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, LightGBM, 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 [11, [7]]. Predicting claim frequency and severity not only supports pricing but also enables insurers to manage portfolio-level risk and optimize resource allocation [16].

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 [6, [4]]. 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 [1, [5]].

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 [11]. 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 [7]. These limitations are particularly pronounced in urban contexts, where crash risk is shaped by heterogeneous population dynamics and localized factors [6, [5]].

[*] Corresponding author
Email address: ajstraumanscott@pm.me (AJ Strauman-Scott)

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 [10, [3]]. 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 [7, [16], [3]]. 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 [15, [8], [17]]. 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 [13].

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 [11, [16]], while systematic reviews highlight that few studies combine open crash data with socio-economic indicators in insurance modeling [2, [3]].

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 [7] 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, [12] 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 [11]. 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. [7] applied gradient boosting to both claim frequency and severity modeling, demonstrating significant performance gains in frequency prediction over Poisson-based GLMs. Similarly, [12] 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 [18]. 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, [1] 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, [8] 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. [5] 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. [6] 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. [1] 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. [9], 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). [11] 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 [15], 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 [16] 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. [8] 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, [17] 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 [7, [11], [12]]. 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 [1, [8]].

Crash frequency is aggregated at the borough and neighborhood level and normalized by population to calculate crashes per 1,000 residents as a proxy for claim frequency [5]. Injury and fatality counts are used to compute injury-to-fatality ratios, which serve as a proxy for claim severity, reflecting the expected cost burden of accidents [7].

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

Socio-economic variables are drawn from the ACS 5-year estimates (2018–2022) at the Neighborhood Tabulation Area (NTA) level, ensuring consistency with NYC's geographic units for urban planning. Variables include: Median income, age distribution, and educational attainment. Mode share (car, subway, walking), average commute times, and vehicle ownership rates. Household size, population density, and residential occupancy patterns. These features have been shown to correlate with traffic risk and claim likelihood, as evidenced by [5] and [6], who found that socio-economic deprivation and commuting behaviors strongly influence crash frequency and severity.

Preprocessing Steps

Crash records are spatially joined to ACS NTAs using NYC Open Data shapefiles. Derived metrics include per-capita crash frequency, vehicle ownership ratios, and log-transformed density variables to capture non-linear scaling effects [6]. Outliers (e.g., extreme crash counts from anomalous events) are removed using interquartile range thresholds. Missing socio-economic values are imputed using median imputation.

Categorical variables (e.g., predominant commute mode) are one-hot encoded, and continuous variables are standardized for input into gradient boosting models [11].

This integrated dataset allows us to examine how socio-economic conditions interact with crash patterns to inform insurance claim frequency and severity modeling, addressing a gap in the literature where these public data sources have rarely been combined [16, [12]].

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 [1, [8]].

The metric crashes per 1,000 residents is employed to approximate claim frequency, following the approach of studies that normalized accident counts by population density to assess relative risk across regions [5, [6]].

This population-adjusted measure accounts for differences in exposure between neighborhoods or boroughs, enabling fair comparisons of risk levels.

To estimate claim severity, we compute injury-to-fatality ratios based on reported crashes, reflecting the likelihood and severity of bodily harm resulting from incidents. This metric is consistent with the severity-focused modeling frameworks used in both insurance [7, [11]] and traffic safety studies [8], where injury severity is treated as a key determinant of financial impact.

3.3. 3.3 Modeling Approach

To capture complex, non-linear relationships between socio-economic variables (e.g., income, commuting patterns) and crash risk, we use gradient boosting algorithms, specifically XGBoost and LightGBM. Gradient boosting has consistently outperformed traditional GLMs and linear models in insurance claim prediction and crash severity modeling due to its ability to handle heterogeneous data types (categorical and continuous) and capture feature interactions [7, [16]].

XGBoost is chosen for its strong track record in insurance risk modeling and interpretability when combined with SHAP [8], while LightGBM is included for its efficiency on large datasets and superior performance on high-cardinality categorical variables. This dual-model approach aligns with studies comparing boosting frameworks for both frequency-severity modeling [11] and urban crash prediction [1].

To optimize model performance, hyperparameter tuning is conducted using Optuna, a state-of-the-art optimization framework leveraging Bayesian search and pruning strategies to efficiently navigate large hyperparameter spaces. This approach is supported by prior research showing that systematic hyperparameter optimization significantly improves boosting model accuracy [8, [14]].

3.4. 3.4 Explainability

Given the regulatory and operational need for transparent, explainable models in insurance [11, [15]], 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

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.