University of Central Florida

Department of Economics

Business Analytics, M.S

Professor: Dr. Harry J. Paarsch

Name: Pooja

Modeling Injury Severity in NYC Traffic Accidents for Insurance Risk

July 2025

1 Introduction and Motivation

New York City has experienced a substantial increase in auto insurance premiums, with year-over-year increases exceeding $750 in 2024, driven primarily by rising crash severity and associated claim costs. This surge reflects broader national trends where the average cost of full coverage car insurance reached $2,543 in 2024, marking a 26 percent annual increase that far exceeds the general inflation rate of 2.9 percent. The primary drivers include escalating vehicle repair expenses and critically, worsening accident severity patterns, as evidenced by bodily injury claims increasing 9.2 percent in the same period.

At the heart of this insurance affordability crisis lies the fundamental challenge of injury severity prediction. Unlike accident frequency, the severity of consequences when crashes occur remains poorly understood and inadequately modeled. This gap is particularly problematic because severe injuries and fatalities drive the majority of insurance claim costs, with fatal accidents carrying economic impacts exceeding $1.8 million per incident according to U.S. Department of Transportation estimates.

From a public safety perspective, New York City's Vision Zero initiative—launched in 2014 with the ambitious goal to eliminate all traffic deaths and serious injuries by 2024—has struggled to achieve its targets despite significant policy efforts and infrastructure investments. The disconnect between policy intentions and outcomes reveals fundamental limitations in current approaches to understanding and predicting crash severity. Traditional safety interventions often rely on historical accident frequency data and basic demographic factors, failing to capture the complex interactions between environmental conditions, driver behaviors, infrastructure characteristics, and temporal patterns that drive severity outcomes.

This study addresses these challenges by developing a comprehensive machine-learning framework for predicting traffic accident injury severity in New York City. I implement a multiclass classification approach that distinguishes between Non-Severe, Severe, and Fatal outcomes.

Since these prediction models will be used internally by insurance companies and government agencies for risk assessment and policy planning, the Lucas Critique is weakened because individual drivers remain unaware of the specific algorithmic factors, preventing behavioral changes that could invalidate the model. The research objectives are threefold: to develop an accurate and interpretable machine-learning model capable of predicting traffic accident severity, to identify and quantify the key factors driving severity outcomes, and to translate model insights into practical recommendations for insurance industry pricing strategies and public safety policy implementation.

2 Literature Review

2.1 H.M. Abdul Aziz, Satish V. Ukkusuri, and Samiul Hasan (2013)

In "Using Random-Parameter Logit Models to Explore Crash Severity in NYC," the authors modeled each borough separately to capture local effects and found variable impacts of gender, parking density, and street features. Their research demonstrated that Manhattan's dense urban environment produces different severity patterns compared to outer boroughs like Queens and Brooklyn.

This suggests modeling each borough separately or including location-based interaction terms. It highlights the predictive relevance of driver demographics and road speed conditions, supporting my decision to include borough-level effects and geographic variables in the severity prediction model.

2.2 Liran Einav, Amy Finkelstein, and Jonathan Levin (2022)

In "Examining How Real-World Telematics Data Affects Driver Behavior and Risk Pricing," Einav, Finkelstein and Levin examined real-world telematics data and found that drivers reduced risky behavior by approximately 30 percent when monitored. Their analysis showed that awareness of being tracked encouraged safer driving habits, leading to fewer and less

severe crashes.

The research validates the effectiveness of behavioral monitoring in insurance design and underscores the strategic value of real-time data in refining pricing models. The research demonstrates the importance of driver behavior factors in crash severity prediction and supports the incorporation of behavioral indicators—such as distraction or intoxication—into injury severity modeling for both risk assessment and incentive-aligned insurance strategies.

2.3 Md. Asif Khan Rifat, Ahmedul Kabir, and Armana Sabiha Huq (2024)

In "Comparing ML Models for Fatality Prediction in Crash Data," Rifat et al. conducted a comprehensive comparison of machine learning algorithms for predicting traffic fatalities using crash data from multiple jurisdictions. The study evaluated various models including Random Forest, XGBoost, and LightGBM for handling the prediction task. LightGBM achieved the highest performance metrics. The researchers employed SHAP (SHapley Additive exPlanations) for model interpretability, validating that temporal factors (time of day), spatial factors (location characteristics), and infrastructure elements (road class) emerged as the most important predictors of fatal outcomes.

