Additive exPlanations) offers a principled framework for feature attribution, allowing insurers and policymakers to understand both global feature importance and instance-level predictions (Lundberg and Lee, 2017, Dong et al., 2022, Ning et al., 2024). This combination of high-performance prediction and explainability provides a strong foundation for modern risk modeling, as demonstrated in other domains such as maritime safety where interpretable models like SHAP have been applied (Kim and Lim, 2022).

Despite the growing body of work applying ML to insurance modeling, few studies integrate publicly available crash data with socio-economic indicators to model claim-related risks. Most research remains limited to proprietary policyholder data (Henckaerts et al., 2021, Mohamed et al., 2025), while systematic reviews highlight that few studies combine open crash data with socio-economic indicators in insurance modeling (Ali et al., 2024, Behboudi et al., 2024).

This study aims to integrate ACS socio-economic features with NYC MVC crash data to develop an explainable gradient boosting framework. The ultimate goal is to identify key socio-economic and transportation predictors that drive claim frequency and severity proxies, offering insights for both insurers and urban policymakers.

The remainder of this paper is organized as follows: Section 2 reviews prior work on machine learning in insurance risk modeling, crash and socio-economic data, geospatial analytics, model explainability, and literature gaps; Section 3 details the data sources, key metrics, modeling approach, and SHAP-based explainability; Section 4 reports the results including model performance, feature importance, and geospatial patterns; Section 5 discusses the findings in relation to existing research and industry applications; and Section 6 concludes with key contributions, limitations, and directions for future research.

2. 2. Related Work

2.1. 2.1 Machine Learning in Insurance Risk Modeling

The transition from traditional actuarial models such as Generalized Linear Models (GLMs) to machine learning (ML) approaches has marked a significant evolution in insurance risk modeling. GLMs have historically served as the backbone for pricing and claim prediction due to their interpretability and regulatory acceptance. However, they are limited by their linearity and inability to naturally capture complex interactions and nonlinear relationships among predictors, such as driver demographics, vehicle characteristics, socio-economic factors, and driving behavior. As Clemente et al. (2023) note, while GLMs remain effective for modeling claim severity with smaller and noisier datasets, they often underperform compared to ensemble methods when modeling claim frequency, where nonlinearities and heterogeneous risk patterns are prevalent. Similarly, Jonkheijm (2023) demonstrated that tree-based models, especially XGBoost, substantially improved predictive accuracy over linear regression, particularly when incorporating both actuarial features (e.g., policyholder age, vehicle value) and behavioral indicators.

Recent studies have validated the predictive superiority of ML methods—such as random forests, gradient boosting machines (GBM), and neural networks—over traditional actuarial models. Gradient boosting methods, such as XGBoost and LightGBM, have emerged as particularly effective tools in auto insurance risk modeling (Henckaerts et al., 2021). Their iterative boosting framework enables them to handle mixed data types (categorical and continuous) and capture intricate patterns that GLMs and single decision trees may miss. Clemente et al. (2023) applied gradient boosting to both claim frequency and severity modeling, demonstrating significant performance gains in frequency prediction over Poisson-based GLMs. Similarly, Jonkheijm (2023) employed XGBoost for forecasting individual claim amounts, outperforming both regression trees and random forests.

2.2. 2.2 Use of Crash and Socio-Economic Data

Crash data has been widely recognized as a reliable proxy for insurance claim frequency and severity, given the direct link between the occurrence of traffic accidents and subsequent claims filed by policyholders. Studies leveraging police crash reports, telematics, and open transportation datasets consistently demonstrate strong correlations between crash frequency and insurance risk metrics (Takale et al., 2022). The integration of socio-economic features—including income levels, commuting patterns, vehicle ownership rates, and population density—has been shown to enhance the explanatory power of crash and claim prediction models.

For example, Adeniyi (2024) utilized a decade of NYC crash data (2013–2023) to identify key predictors of accident severity—such as unsafe speed, alcohol involvement, and adverse weather—which align closely with the variables insurers use to model claim likelihood. Similarly, Dong et al. (2022) applied boosting-based ensemble models to traffic injury severity prediction, finding that vehicle type, collision mode, and

environmental conditions strongly influenced both injury outcomes and, by extension, potential claim costs. Brubacher et al. (2016) conducted a geospatial analysis of 10 years of crashes in British Columbia and found that regions with lower income and higher socio-economic deprivation exhibited higher rates of pedestrian crashes, severe injuries, and fatalities, reflecting disparities in road safety linked to infrastructure quality and enforcement intensity. Cabrera-Arnau et al. (2020) expanded on this by identifying superlinear scaling of road accidents in urban areas, where higher population densities led to disproportionate increases in crash frequency, especially for minor collisions. These findings are directly relevant for insurers, as they imply that socio-economic and urban structural factors—such as commuting patterns or access to public transit—can serve as proxies for underlying risk exposure.

