Predicting Auto Insurance Risk Using Gradient Boosting
Analyzing Socio-Economic Predictors in Crash Data for 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,*

Abstract

This study explores the use of the gradient boosting model XGBoost to predict auto insurance risk by integrating socio-economic variables from publicly available data. By treating crash frequency as proxy for insurance claims, the project aims to identify key neighborhood-level factors influencing risk. The dataset, encompassing 13,518 tract-year observations from 2018 to 2023, captures demographic, economic, housing, and commuting indicators alongside engineered interaction variables. Optuna hyperparameter tuning and SHAP-based explainability reveal that post-pandemic traffic dynamics, median gross rent, percent of a population in the labor-force, and the interaction of poverty with vehicle ownership are significant predictors of crash risk. While the model achieves moderate predictive accuracy ($R^2 = 0.26$), its interpretability highlights socio-economic disparities that influence urban traffic safety. The findings underscore the potential of open data-driven models for portfolio-level risk assessment and urban safety planning, while cautioning against direct use for individual underwriting due to fairness and legal concerns. Future work should incorporate telematics data and fairness-aware algorithms to improve granularity and reduce bias.

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

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).

---

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

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 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 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 methods, particularly ensemble models like Gradient Boosting Machines (GBMs), 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 (Kim and Lim, 2022).

Despite the growing body of work applying GBMs to insurance modeling, few studies integrate publicly available crash data with socio-economic indicators to model claim-related risks specifically for the automotive insurance sector. 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 for measureing social risk for automotive insurance and urban policy. The ultimate goal is to identify key socio-economic and transportation predictors that drive claim frequency.

The remainder of this paper is organized as follows: Section 2 reviews prior work on ML 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 hyperparameter tuning results, model performance, and feature importance; 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. Related Work

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 (Chen and Guestrin, 2016), 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, GBMs, and neural networks—over traditional actuarial models. GBMs, 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.

3

2.2. *Use of Crash and Socio-Economic Data*

Crash data has been widely recognized as a reliable proxy for insurance claim frequency, given the direct link between the occurrence of traffic accidents and subsequent claims filed by policyholders. Studies utilizing 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.

2.3. *Explainability in GBM Models*

In high-stakes fields such as insurance pricing, underwriting, and claims management, the interpretability of 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 (shortened to 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. Gaps in the Literature

While ML 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 crash frequency) and socio-economic indicators. By integrating geospatially-coded ACS data with MVC 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. Materials and Methods

3.1. Data Sources

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 metric 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, % Hispanic, % foreign-born), age distribution, and income indicators. Additional features include median gross rent, housing tenure, educational attainment, employment metrics, and transportation factor. Interaction features poverty × vehicle ownership and unemployment × vehicle ownership were engineered to capture compound effects on risk exposure.

Table 1: ACS tables used and derived variables.

ACS	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
B08301	Transportation to Work	drive_alone, carpool, public_transit, walk, bike, work_from_home
B08303	Travel Time to Work	commute_short, commute_medium, commute_long
B19001	Household Income	income_under_25k, income_25k_75k, income_75k_plus, median_income
B25010	Average Household Size	average_household_size
B25044	Vehicles Available	no_vehicle, one_vehicle, two_plus_vehicles
C24010	Occupation	occupation variables (aggregated)
C24030	Industry	industry variables (aggregated)
B15003	Education	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	hispanic_population
B18101	Disability Status	disability variables (aggregated)
B16005	Language Spoken at Home	foreign_born
B23025	Employment Status	in_labor_force, employed, unemployed, not_in_labor_force, unemployment_rate
B25064	Median Gross Rent	median_gross_rent

3.1.3. *Preprocessing*

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, 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.

After examination, a subset of highly correlated variables were removed. Measures of poverty level and population above the poverty line, as well as employment and unemployment percentages, were closely tied to broader income and labor force indicators already included in the model, creating redundancy without improving predictive power. Similarly, metrics describing commuting alone by car and the distribution of vehicle ownership (households with no vehicles, one vehicle, or multiple vehicles) were binned to prevent issues from being strongly interrelated. The precentage of female share of the population was excluded because of its near-perfect correlation with the male share of the population. The same was true for the share of high-income households (earning above $75,000), which closely overlapped with median income levels.

