Predicting Auto Insurance Risk Using Gradient Boosting
Analyzing Socio-Economic Factors in Car Crashes 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 census tract-by-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 ($R^2 = 0.425$), 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.

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 supports pricing and 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-

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

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, 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).

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 measuring social risk for automotive insurance and urban policy. The ultimate goal is to identify key socio-economic, transportation, and demographic predictors that drive claim frequency, and to determine how much crash risk can be predicted solely by these factors using an XGBoost model, without incorporating weather, road pattern or individual-level driver data.

The remainder of this paper is organized as follows: Section 2 reviews prior work on ML in insurance risk modeling, 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, 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.

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.

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 partial dependence plots (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.

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.

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). This gap limits the development of robust, regionally sensitive models that capture the real-world interaction between crash frequency and socio-economic indicators.

3. Materials and Methods

3.1. Data Sources

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. 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 , age distribution, and income indicators. Additional features include median gross rent, housing tenure, educational attainment, employment metrics, and transportation factors. The full table of ACS derived variables in available in Section 7.

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 variable is modeled alongside the transformed and selected socio-economic and transportation variables detailed in Section 8. 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).

3.2.1. ***Preprocessing***

Crash records from the NYC Open Data MVC dataset are 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 annual per-capita crash rates for each census tract in NYC for each year.

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-poverty rate and vehicle ownership, as well as unemployment rate and vehicle ownership-were engineered to capture compound effects on risk exposure.

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 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, which closely overlapped with median income levels.

No categorical encoding besides `year` or feature standardization was performed, as all ACS predictors are expressed as continuous percentages or numeric values, and gradient boosting models (XGBoost) handle raw scales effectively (Henckaerts et al., 2021). However, the response variable - number of crashes per 1,000 people - was log-transformed to stabilize variance and reduce the impact of extreme values.

3.3. Modeling Approach

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.

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.