This provides direct methodological validation for my research approach. The confirmed effectiveness of the LightGBM + SHAP pipeline supports my model selection strategy. Additionally, their emphasis on model interpretability through SHAP analysis aligns with the need for actionable insights in severe crash prediction, providing direct support for my choice of interpretable machine learning methods in safety-critical applications.

2.4 Kenny Santos, João P. Dias, and Conceição Amado (2021)

In their "Comprehensive Literature Review of ML Algorithms for Crash Injury Severity Prediction," Santos et al. conducted a systematic review analyzing 56 peer-reviewed studies published between 2001-2021 focusing on machine learning applications in crash injury

severity prediction. The meta-analysis revealed that ensemble methods, particularly Random Forest and gradient boosting variants (XGBoost, LightGBM), consistently outperformed individual classifiers across diverse datasets and geographic contexts. The review identified that most successful studies achieved AUC scores between 0.65-0.80, with ensemble methods showing superior performance in handling complex feature interactions and non-linear relationships inherent in crash severity data. The analysis also highlighted the importance of feature engineering and the growing trend toward interpretable machine learning approaches in transportation safety research.

The comprehensive literature review provides crucial context for interpreting my model performance results and validates my methodological choices. The finding that ensemble methods consistently achieve AUC scores in the 0.65-0.80 range for crash severity prediction establishes realistic performance benchmarks for evaluating my results. The review's emphasis on feature engineering importance supports my comprehensive approach to creating behavioral, temporal, and environmental features for severity prediction.

2.5 Jungsoo Lee, Taehoon Yoon, Seonho Kwon, and Jangmyung Lee (2020)

In their "Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning," the authors analyzed nine years of traffic accident data from Seoul, Korea, specifically focusing on the impact of weather conditions on crash severity during rainy seasons. The study systematically evaluated multiple machine learning algorithms including Decision Trees, Random Forest, SVM, and Neural Networks for predicting accident severity outcomes. Their comprehensive feature set included meteorological variables (rainfall intensity, wind speed, visibility), temporal factors (time of day, season), and infrastructure characteristics (road geometry, surface type). The research revealed that rainfall intensity and road geometry characteristics emerged as the top predictive factors, with Random Forest models achieving the highest accuracy when environmental variables were properly integrated into the feature space.

The study provides direct empirical validation for my decision to incorporate NOAA weather data as core predictive features rather than treating them as control variables. Their finding that meteorological factors rank among the most important predictors strongly supports my approach to treat weather conditions as primary severity determinants. The study's methodology of integrating hourly weather observations with crash records directly parallels my data integration approach, providing methodological validation for the environmental feature engineering strategy employed in my research.

3 Theoretical Framework

3.1 Economic Theory

The Value of Statistical Life (VSL) framework provides additional economic foundation for quantifying severe injury impacts: $VSL = \frac{\Delta WTP}{\Delta Risk}$, where ΔWTP represents additional willingness to pay for safety improvements and $\Delta Risk$ represents the reduction in injury or fatality risk. The insurance pricing framework that emerges can be expressed as: $Premium = Pr(Accident) \times Pr(Severe|Accident) \times E[Loss] + RiskLoading$. This decomposition reveals that accurate severity prediction captured by $Pr(Severe|Accident)$ directly influences premium determination.

The VSL approach enables insurers to monetize the social value of preventing severe injuries and fatalities, providing economic justification for investments in safety technologies and risk reduction measures. By incorporating VSL estimates into pricing models, insurance companies can better align premiums with the true societal costs of traffic accidents, creating incentives for policyholders to adopt safer driving behaviors and vehicle safety features.

3.2 Reverse Application of Discrete Choice Theory

My analysis employs DCT in a novel reverse application. Rather than predicting prospective choice behavior, I observe realized injury severity outcomes (Non-Severe, Severe, Fatal) and

work backward to infer which crash characteristics most likely contributed to the observed severity level. This methodological adaptation transforms the traditional choice prediction framework into a severity attribution model.

In this application, the "decision-maker" becomes the underlying data-generating process that determines crash outcomes. Each collision represents a realization of a latent severity process influenced by multiple risk factors: (1) Behavioral Factors — driver actions including rule violations, distraction, and impairment; (2) Environmental Conditions — weather patterns, visibility, and temporal factors; (3) Vehicle Characteristics — type, age, and safety technology presence; and (4) Spatial Context — borough-specific infrastructure and traffic patterns.