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).

3.2. *Key Metrics*

The primary risk metric, `crash_rate_per_1000`, measures the number of crashes per 1,000 residents in each census tract by 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).

This response variableis modeled alongside the socio-economic and transportation variables detailed in Table 1. 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).

3.3. *Modeling Approach*

To model the relationship between socio-economic characteristics and crash risk, we implemented a single gradient boosting framework using XGBoost.

XGBoost (Chen and Guestrin, 2016) 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).

Model performance was optimized with hypermater tuning using the automated Bayesian optimization framework Optuna (Akiba et al., 2019). This approach is supported by prior research showing that systematic hyperparameter optimization significantly improves boosting model accuracy (Liu et al., 2025).

Each configuration of Optuna tuning 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 approach was sufficient for our dataset size and feature set, yielding robust improvements in predictive accuracy.

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. Results

4.1. *Descriptive Statistics*

The dataset comprises 13,518 census tract–year observations from 2018 to 2023. Population counts vary widely across tracts, with a median of approximately 42,979 residents and extremes ranging from fewer than 100 to over 220,000. Demographically, the average tract population is predominantly White (mean 3.17%), Hispanic (mean 2.21%), and Black (mean 2.02%), though these percentages vary considerably across the five boroughs.

Table 2: Summary Statistics of Key Variables (Avg Across Tracts and Years)

Variable	Mean	SD	Max
Crash Rate per 1,000 People	1.43	1.83	23.29
Median Gross Rent	$1,613.24	$623.12	$3,501.00
Median Annual Income	$74,681.04	$40,140.23	$250,001.00
Percent Below Poverty	1.36%	1.05%	11.11%
Percent Commute Public Transit	3.72%	1.41%	12.31%
Percent Own a Vehicle	1.48%	0.66%	7.69%

Crash-related statistics show an average of 1.43 crashes per 1,000 residents, with the highest-risk tracts exceeding 23 crashes per 1,000 residents—nearly 20 times the citywide mean. Commuting behaviors show a multi-modal distribution, with public transit usage averaging 3.7%, but certain tracts—particularly in Manhattan—exhibit much higher pedestrian and transit activity.

Economic indicators reveal stark disparities. The median household income across all tracts and years is $74,681, but with a standard deviation of $40,140 and a maximum of $250,001, indicating substantial variation in neighborhood wealth. Housing costs are also highly variable, with a median gross rent of $1,589 and a maximum of $3,501.

Annual Mean Crash Rate per 1K Residents by Borough

Figure 1: Crash Rate by Borough, by Year

Crash Rate by Borough shows a clear downward trend in crash rates across all boroughs over time, with a sharp reduction during the COVID-19 pandemic period (2020) and gradual stabilization afterward. Bronx and Queens consistently show higher crash rates per 1,000 residents compared to Staten Island and Manhattan.
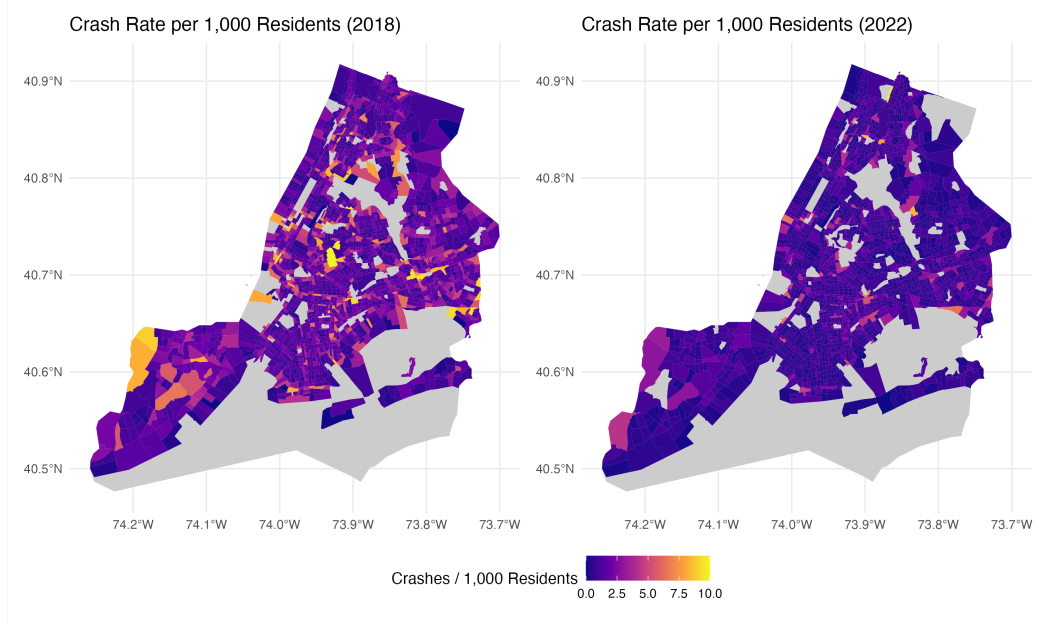
Figure 2: Crash Rate in 2018 vs 2022