Urban-focused studies have further illuminated the unique risk dynamics in metropolitan environments like New York City, Chicago, and London, where complex traffic patterns, dense road networks, and high pedestrian activity elevate accident risk. Adeniyi (2024) analyzed NYC crash data to show how the COVID-19 pandemic altered accident patterns, with fewer total crashes but an increase in injury severity due to higher vehicle speeds on less congested roads. Feng et al. (2020), studying UK traffic data, emphasized the value of big data platforms and spatial clustering techniques (e.g., accident hotspot detection) to identify urban risk zones, a concept that parallels insurer efforts to assess region-based risk for underwriting.

Collectively, these studies support the notion that combining crash data with socio-economic indicators offers a powerful means of modeling insurance claim frequency and severity. By integrating open data sources—such as NYC's Vision Zero crash records and U.S. Census-derived socio-economic attributes—researchers and insurers can capture a more holistic view of driver risk behavior, infrastructure quality, and regional safety disparities.

2.3. 2.3 Explainability in Machine Learning Models

In high-stakes fields such as insurance pricing, underwriting, and claims management, the interpretability of machine learning (ML) models is not only a technical preference but also a regulatory and business requirement. Insurers must be able to justify rating factors and risk scores to regulators, policyholders, and internal stakeholders. Traditional actuarial models like GLMs are naturally interpretable due to their linear structure and explicit coefficient estimates. However, modern ML models—such as gradient boosting or neural networks—are often criticized as "black boxes," complicating the explanation of predictions that influence financial decisions or customer premiums. Regulatory frameworks, including the EU's General Data Protection Regulation (GDPR) and U.S. state-level insurance guidelines, increasingly require transparency in algorithmic decision-making, further amplifying the need for explainable AI (XAI). Henckaerts et al. (2021) further underscore this, showing that variable importance plots and PDPs can yield actionable insights into driver and policyholder risk factors, blending predictive power with interpretability.

Among XAI methods, SHAP (SHapley Additive exPlanations) has become the state-of-the-art framework for interpreting complex ML models. Developed by Lundberg and Lee (2017), SHAP is grounded in cooperative game theory, assigning each feature a Shapley value that quantifies its contribution to individual predictions. Unlike traditional feature importance metrics—such as Gini importance in random forests or split gain in XGBoost—SHAP accounts for both main effects and feature interactions, offering a consistent and additive explanation of how variables drive model outputs.

Tools like SHAP allow practitioners to interpret complex models by quantifying the contribution of each variable to the predictions. Studies like Mohamed et al. (2025) highlight the value of such interpretability when using gradient boosting for pricing and fraud detection, as insurers must justify rating factors for regulatory compliance.

In the insurance domain, SHAP has been widely applied to interpret models for claims prediction, fraud detection, and risk scoring. Dong et al. (2022) used SHAP in conjunction with boosting-based models (LightGBM and CatBoost) to analyze the contribution of driver age, vehicle type, and collision type to injury severity predictions, providing insights that aligned with domain expertise. Similarly, Ning et al. (2024) demonstrated how Shapley Variable Importance Cloud (ShapleyVIC) builds on SHAP principles to assess variable significance with uncertainty intervals, enabling fairer and more transparent risk predictions. These approaches not only improve trust in ML-driven decision-making but also help insurers identify the most actionable risk factors influencing claims.

2.4. 2.4 Gaps in the Literature

While machine learning methods—particularly ensemble models like gradient boosting—have gained traction in insurance risk modeling, there is a notable absence of studies that combine socio-economic and crash data for claim risk prediction. Most existing research focuses on proprietary insurance datasets containing policyholder and vehicle information (Clemente et al., 2023, Henckaerts et al., 2021, Jonkheijm, 2023). However, publicly available crash datasets, such as NYC's Motor Vehicle Collision (MVC) reports, and socio-economic features from the American Community Survey (ACS) remain underutilized in insurance modeling. This gap limits the development of robust, regionally sensitive models that capture the real-world interaction between driving risk factors (e.g., crash frequency) and socio-economic indicators (e.g., income, commuting patterns, and vehicle ownership rates). By integrating ACS data with urban crash records, it becomes possible to construct granular, location-aware risk models that better reflect variations in driving exposure, infrastructure quality, and neighborhood-level risk factors.