Each severity category represents a potential "outcome" determined by the combination of crash conditions, with higher-utility scenarios corresponding to more severe injury levels given the risk environment.

Strategic Value for Insurance Analytics

This theoretical framework provides several analytical advantages for insurance applications. Behavioral interpretation enables insurers to understand injury outcomes beyond simple correlation, revealing how specific combinations such as nighttime driving conditions or environmental hazards systematically increase the probability of severe outcomes. Heterogeneous risk modeling accommodates the diverse range of crash-level predictors that vary substantially across NYC's complex urban environment, from Manhattan's dense traffic patterns to Staten Island's suburban characteristics.

Most importantly, the framework establishes a foundation for ordered modeling approaches. Since injury severity represents an inherently ordered outcome with natural progression from property damage through severe injury to fatality, DCT extends seamlessly into an Ordered Logit specification. This connection provides economic justification for the baseline model while supporting the transition to machine learning approaches that can capture

complex non-linear relationships while maintaining interpretability requirements essential for insurance regulatory compliance and business decision-making.

4 Data Preparation

My analysis relies on comprehensive data integration combining multiple high-quality sources spanning temporal, environmental, behavioral, and demographic dimensions: (1) Motor Vehicle Collisions, Crashes — comprehensive collision incident information including timestamps, geographic coordinates, contributing factors, and initial severity assessments; (2) Motor Vehicle Collisions, People — detailed individual information including demographics, injury severity outcomes, person type, and license information; (3) Motor Vehicle Collisions, Vehicles — vehicle characteristics including type, year of manufacture, registration state, and identification numbers; and (4) NOAA Weather Data — hourly weather observations matched to collision records providing temperature, humidity, precipitation, wind speed, and weather descriptions.

The integration process employs a multi-step join strategy: first joining People and Vehicles datasets using vehicle identification numbers, then merging with Crashes dataset using collision identification numbers, creating comprehensive records with incident, person, and vehicle information for each collision event.

4.1 Dataset Construction

My analysis employs a driver-focused approach filtering the dataset to concentrate on at-fault drivers. The filtering process identifies driver records within the People dataset, excluding passengers, pedestrians, and cyclists. At-fault driver identification utilizes contributing factor fields, with drivers associated with specific factors such as “Unsafe Speed,” “Failure to Yield,” or “Driver Distraction” classified as at-fault.

To ensure data quality and avoid potential biases, the analysis focuses on the most recent

three years of collision data, deliberately excluding the COVID-19 period. This temporal restriction prevents the inclusion of anomalous traffic patterns and behavioral changes that occurred during the pandemic, which would introduce bias from this exceptional one-time event that significantly altered normal driving patterns and traffic volumes.

4.2 Final Target Variable Definition

The final analysis employs multiclass classification distinguishing between three distinct categories based on injury outcomes and their economic implications: (1) Non-Severe (63,066 observations) — collisions with no injuries to any person involved in the accident, resulting in property damage only, these incidents do not require medical attention for any participants; (2) Severe (51,777 observations) — collisions where anyone involved in the accident sustained injuries, ranging from minor to severe injuries, this category includes any level of physical harm requiring medical attention, from minor cuts to serious injuries requiring hospitalization; and (3) Fatal (21,123 observations) — collisions resulting in one or more fatalities among any participants in the accident, these represent the most severe outcomes with maximum economic and social costs, including life insurance claims, wrongful death settlements, and immeasurable human loss.

This tripartite classification provides insurance companies with the granularity needed for accurate premium calculation while supporting public safety agencies in targeted intervention planning. The distribution shows approximately 46.4 percent Non-Severe, 38.1 percent Severe, and 15.5 percent Fatal outcomes, reflecting the serious nature of at-fault driver incidents in NYC traffic.