The geospatial heatmaps illustrate how crash rates are spatially clustered within the city. In 2018, high crash rates were concentrated in central Brooklyn, the South Bronx, and sections of northern Manhattan, while in 2022, these hotspots persisted but appeared less intense overall, consistent with the downward temporal trend post-Covid-19 pandemic. By applying the same color scale across both maps, it is evident that most census tracts saw a reduction in crash intensity, although isolated high-risk corridors remain.

4.2. *Hyperperameter Tuning*

Optuna hyperparameter tuning (Akiba et al., 2019) enhanced the predictive accuracy and generalization capability of the XGBoost model. The final configuration represents a balance between model complexity and overfitting risk, as determined by performance on the training and validation subsets.

Table 3: Optimal Parameters as per Optuna

Action	Parameter	Value
Learning Rate	`eta`	1.549545
Tree Depth	`max_depth`	4
Row Sampling	`subsample`	0.5583975
Feature Sampling	`colsample_bytree`	0.8797697
Minimum Child Weight	`min_child_weight`	3

11

Action	Parameter	Value
Minimum Loss Reduction	gamma	3.35839
L2 Regularization	lambda	6.131609
L1 Regularization	alpha	2.805485

A tree depth of 4 means the model only splits features a few times before making predictions. The socio-economic and crash-rate features capture patterns that can be learned with relatively few decision boundaries.

This high learning rate is unusually high compared to common XGBoost defaults (0.01–0.3), indicating the model is taking larger steps while fitting each tree. The relatively small dataset (13,518 observations) allows a faster convergence without needing a very gradual learning rate. However, high `eta` combined with shallow trees requires careful monitoring for overfitting, which the regularization terms help mitigate.

The subsampling ratios (`subsample = 0.56`, `colsample_bytree = 0.88`) show the model uses just over half of the rows for each boosting iteration and nearly all features. This randomness prevents any one subset of data from dominating the model and improves generalization. This confirms the dataset has sufficient size and diversity to benefit from row sampling, but not so many redundant features that aggressive column sampling is needed.

Regularization parameters (`lambda = 6.13`, `alpha = 2.81`, `gamma = 3.36`) are relatively strong regularization values, indicating the model needed constraints to avoid overfitting. In practice, this means there is enough correlation and redundancy among features that despite removing highy correlated variables, the model benefits from being penalized for complex splits or large feature weights.