3. 3. Materials and Methods

3.1. 3.1 Data Sources and Preprocessing

This study integrates publicly available crash data from New York City with socio-economic features from the American Community Survey (ACS) to develop a proxy model for insurance claim risk. The data sources and preprocessing steps are designed to replicate key factors used in actuarial risk models while incorporating broader socio-economic and regional variables.

3.1.1. ***Crash Data (Claim Proxies)***

Crash data is obtained from the NYC Motor Vehicle Collisions (MVC) Open Data Portal, covering the years 2018–2023. Each record includes details such as crash location, number of injuries and fatalities, vehicle type, and contributing factors (e.g., driver behavior, environmental conditions). These variables are well-documented predictors of both accident severity and insurance claims (Adeniyi, 2024, Dong et al., 2022).

Crash frequency was aggregated at the 2020 census tract level and normalized by tract-level population to compute crashes per 1,000 resident. This mentric will replace claim frequency (Brubacher et al., 2016).

3.1.2. ***Socio-Economic Data (ACS Features)***

Socio-economic variables are drawn from the ACS 5-year estimates (2018–2023) at the 2020 census tract level. The variables include demographic composition (e.g., % male, % white, % Black, % Asian, % Hispanic, % foreign-born), age distribution (% under 18, % 18–34, % 35–64, % 65+), and income indicators (median income, % households earning <$25,000, % households earning $25,000–$75,000, % below the poverty line). Additional features include median gross rent, housing tenure (% owner- vs. renter-occupied), educational attainment (% with high school diploma, % with bachelor's or graduate degrees), employment metrics (% in labor force, unemployment rate), and transportation factors (% driving alone, % carpooling, % using public transit, % walking, % biking, % working from home, and commute time distributions). Interaction features poverty × vehicle ownership and unemployment × vehicle ownership were engineered to capture compound effects on risk exposure. These features have been shown to correlate with traffic risk and claim likelihood, as evidenced by Brubacher et al. (2016) and Cabrera-Arnau et al. (2020), who found that socio-economic deprivation and commuting behaviors strongly influence crash frequency and severity.

Table 1: Table 1: ACS tables used and derived variables.

ACS.Table	Description	Derived.Variables
B01001	Age and Sex	total_population, male_population, female_population, age_under_18, age_18_34, age_35_64, age_65_plus
B01003	Total Population	total_population
B08134	Means of Transportation to Work by Vehicle Occupancy	drive_alone, carpool
B08301	Means of Transportation to Work	public_transit, walk, bike, work_from_home
B08303	Travel Time to Work	commute_short, commute_medium, commute_long
B19001	Household Income Distribution	income_under_25k, income_25k_75k, income_75k_plus
B19013	Median Household Income	median_income
B25010	Average Household Size	average_household_size
B25044	Tenure by Vehicles Available	no_vehicle, one_vehicle, two_plus_vehicles
C24010	Occupation by Sex and Median Earnings	occupation variables (aggregated)
C24030	Industry by Sex and Median Earnings	industry variables (aggregated)
B15003	Educational Attainment	less_than_hs, hs_diploma, some_college, associates_degree, bachelors_degree, graduate_degree
B17001	Poverty Status	below_poverty, above_poverty, poverty_rate
B02001	Race	white_population, black_population, asian_population
B03002	Hispanic or Latino Origin by Race	hispanic_population
B08201	Household Size by Vehicles Available	vehicle availability (aggregated)
B18101	Sex by Age by Disability Status	disability variables (aggregated)
B16005	Nativity by Language Spoken at Home	foreign_born
B23025	Employment Status for Population 16+	in_labor_force, employed, unemployed, not_in_labor_force, unemployment_rate
B25064	Median Gross Rent	median_gross_rent
B09005	Household Type (Families vs Non-families)	household_type variables

ACS.Table	Description	Derived.Variables
B11001	Household Type by Presence of Children	household_type_children variables