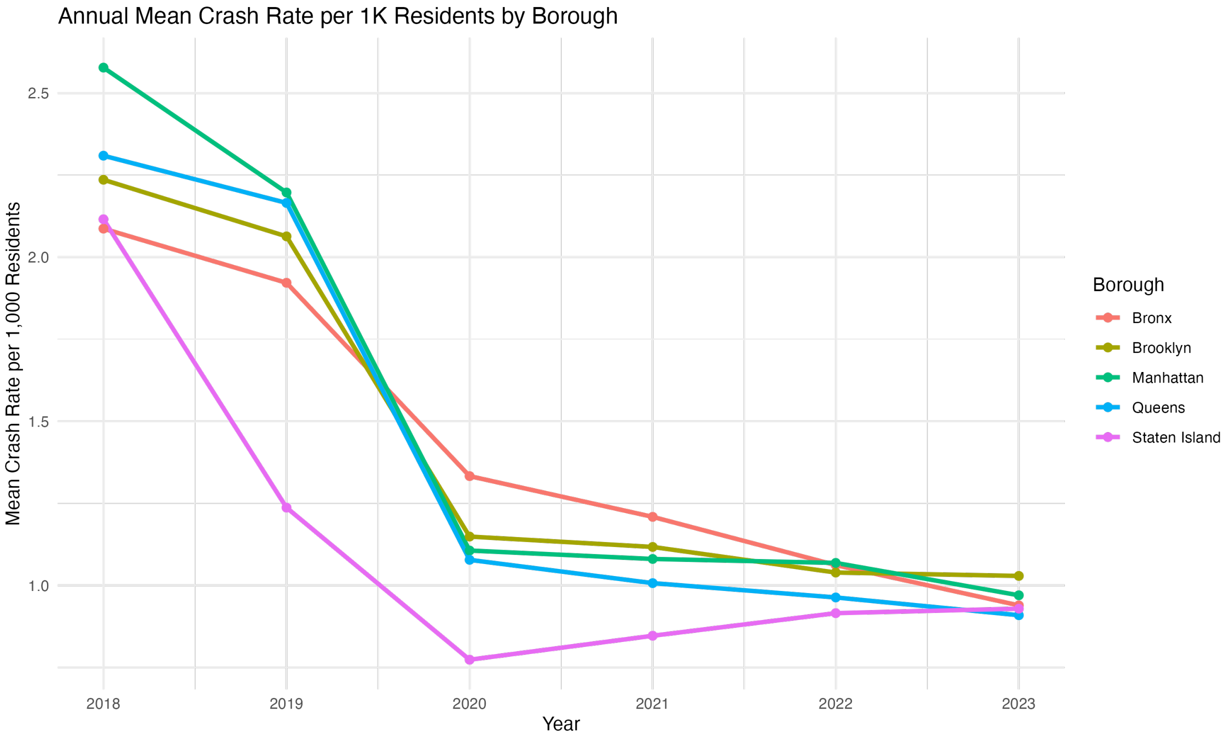

Figure 1: Crash Rate by Borough, by Year

In Figure 1, 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.

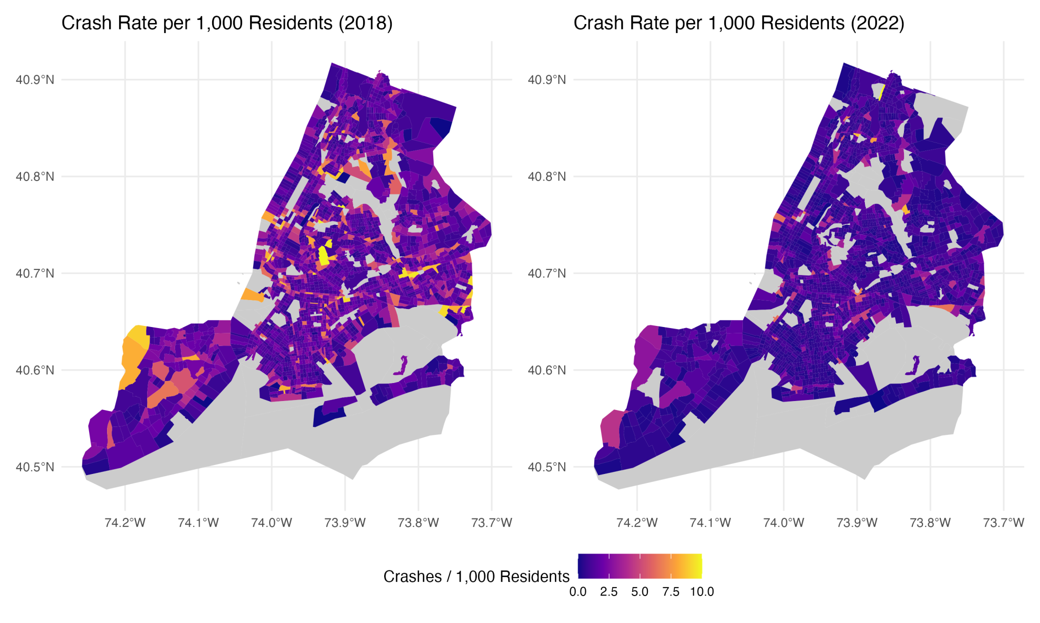
Crash Rate per 1,000 Residents (2018)
Crash Rate per 1,000 Residents (2022)
40.9°N
40.8°N
40.7°N
40.6°N
40.5°N
74.2°W
74.1°W
74.0°W
73.9°W
73.8°W
73.7°W
Crashes / 1,000 Residents
0.0
2.5
5.0
7.5
10.0

Figure 2: Crash Rate in 2018 vs 2022

The geospatial heatmaps in Figure 2 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 across both total and commuting population post-Covid-19 pandemic. 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 1: Optimal Parameters as per Optuna

Action	Parameter	Value
Learning Rate	`eta`	0.2586295
Tree Depth	`max_depth`	12
Row Sampling	`subsample`	0.9921435
Feature Sampling	`colsample_bytree`	0.5823858

Action	Parameter	Value
Minimum Child Weight	`min_child_weight`	4
Minimum Loss Reduction	`gamma`	0.08316905
L2 Regularization	`lambda`	4.260907
L1 Regularization	`alpha`	0.2049112

The tree depth of 12 allows the model to make much deeper splits, capturing highly granular interactions between socio-economic and crash-related variables. While deep trees can increase the risk of overfitting, this is balanced by the other constraints in the model. The minimum child weight places a threshold on the minimum sum of instance weights required to create a split. This prevents the model from building branches that explain only a small subset of observations, helping to maintain focus on broader patterns in the data.