4.3. Model Performance and Diagnostics

The XGBoost model achieved robust predictive performance on the holdout test set, with the following metrics:

Table 4: XGBOOST Model Evaluation Metrics

Metric	Score
RMSE	1.549545
MAE	0.8372956
R^2	0.2595503

The XGBoost model's evaluation metrics suggest moderate predictive power with stable error bounds on the holdout test set. An RMSE of 1.55 indicates that the model's predicted crash rates deviate, on average, by roughly 1.5 crashes per 1,000 residents from observed values, with an MAE of 0.84 confirming that most prediction errors are below 1 crash per 1,000 residents. While the $R^2 = 0.26$ shows that the model explains only about 26% of the variance in crash rates across census tracts, this is consistent with the high degree of randomness and unobserved factors (e.g., driver behavior, weather) inherent in crash data.
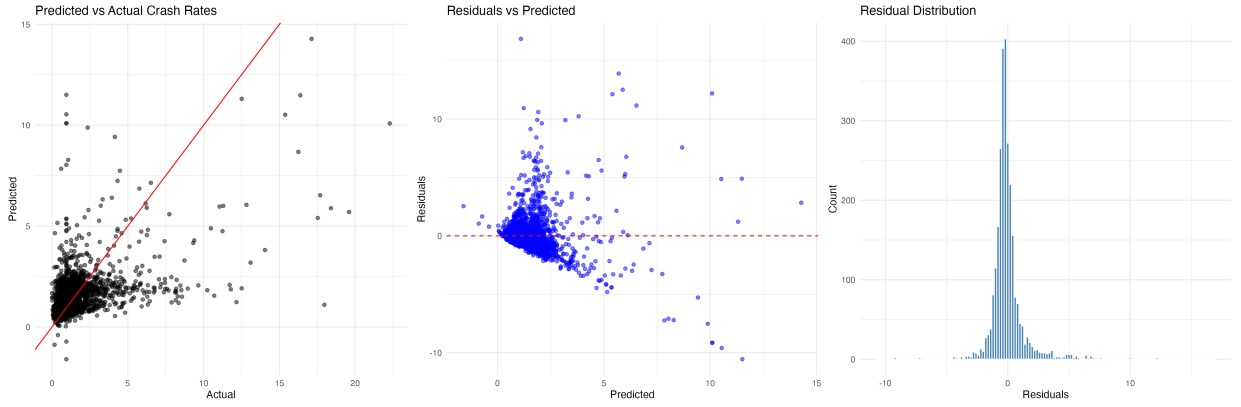

Figure 3: Residual Diagnostic Plots

4.3.1. *Examination of the Residuals*

The predicted vs. actual values plot shows that the model captures the general trend of observed crash rates across census tracts, with most predictions clustering closely around the 45-degree reference line. While a slight underestimation is evident for the highest crash-rate tracts, this is typical for gradient boosting models where extreme outliers are smoothed during ensemble averaging.

The residual density plot indicates residuals are centered near zero with a narrow peak, suggesting minimal systemic bias.