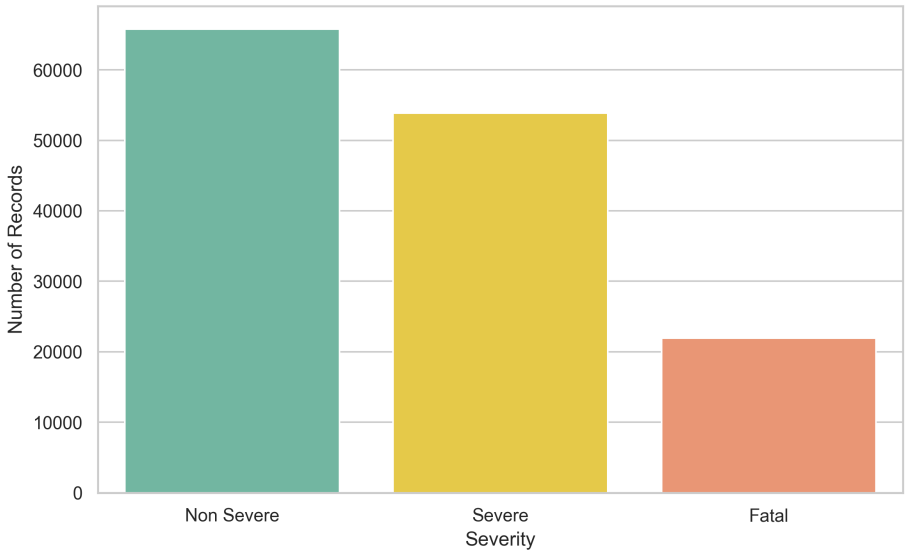

Figure 1: Distribution of Crash Severity Outcomes

5 Data Cleaning and Preprocessing

5.1 Comprehensive Data Cleaning Process

The data cleaning process was designed to maintain data integrity while maximizing the analytical value of the integrated dataset. Key cleaning procedures included:

Missing Value Treatment:

1. **Categorical Variables:** Created explicit "missing" levels for categorical variables including License Status, State Registration, and Vehicle License Status. This approach enables the model to learn patterns associated with missing information rather than discarding potentially valuable observations.
2. **Continuous Variables:** For continuous weather variables where missing data represented less than two percent of observations, these records were removed from the dataset rather than applying imputation techniques. This approach was chosen to avoid introducing bias that could result from imputation methods, maintaining the integrity of the weather data which is critical for severity prediction.

Categorical Variable Consolidation:

1. **Vehicle Type Consolidation:** The original Vehicle Type field contained over 200 unique categories, many with very few observations. Sparse categories were systematically consolidated into broader vehicle type groups, with remaining low-frequency categories assigned to an "Other" category. This consolidation improves model stability while preserving meaningful vehicle type distinctions.

Outlier Detection and Removal:

1. **Age Restrictions:** Driver ages were filtered to the realistic range of 16-114 years, removing clear data entry errors while retaining the full spectrum of legal drivers.
2. **Vehicle Year Validation:** Removed vehicles with manufacture dates after 2025, eliminating impossible future dates that indicate data entry errors.
3. **Geographic Validation:** Verified that all collision coordinates fall within NYC boundaries, removing erroneous location data.

Data Type Corrections:

1. Converted Person ID, Vehicle ID, and Crash ID fields from numeric to object data types to prevent unintended mathematical operations on identifier variables.
2. Standardized date and time formats across all temporal variables to ensure consistent time-based analysis.

6 Feature Engineering and Data Preparation

6.1 Comprehensive Feature Engineering Process

Temporal Features:

1. **Daylight Presence (Y/N):** Engineered using astronomical data for New York City, imported suntime calculations to determine precise sunrise and sunset times for each

crash date, then cross-referenced with crash time to create a binary daylight indicator. This feature captures the critical safety implications of visibility conditions.

2. **Time of Day Categories:** Extracted from crash timestamp and categorized into six distinct periods: Early Morning (5:00-7:59), Morning (8:00-11:59), Afternoon (12:00-16:59), Evening (17:00-19:59), Night (20:00-23:59), and Late Night (0:00-4:59). These categories capture circadian rhythm effects and traffic pattern variations.
3. **Weekday vs. Weekend:** Binary indicator extracted from crash date to capture differential risk patterns between weekday commuting periods and weekend recreational driving.
4. **Seasonal Variables:** Extracted from crash date and categorized into four seasons: Winter (Dec-Feb), Spring (Mar-May), Summer (Jun-Aug), and Fall (Sep-Nov). These categories capture weather pattern effects and seasonal driving behavior variations.

Geographic and Licensing Features:

1. **In-State License Indicator:** Binary variable indicating whether the driver holds a New York State license, extracted from Driver License Jurisdiction field. This feature captures familiarity with local traffic patterns and regulations.
2. **In-State Vehicle Registration:** Binary indicator for vehicles registered in New York State, providing additional context for driver-vehicle familiarity.
3. **Borough Indicators:** Categorical variables for each of NYC's five boroughs, capturing distinct traffic patterns, infrastructure characteristics, and demographic factors.