The learning rate is moderate, enabling the model to update predictions in steady increments without being overly aggressive. This value strikes a balance between convergence speed and stability.

The row sampling rate (`subsample = 0.992`) indicates that nearly all rows are used in each boosting iteration, which reduces variance but increases the risk of memorizing noise in the data. In contrast, the feature sampling rate (`colsample_bytree = 0.582`) introduces meaningful randomness by selecting slightly more than half of the features for each tree.

Regularization parameters, `lambda` (L2) and `alpha` (L1), apply moderate constraints to the feature weights, discouraging overly complex or extreme splits without overly penalizing flexibility. The minimum loss reduction (`gamma` = 0.083) is relatively low, which allows the model to explore more potential splits during training. This can enhance the model's ability to capture subtle relationships between features.

4.3. ***Model Performance***

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

Table 2: XGBOOST Model Evaluation Metrics

Metric	Score
RMSE	0.3396146
MAE	0.2422538
R^2	0.425285

These metrics indicate solid predictive performance with tight error bounds relative to the normalized crash rate values. An RMSE of 0.34 suggests that predictions deviate by approximately 0.34 crashes per 1,000 residents on average, which is a low margin of error given the variability across census tracts. The MAE of 0.24 confirms that typical prediction errors remain well below half a crash per 1,000 residents.

Most critically, the R^2 value of 0.43 means the model explains approximately 43% of the variance in crash rates, a strong result for a model built only on socio-economic and transportation predictors, without direct behavioral or vehicle-level data.

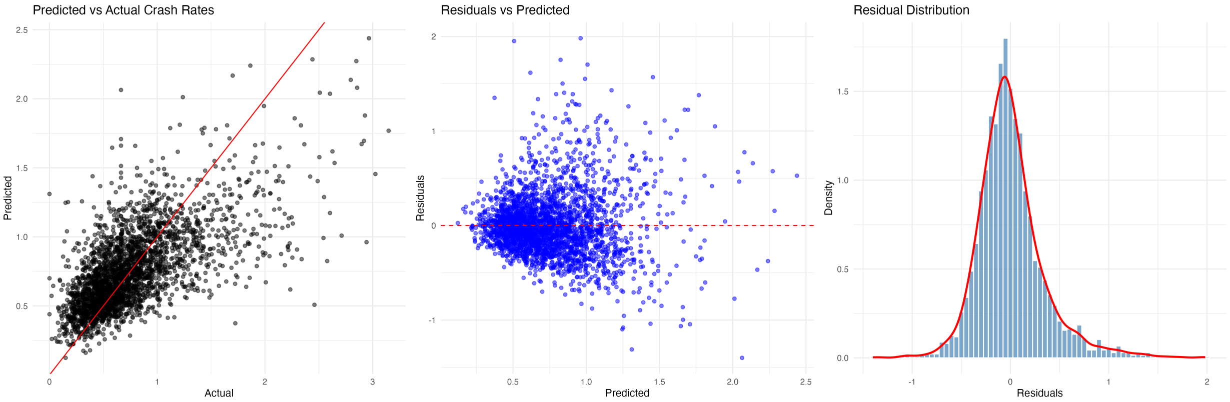

Figure 3: Residual Diagnostic Plots

4.3.1. ***Residuals***

The predicted vs. actual values plot shows that while the model generally tracks the trend of observed crash rates, there is a clear cone-shaped spread. This indicates that predictions tend to be compressed toward the mean—with underestimation for high crash-rate tracts and overestimation for low crash-rate tracts. This is a common behavior for gradient boosting models trained on noisy data, where extreme values are smoothed due to ensemble averaging.

The residuals vs. predicted values plot also displays the same cone shape. This pattern suggests variance increases at the extremes, meaning the model is less confident and less accurate when predicting very high or very low crash rates.

The residual distribution plot remains centered around zero with a sharp peak, showing minimal systemic bias.