The residuals vs. predicted values plot shows a random scatter around zero, with no strong patterns of heteroskedasticity or underfitting, although a few outlier tracts exhibit residuals above ± 10.
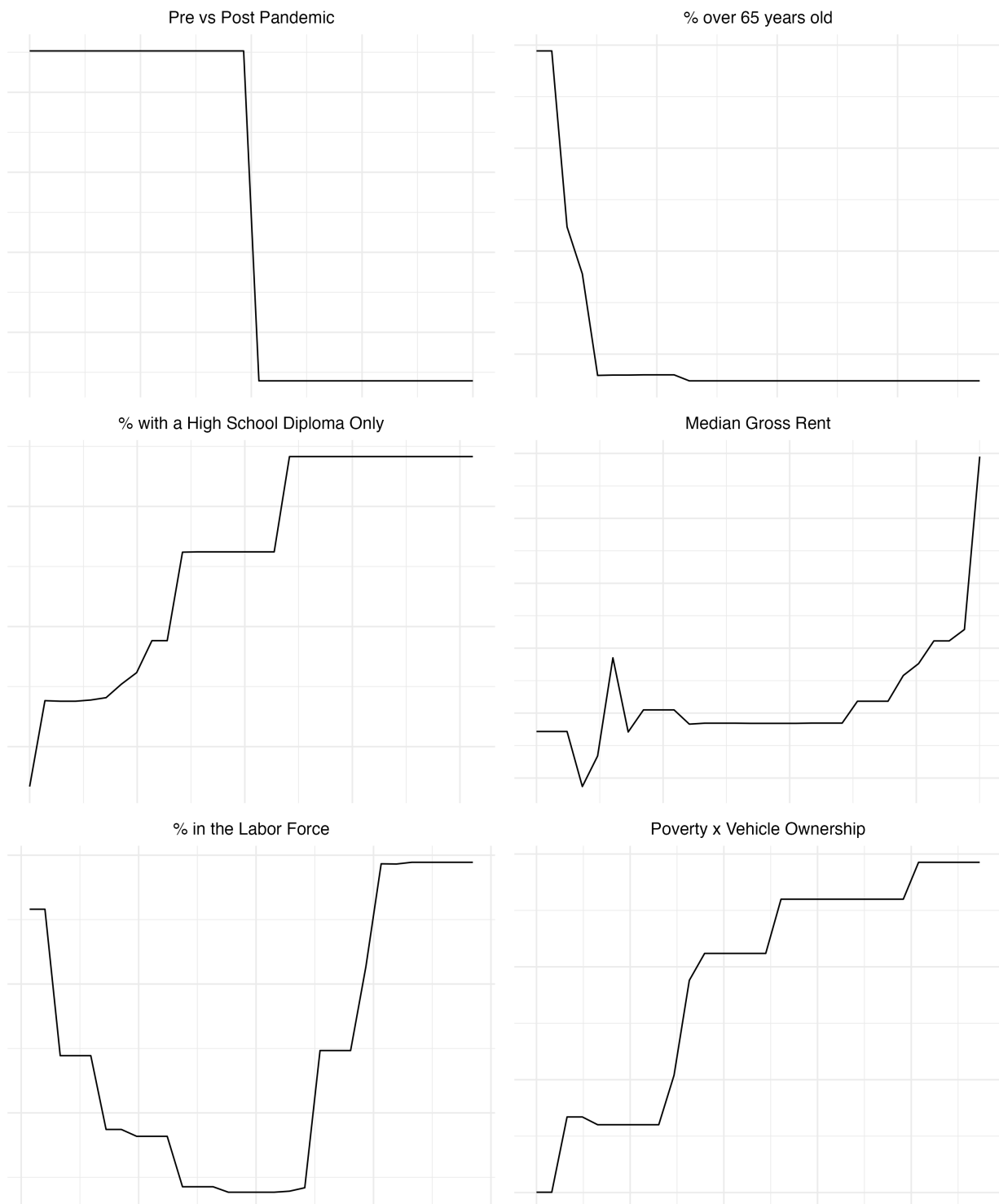
13

4.4. Global Feature Importance (SHAP)

Figure 4: Global Feature Importance with SHAP

The SHAP plot highlights the most influential features on crash risk predictions. The top six variables, in order of importance, are:

Post-Pandemic Indicator

The pre vs post Covid pandemic indicator variable hows a marked upward shift in predicted crash rates from 2020 onward, consistent with observed pandemic-era changes in traffic dynamics, also observed by Adeniyi (2024).

Aging Population

The PDP plot suggests that areas with a moderate share of elderly residents (10–20%) have slightly lower crash risks, potentially due to lower driving exposure. However, beyond 20%, predicted crash risk rises.

High School Graduate Population

For percent of the population with a high school diploma (but no further education), the PDP exhibits a U-shaped relationship: tracts with either very low or very high shares of residents holding a high school diploma are associated with higher crash risk.

Median Gross Rent

Median gross rent displays a positive gradient: tracts with higher rents—often denser and more urbanized—show higher crash rates.

Working Population

The PDP for percentage of the population in the workforce indicates that crash risk peaks around 60–70% labor force participation. Lower participation areas may have fewer commuters, while areas with higher participation rates may experience higher traffic volumes.