3.1.3. ***Preprocessing Steps***

Crash records from the NYC Open Data MVC dataset are cleaned (removing rows with missing or zero coordinates) and spatially joined to 2020 Census Tracts using official census tract shapefiles. Annual summaries of total crashes, injuries, and fatalities are then aggregated by tract and normalized by tract-level population to compute per-capita crash, injury, and fatality rates. The ACS socio-economic data are harmonized to 2020 tract boundaries (via crosswalks for 2018–2019), binned into interpretable categories (e.g., income brackets, age groups, education levels), and converted to percentages of total population where applicable. Interaction features—such as poverty × vehicle ownership and unemployment × vehicle ownership—are engineered to capture compounded socio-economic risk factors.

No categorical encoding besides `year` or standardization was performed at this stage since all ACS features are already expressed as continuous percentages or numeric values, and gradient boosting models (XGBoost) handle raw scales effectively (Henckaerts et al., 2021). The resulting integrated dataset combines socio-economic indicators with tract-level crash metrics, allowing us to explore how demographic and transportation characteristics interact with crash patterns to inform insurance claim frequency and severity modeling, filling a gap in the literature where public ACS and crash data are rarely combined (Mohamed et al., 2025, Jonkheijm, 2023).

3.2. 3.2 Key Metrics

To model insurance risk in the absence of proprietary claims data, publicly available crash data is used as a proxy for both claim frequency and severity, a practice supported by prior research that links traffic accidents directly to insurance losses (Adeniyi, 2024, Dong et al., 2022).

Our primary risk metric, crash_rate_per_1000, measures the number of crashes per 1,000 residents in each census tract-year. This population-adjusted rate follows the methodology of studies that normalize crash counts by population to ensure fair comparisons of relative risk across areas with varying exposure levels (Brubacher et al., 2016, Cabrera-Arnau et al., 2020).

These crash metrics are modeled alongside the socio-economic and transportation variables detailed in Table 2, which include demographic distributions (e.g., % white, % Hispanic), income and poverty indicators (e.g., median income, poverty rate), commuting and transportation patterns (e.g., % public transit use, % walk,

% bike), and engineered interaction features (e.g., poverty × vehicle ownership). Together, these variables allow the model to capture both the exposure risk (frequency) and potential cost severity of accidents, aligning with the frameworks used in both insurance (Clemente et al., 2023, Henckaerts et al., 2021) and traffic safety research (Dong et al., 2022).

Table 2: Table 2: Key variables, descriptions, and transformations in the final dataset.

Variable	Description	Type	Transformation
pct_male_population	Percentage of population that is male	Demographic	Percentage
pct_white_population	Percentage of population identifying as white	Demographic	Percentage
pct_black_population	Percentage of population identifying as black	Demographic	Percentage
pct_asian_population	Percentage of population identifying as Asian	Demographic	Percentage
pct_hispanic_population	Percentage of population identifying as Hispanic/Latino	Demographic	Percentage
pct_foreign_born	Percentage of population that is foreign-born	Demographic	Percentage
pct_age_under_18	Percentage of population under 18	Age	Percentage
pct_age_18_34	Percentage of population aged 18-34	Age	Percentage
pct_age_35_64	Percentage of population aged 35-64	Age	Percentage
pct_age_65_plus	Percentage of population aged 65 and above	Age	Percentage
median_income	Median household income (inflation-adjusted)	Income/Poverty	Raw value (USD)
pct_income_under_25k	Percentage of households earning less than $25,000	Income/Poverty	Percentage
pct_income_25k_75k	Percentage of households earning $25,000-$75,000	Income/Poverty	Percentage
pct_below_poverty	Percentage of population below the poverty line	Income/Poverty	Percentage
median_gross_rent	Median gross rent (USD)	Housing	Raw value (USD)
pct_owner_occupied	Percentage of owner-occupied housing units	Housing	Percentage
pct_renter_occupied	Percentage of renter-occupied housing units	Housing	Percentage