4.4. ***Feature Importance (SHAP)***

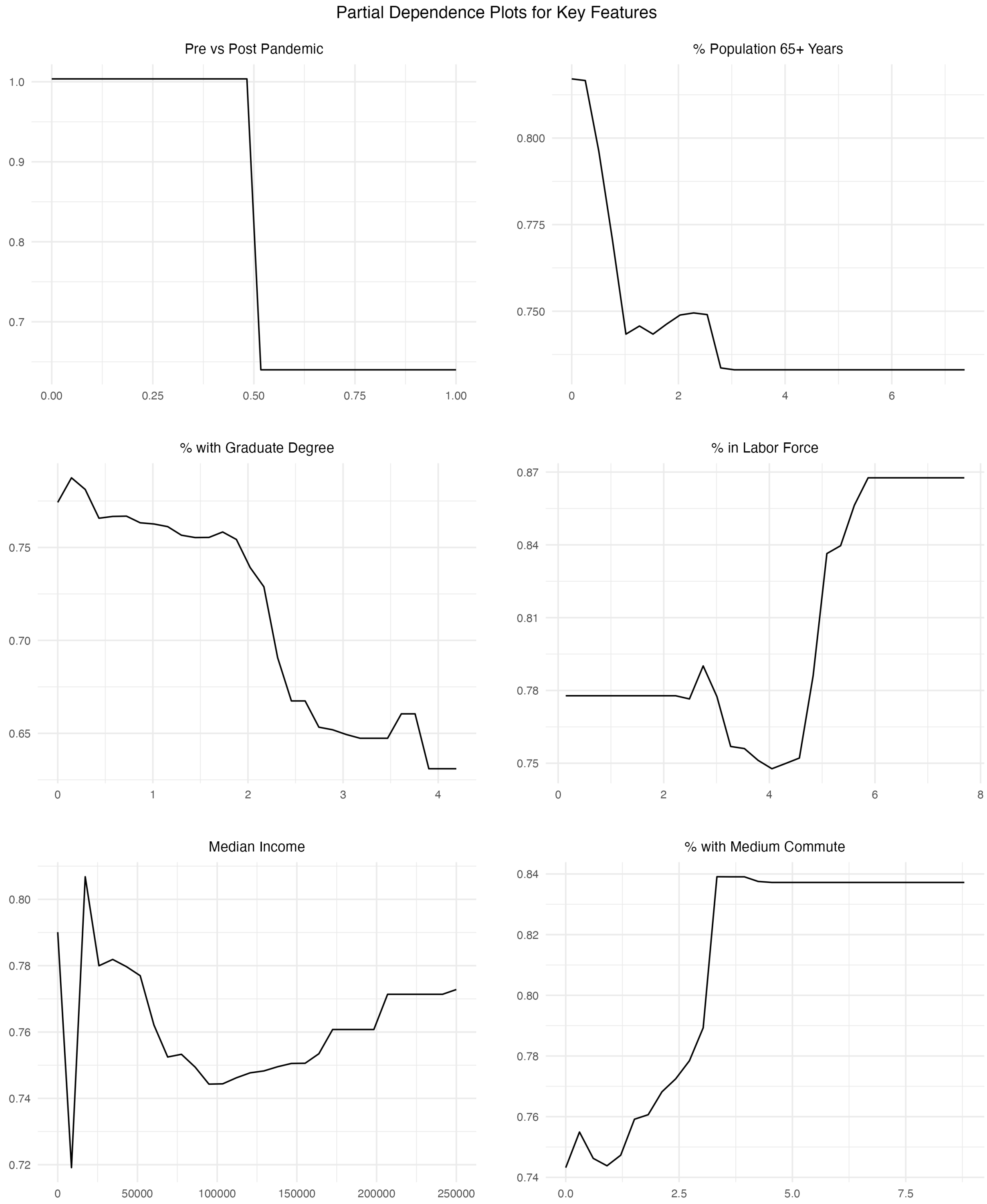

Figure 4: Global Feature Importance with SHAP

The SHAP plot highlights the six most influential variables shaping crash risk predictions in the XGBoost model. These features reveal complex, non-linear relationships with crash rates:

Post-Pandemic Indicator

The pre vs. post-COVID-19 variable shows a sharp drop in predicted crash rates after the pandemic began, consistent with the reduction in traffic volumes and changes in driving patterns observed citywide, also observed by Adeniyi (2024).

Aging Population

Crash risk declines as the proportion of residents aged 65 and older increases up to roughly 2%, likely due to reduced driving exposure among elderly populations. Beyond this point, the trend stabilizes, suggesting diminishing marginal effects.