Poverty Rate x Vehicle Ownership Interaction

Th variable marking the interaction of vehicle ownership with poverty rate shows that high poverty rates combined with high vehicle ownership strongly elevate crash risk. This finding suggests that economically vulnerable drivers may face both infrastructure and behavioral risks.
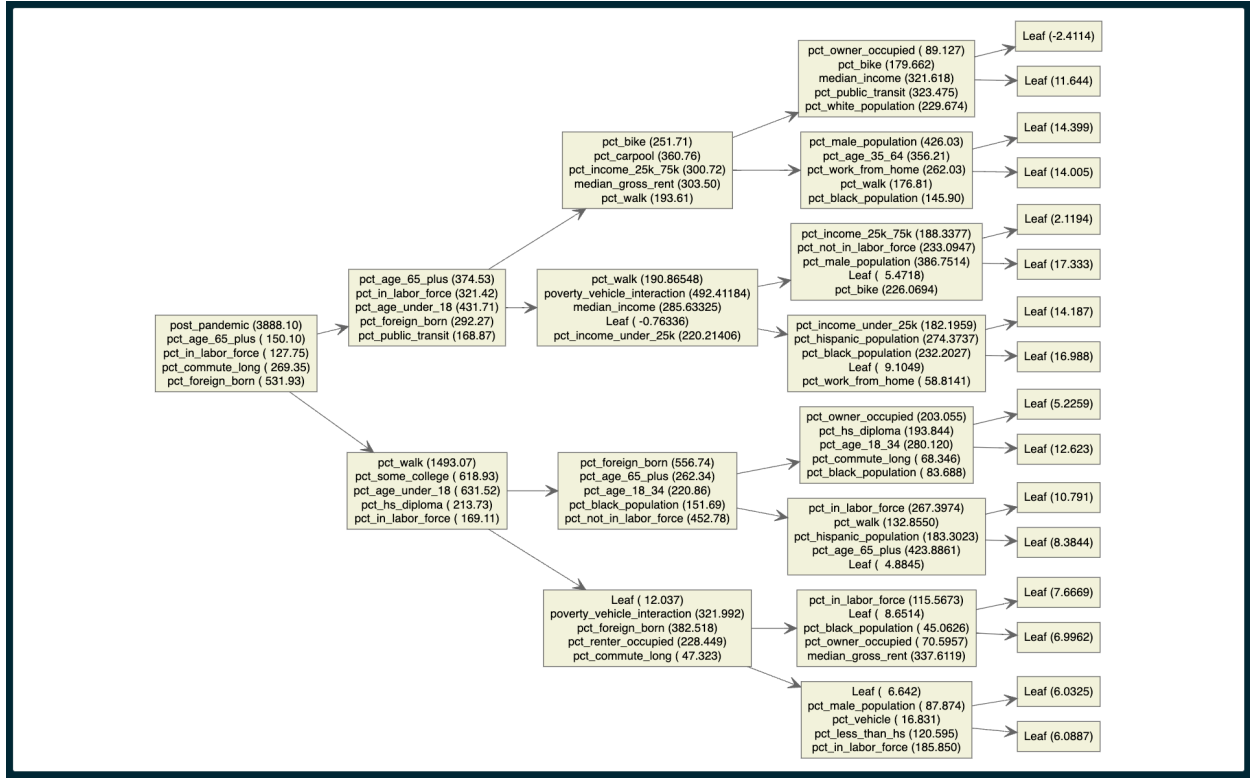
Figure 5: Figure X: Multi-Tree Plot

The multi-tree plot provides a detailed view of how the XGBoost model partitions the socio-economic and transportation variables to predict crash risk, revealing intricate interactions that are not visible in single-variable importance metrics. For insurers and policymakers, this tree-based view offers actionable insights into which combinations of conditions—such as low-income, high-commute areas with dense walking activity—are most predictive of risk.

At the top of the tree hierarchy, the model consistently splits on the post-pandemic indicator. Subsequent splits emphasize variables such as age distribution, labor force participation, and foreign-born population, all of which indicate that neighborhoods with specific demographic compositions face distinct patterns of risk exposure. For instance, branches that combine higher shares of elderly residents with elevated public transit use or lower workforce participation show heightened predicted crash rates.