Variable	Description	Type	Transformation
pct_less_than_hs	Percentage with less than high school education	Education	Percentage
pct_hs_diploma	Percentage with a high school diploma	Education	Percentage
pct_some_college	Percentage with some college education	Education	Percentage
pct_associates_degree	Percentage with an associate's degree	Education	Percentage
pct_bachelors_degree	Percentage with a bachelor's degree	Education	Percentage
pct_graduate_degree	Percentage with a graduate or professional degree	Education	Percentage
pct_in_labor_force	Percentage of population in the labor force	Employment	Percentage
pct_not_in_labor_force	Percentage of population not in the labor force	Employment	Percentage
unemployment_rate	Unemployment rate	Employment	Percentage
pct_commute_short	Percentage with commute under 15 minutes	Transport/Commute	Percentage
pct_commute_medium	Percentage with commute between 15-30 minutes	Transport/Commute	Percentage
pct_commute_long	Percentage with commute longer than 30 minutes	Transport/Commute	Percentage
pct_carpool	Percentage commuting by carpool	Transport/Commute	Percentage
pct_public_transit	Percentage commuting by public transit	Transport/Commute	Percentage
pct_walk	Percentage commuting by walking	Transport/Commute	Percentage
pct_bike	Percentage commuting by biking	Transport/Commute	Percentage
pct_work_from_home	Percentage working from home	Transport/Commute	Percentage
pct_vehicle	Percentage of households with at least one vehicle	Transport/Commute	Percentage
post_pandemic	Post-pandemic indicator (1 = 2020 and later)	Engineered Feature	Binary
poverty_vehicle_interaction	Interaction term: poverty rate × vehicle ownership	Engineered Feature	Interaction
unemployment_vehicle_interaction	Interaction term: unemployment rate × vehicle ownership	Engineered Feature	Interaction
year2018	Year dummy: 2018	Year Indicator	Indicator
year2019	Year dummy: 2019	Year Indicator	Indicator
year2020	Year dummy: 2020	Year Indicator	Indicator

Variable	Description	Type	Transformation
year2021	Year dummy: 2021	Year Indicator	Indicator
year2022	Year dummy: 2022	Year Indicator	Indicator
year2023	Year dummy: 2023	Year Indicator	Indicator

3.3. 3.3 Modeling Approach

To model the relationship between socio-economic characteristics and crash risk, we implemented a single gradient boosting framework using XGBoost, rather than multiple boosting algorithms. Gradient boosting was selected because of its proven ability to model complex, non-linear interactions and handle heterogeneous input variables (for this project, percentages, continuous income values, and engineered features) without requiring variable standardization or heavy preprocessing (Clemente et al., 2023, Mohamed et al., 2025).

XGBoost is chosen for its strong track record in insurance risk modeling and interpretability when combined with SHAP (Dong et al., 2022). This model selection aligns with studies comparing boosting frameworks for both frequency-severity modeling (Henckaerts et al., 2021) and urban crash prediction (Adeniyi, 2024).

To optimize model performance, we performed hyperparameter tuning through exploratory grid searches rather than using automated Bayesian optimization frameworks like Optuna. The tuning process focused on key parameters such as `max_depth`, `learning_rate`, `subsample`, and `colsample_bytree`, which significantly influence the complexity and generalization ability of gradient boosting models. This approach is supported by prior research showing that systematic hyperparameter optimization significantly improves boosting model accuracy (Liu et al., 2025).

Each configuration was evaluated using spatial cross-validation at the borough level on the training data to balance bias and variance, ensuring that the model captured meaningful patterns without overfitting or overgeneralizing across geography. This iterative approach, while more manual than advanced optimization tools, was sufficient for our dataset size and feature set, yielding robust improvements in predictive accuracy. Model performance was then validated on the holdout test set using RMSE, MAE, and R^2 metrics.

3.3.1. Explainability

Given the regulatory and operational need for transparent, explainable models in insurance (Henckaerts et al., 2021, Lundberg and Lee, 2017), we employ SHAP (SHapley Additive exPlanations) for both global and local feature analysis. SHAP values are aggregated across the dataset to quantify overall feature importance, revealing which socio-economic and crash-related variables most influence predicted claim frequency and severity.

4. 4. Results

- Model performance metrics (AUC, RMSE).
- Feature importance rankings (SHAP).
- Visual findings (maps, animations, dashboards).

5. 5. Conclusions and Future Work

- Key insights and implications for insurance risk modeling.
- Limitations and suggested extensions (e.g., telematics, temporal models).
- Potential improvements in visualization and explainability.

References

Adeniyi, A.P., 2024. Understanding road accident patterns using exploratory data mining: A case study of nyc. Alabama A&M University URL: https://www.proquest.com/openview/6a23e635f5211337c03fd3ed364b0297/1. master's Thesis, Department of Community and Regional Planning.