Graduate Degree Share

The percentage of residents with graduate degrees exhibits a steady downward relationship with crash risk. Higher educational attainment may correlate with neighborhoods that have better infrastructure or lower exposure to high-risk driving.

Working Population

The PDP for labor force participation indicates a non-linear trend. Crash risk decreases slightly between 2% and 5% workforce participation but increases sharply as participation exceeds 5%, reflecting higher commuter activity and vehicle miles traveled.

Median Income

Median income shows a non-monotonic relationship: risk is elevated in very low-income areas, dips in middle-income neighborhoods, and then rises again at high-income levels, possibly reflecting denser, high-cost urban areas with complex traffic dynamics.

Medium Commute Share The percentage of residents with medium-length commutes (15–30 minutes) is strongly associated with higher crash risk once it exceeds approximately 3%. This likely reflects regions with heavier traffic flow and greater exposure to collisions.

5. Discussion

The XGBoost model achieved a notable performance with an R^2 of 0.43, meaning it explains over 40% of the variance in crash rates across New York City census tracts. This is a strong result given the absence of driver-level or vehicle-level data and the inherently stochastic nature of traffic collisions. The low RMSE (0.34 crashes per 1,000 residents) and MAE (0.24) indicate that the model's predictions are generally close to observed values, with small, consistent errors across most census tracts.

From a social risk modeling perspective, the results highlight the importance of broader socio-economic and transportation-related conditions in shaping crash risk. The SHAP analysis reveals that the post-pandemic indicator is the single strongest predictor, showing a significant shift in traffic patterns starting in 2020. Reduced congestion but increased vehicle speeds during and after the pandemic contributed to changes in crash risk that conventional actuarial models might fail to capture.

Key socio-economic variables also play prominent roles. Median income exhibits a non-linear relationship with crash risk, with both very low-income and high-income areas seeing higher risk—likely due to a mix of infrastructure disparities and dense traffic flows. Labor force participation and commute-related metrics, particularly the share of medium-length commutes, are also strongly predictive, indicating that commuter activity and roadway exposure remain fundamental drivers of crash frequency. The negative association between graduate degree share and crash risk may reflect safer traffic environments or reduced reliance on driving in these neighborhoods.

These findings emphasize the interplay between economic conditions, urban infrastructure, and mobility behaviors in shaping neighborhood-level crash risks. High-income and low-income areas may share common risk factors, such as higher vehicle ownership rates, complex traffic flows, or limited access to safe pedestrian infrastructure, albeit for different reasons. Similarly, commute-related variables underscore the importance of traffic volume and duration as core determinants of crash likelihood, suggesting that transportation policies—such as investment in public transit or traffic-calming measures—could meaningfully reduce risks

The cone-shaped residual patterns suggest that while the model performs well across the midrange of crash rates, it tends to compress predictions toward the mean—underestimating high-risk tracts and overestimating low-risk ones. This limitation is typical for tree-based ensembles trained on data with sparse extremes, but it indicates an opportunity for refinement through log-transformed targets or quantile regression techniques.

Although the model provides actionable insights into spatial risk patterns, its direct use for individual insurance pricing is neither appropriate nor ethical. Socio-economic and demographic factors, while predictive, are not permissible as rating factors due to potential for proxy discrimination. Instead, these findings are better suited for identifying high-risk areas for targeted safety interventions, urban planning efforts, or high-level portfolio analysis.

6. Conclusions and Future Work

This study demonstrates that gradient boosting models, when combined with open crash and socio-economic data, can produce accurate and interpretable models of neighborhood-level crash risk in New York City. By identifying factors such as post-pandemic traffic changes, income patterns, educational attainment, and commuting behaviors, the model offers a rich, data-driven understanding of how socio-economic context influences crash frequency.

However, the model's focus on aggregated census tract data and its moderate $R^$2 highlight the limits of what can be achieved without driver-specific or telematics data. While it effectively captures macro-level trends, it is not a replacement for traditional actuarial models but rather a complementary tool for exploring spatial and demographic risk factors.

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: ACS Variables

Table 3: ACS tables 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
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)

ACS	Description	Derived.Variables
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
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

8. Appendix B: Variables Modeled

Table 4: 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

Variable	Description	Type	Transformation
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
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

Variable	Description	Type	Transformation
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.

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.