Another prominent theme in the tree structure is the role of housing and economic indicators, such as median gross rent and income brackets, which appear alongside transportation variables. The multi-tree pathways involving poverty rate and vehicle ownership interaction are particularly illuminating, as they reveal how economic stress combined with high vehicle reliance amplifies crash likelihood. For example,

nodes where the model isolates high poverty rates, high vehicle ownership, and long commutes lead to leaves with some of the highest predicted crash values.

The branching structure also highlights the interplay between transportation modes and demographics. Variables like walking, biking, and working from home repeatedly appear as secondary splits, indicating that these behaviors act as modifiers of the primary socio-economic risk factors. Areas with high pedestrian activity combined with mid-range incomes or high percentages of young adults tend to diverge from the patterns observed in primarily car-dependent neighborhoods. This reflects the unique urban mobility landscape of New York City, where mixed transportation modes and densely populated corridors create more opportunities for vehicle-pedestrian conflicts.

5. Discussion

Although the model's R^2 value of approximately 0.26 indicates that it explains just over one-quarter of the variance in crash rates across New York City census tracts, this is a meaningful achievement considering the inherent randomness of crash events and the absence of individual-level data on drivers, vehicles, road conditions, weather, road design, etc. Auto collisions are influenced by many unobserved factors that cannot be captured through aggregated socio-demographic measures alone. Despite these constraints, the model's relatively low root mean square error (RMSE = 1.55 crashes per 1,000 residents) and mean absolute error (MAE = 0.84 crashes per 1,000 residents) suggest that it has successfully captured consistent, broad patterns in crash risk that align with socio-economic disparities and urban traffic dynamics.

From a social risk modeling perspective, these results are significant. The SHAP analysis and multi-tree plot reveal that the most influential predictors are not strictly transportation variables but rather socio-economic indicators that shape exposure and vulnerability. The post-pandemic indicator emerges as the single strongest factor, reflecting the profound shift in traffic patterns since 2020, when less congestion but higher average speeds upended previous risk models by changing nearly all traffic conditions, especially in densely populated urban areas like NYC.

Housing and economic variables also feature prominently. Median gross rent is positively correlated with crash rates, suggesting that denser, higher-cost urban areas within NYC, with more vehicle-pedestrian interactions and complex traffic flows, have elevated risk levels. Similarly, labor force participation peaks as a predictor around 60–70%, reflecting that areas with higher commuter activity experience more traffic and thus more collisions.

One of the most striking findings is the role of poverty rate and vehicle ownership interaction, which appears consistently in the multi-tree plot. Areas with both high poverty rates and high vehicle availability

17

show notably higher crash risks, likely due to a combination of older vehicles, reduced access to safety infrastructure, and potentially riskier driving environments.

While the model demonstrates that neighborhood-level demographic and economic factors can act as strong proxies for crash risk, it also highlights critical limitations for deployment in insurance pricing. The moderate R^2 indicates that a significant portion of risk remains unaccounted for, largely due to unobserved individual-level factors such as driver age, prior claims, and vehicle safety features. Moreover, the use of variables like income, race, or foreign-born population—though statistically predictive—would be highly problematic if directly incorporated into premium calculations due to both legal prohibitions and ethical concerns. These variables risk proxy discrimination, where protected classes are indirectly penalized through correlated socio-economic attributes.

The model answers the primary research question-what socio-economic variables are most strongly predictive of auto insurance risk. It shows that the environmental and demographic facotrs can explain spatial patterns of crashes and provide insurers or policymakers with macro-level risk insights. However, the model cannot, and should not, be used as-is for individual risk assessment. It is better suited for regional portfolio analysis, identifying high-risk neighborhoods for targeted safety interventions, or supplementing traditional actuarial models rather than replacing them.

6. Conclusions and Future Work

This research demonstrates that gradient boosting, combined with socio-economic and crash data, can provide interpretable, data-driven insights into auto insurance risk patterns in NYC. By identifying key drivers of crash risk—such as post-pandemic shifts, commuting intensity, and socio-economic vulnerability— the model offers a foundation for both urban safety planning and high-level risk assessment.