Ali, Y., Hussain, F., Haque, M.M., 2024. Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. Accident Analysis & Prevention 194, 107378. URL: https://doi.org/10.1016/j.aap.2023.107378.

Behboudi, N., Moosavi, S., Ramnath, R., 2024. Recent advances in traffic accident analysis and prediction: A comprehensive review of machine learning techniques. arXiv preprint arXiv:2406.13968 URL: https://arxiv.org/abs/2406.13968.

Bettencourt, L.M.A., Lobo, J., Helbing, D., Kühnert, C., West, G.B., 2007. Growth, innovation, scaling, and the pace of life in cities. Proceedings of the National Academy of Sciences 104, 7301–7306. URL: https://doi.org/10.1073/pnas.0610172104.

Brubacher, J.R., Chan, H., Erdelyi, S., Schuurman, N., Amram, O., 2016. The association between regional environmental factors and road trauma rates: A geospatial analysis of 10 years of road traffic crashes in british columbia, canada. PLoS ONE 11, e0153742. URL: https://doi.org/10.1371/journal.pone.0153742.

Cabrera-Arnau, C., Prieto Curiel, R., Bishop, S.R., 2020. Uncovering the behaviour of road accidents in urban areas. Royal Society Open Science 7, 191739. URL: https://doi.org/10.1098/rsos.191739.

Clemente, C., Guerreiro, G.R., Bravo, J.M., 2023. Modelling motor insurance claim frequency and severity using gradient boosting. Risks 11, 163. URL: https://doi.org/10.3390/risks11090163.

Dong, S., Khattak, A., Ullah, I., Zhou, J., Hussain, A., 2022. Predicting and analyzing road traffic injury severity using boosting-based ensemble learning models with shapley additive explanations. International Journal of Environmental Research and Public Health 19, 2925. URL: https://doi.org/10.3390/ijerph19052925.

Feng, M., Zheng, J., Ren, J., Liu, Y., 2020. Towards big data analytics and mining for uk traffic accident analysis, visualization & prediction. Proceedings of the 2020 12th International Conference on Machine Learning and Computing (ICMLC) , 225–229URL: https://doi.org/10.1145/3383972.3384034.

Grigorev, A., Mihaita, A.S., Chen, F., 2024. Traffic incident duration prediction: A systematic review of techniques. Journal of Advanced Transportation 2024, Article ID 3748345, 36 pages. URL: https://doi.org/10.1155/atr/3748345.

Henckaerts, R., Côté, M.P., Antonio, K., Verbelen, R., 2021. Boosting insights in insurance tariff plans with tree-based machine learning methods. North American Actuarial Journal 25, 255–285. URL: https://doi.org/10.1080/10920277.2020.1745656.

Jonkheijm, T., 2023. Forecasting insurance claim amounts in the private automobile industry using machine learning algorithms. Tilburg University .

Kim, G., Lim, S., 2022. Development of an interpretable maritime accident prediction system using machine learning techniques. IEEE Access 10, 41313–41329. URL: https://doi.org/10.1109/ACCESS.2022.3168302.

Liu, P., Zhang, W., Wu, X., Guo, W., Yu, W., 2025. Driver injury prediction and factor analysis in passenger vehicle-to-passenger vehicle collision accidents using explainable machine learning. Vehicles 7, 42. URL: https://doi.org/10.3390/vehicles7020042.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017) , 4765–4774URL: https://arxiv.org/abs/1705.07874.

Mohamed, H.S., Abdelhamed, F.S., Mahdy, H.K., 2025. Machine learning algorithms to improve insurance claim prediction. Faculty of Business Administration 1, 20–36.

Ning, Y., Li, S., Ng, Y.Y., Chia, M.Y.C., Gan, H.N., Tiah, L., Mao, D.R., Ng, W.M., Leong, B.S.H., Doctor, N., Ong, M.E.H., Liu, N., 2024. Variable importance analysis with interpretable machine learning for fair risk prediction. PLOS Digital Health 3, e0000542. URL: https://doi.org/10.1371/journal.pdig.0000542.

Takale, D.G., Gunjal, S.D., Khan, V.N., Raj, A., Gujar, S.N., 2022. Road accident prediction model using data mining techniques. NeuroQuantology 20, 2904–2911. URL: https://doi.org/10.48047/NQ.2022.20.16.NQ880299.