The study's findings also underscore the limitations of using aggregated, public data for underwriting. While the model captures broad trends, it lacks the precision, granularity, and fairness safeguards required for production-level insurance applications. Therefore, the model is best suited as a complementary tool for insurers, providing neighborhood-level risk insights or supporting reinsurance and portfolio management rather than individual pricing.

Future research should focus on three fronts: (1) integrating behavioral data, such as telematics, and weather patterns-among other possible additions-to bridge the gap between macro-level socio-economic patterns and micro-level driving behavior; (2) developing fairness-aware modeling approaches to mitigate bias from socio-economic proxies; and (3) exploring temporal extensions that incorporate evolving risk factors, including post-pandemic traffic patterns and climate-related hazards. These directions will help transition

from descriptive social risk modeling to actionable, ethically sound insurance applications.

7. Appendix A: Variables Modeled

Table 5: Key variables, descriptions, and transformations in the final dataset.

Variable	Description	Type	Transformation
Demographic	pct_male_population	Men	Percentage
Demographic	pct_white_population	Identifying as white	Percentage
Demographic	pct_black_population	Identifying as black	Percentage
Demographic	pct_asian_population	Identifying as Asian	Percentage
Demographic	pct_hispanic_population	Identifying as Hispanic/Latino	Percentage
Demographic	pct_foreign_born	Foreign-born	Percentage
Age	pct_age_under_18	Under 18	Percentage
Age	pct_age_18_34	Aged 18-34	Percentage
Age	pct_age_35_64	Aged 35-64	Percentage
Age	pct_age_65_plus	Aged 65 and above	Percentage
Income/Poverty	median_income	Median household income (inflation-adjusted)	Raw value (USD)
Income/Poverty	pct_income_under_25k	Households earning less than $25,000	Percentage
Income/Poverty	pct_income_25k_75k	Households earning $25,000-$75,000	Percentage
Income/Poverty	pct_below_poverty	Below the poverty line	Percentage
Housing	median_gross_rent	Median gross rent (USD)	Raw value (USD)
Housing	pct_owner_occupied	Owner-occupied housing units	Percentage
Housing	pct_renter_occupied	Renter-occupied housing units	Percentage
Education	pct_less_than_hs	Less than high school education	Percentage
Education	pct_hs_diploma	High school diploma	Percentage
Education	pct_some_college	Some college education	Percentage
Education	pct_associates_degree	Associate's degree	Percentage
Education	pct_bachelors_degree	Bachelor's degree	Percentage
Education	pct_graduate_degree	Graduate or professional degree	Percentage
Employment	pct_in_labor_force	In the labor force	Percentage
Employment	unemployment_rate	Unemployment rate	Percentage
Transport	pct_commute_short	Commute under 15 minutes	Percentage

Variable	Description	Type	Transformation
Transport	pct_commute_medium	Commute between 15-30 minutes	Percentage
Transport	pct_commute_long	Commute longer than 30 minutes	Percentage
Transport	pct_carpool	By carpool	Percentage
Transport	pct_public_transit	By public transit	Percentage
Transport	pct_walk	By walking	Percentage
Transport	pct_bike	By biking	Percentage
Transport	pct_work_from_home	Working from home	Percentage
Transport	pct_vehicle	Owns a vehicle	Percentage
Engineered	post_pandemic	Post-pandemic indicator (1 = 2020 and later)	Binary
Engineered _interaction	poverty_vehicle Interaction term: poverty rate × vehicle ownership	Interaction	
Engineered _interaction	unemployment_vehicle Interaction term: unemployment rate × vehicle ownership	Interaction	
Year	year2018	Year dummy: 2018	Indicator
Year	year2019	Year dummy: 2019	Indicator
Year	year2020	Year dummy: 2020	Indicator
Year	year2021	Year dummy: 2021	Indicator
Year	year2022	Year dummy: 2022	Indicator
Year	year2023	Year dummy: 2023	Indicator

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.

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. ACM. URL: https://doi.org/10.1145/3292500.3330701.

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.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. ACM. URL: https://doi.org/10.1145/2939672.2939785.

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.