Environmental Integration:

1. **Weather Variables:** Temperature (°C), Humidity Percentage, Precipitation (mm), Wind Speed (kph), and categorical Weather Descriptions from NOAA hourly observations matched to crash time and location.
2. **Weather Categorization:** Systematic categorization of weather descriptions into interpretable groups including Clear, Cloudy, Rain, Snow, Fog, and extreme weather

conditions.

6.2 Driving Behavior Group Classification System

A key innovation involves transforming the complex contributing factor information into seven interpretable behavioral risk categories. The original contributing factors, written by police officers using free-form text, were systematically converted into standardized behavioral groups:

1. **Driver Distraction** — Behaviors where the driver's attention was diverted from the primary task of driving, including various forms of cognitive, visual, or manual distractions that compromise driving performance.
2. **Under Influence** — Incidents where the driver's cognitive or physical abilities were impaired due to substance use, including alcohol, illegal drugs, or prescription medications affecting driving capability.
3. **Driver Health Issue** — Physical or mental health conditions that compromised the driver's ability to operate the vehicle safely, including fatigue, medical episodes, or physical disabilities affecting driving performance.
4. **Rule Violation** — Deliberate or inadvertent violations of traffic laws and regulations, including speed violations, failure to yield, and disregard for traffic control devices.
5. **Unsafe Maneuver** — Aggressive or improper driving behaviors that create hazardous conditions, including dangerous lane changes, following too closely, and other risky driving actions.
6. **Environment Risk** — External environmental factors that contributed to the crash, including adverse road conditions, weather-related hazards, debris, or defective infrastructure elements.
7. **Vehicle Defect** — Mechanical failures or defects in the vehicle that contributed to the crash, including brake systems, steering components, tires, or other critical vehicle components.

8. **Other Factors** — Contributing factors not fitting the above categories, including unspecified circumstances and unique crash causation factors that cannot be classified into standard behavioral groups.

This behavioral categorization system enables direct translation of model insights into actionable risk management strategies for both insurance pricing and safety interventions.

7 Data Quality Assessment and Feature Selection

7.1 Irrelevant Variable Removal

Several categories of variables were systematically removed from the dataset as they were either irrelevant to the modeling objectives or contained excessive categorical levels that would create computational inefficiency without analytical value.

1. **Identifier Variables** — Person ID, Vehicle ID, and Crash ID fields were removed as these are unique identifiers serving as joining keys rather than predictive features. These variables have no relationship to crash severity outcomes and would only introduce noise into the modeling process.
2. **Geographic Granularity Variables** — ZIP code was removed due to excessive categorical levels that would create computational burden without meaningful predictive improvement over the borough-level geographic indicators already included in the feature set.
3. **Post-Incident Variables** — Variables capturing information recorded after the crash event, such as point of impact details, were excluded as they represent consequences rather than causal factors. These variables are not available at the time of prediction and including them would create data leakage issues.
4. **Redundant Temporal Variables** — Raw datetime, longitude, and latitude variables were removed after extracting engineered features from them. The original timestamp and coordinate data served their purpose in creating meaningful temporal and geo-

graphic features and were no longer needed for the modeling process.

7.2 Feature Selection

To optimize model performance and reduce computational complexity, feature selection was conducted using Random Forest feature importance rankings. This approach identifies variables that contribute most significantly to predicting crash severity outcomes while removing features with minimal predictive value. This data-driven feature selection approach does not affect post-pruning inference validity, as the focus remains on predictive performance rather than causal inference, which is appropriate given the machine learning focus of this analysis.

The Random Forest algorithm was applied to calculate feature importance scores for all 53 variables in the dataset. Based on these rankings, the bottom seven variables with the lowest feature importance scores were systematically removed from the final modeling dataset. Model performance was evaluated both before and after feature removal to ensure that eliminating low-importance variables did not compromise predictive accuracy.

The validation confirmed that removing these seven low-importance features maintained equivalent model performance while reducing dataset dimensionality, resulting in a more efficient and interpretable final model with 46 features for the machine learning analysis.

8 Final Data Structure

The final dataset prepared for machine learning analysis contains 135,966 observations across 45 variables, representing a comprehensive foundation for injury severity prediction. The variables are organized into several key categories that capture different dimensions of crash risk factors. For each category, we present key insights discovered when analyzing the relationship between these variables and crash severity outcomes, providing empirical validation of their predictive relevance.

Demographic Variables (two features): Age and Sex provide essential driver demographic information that influences crash severity patterns through physiological and behavioral factors. A key insight shown in the figure below reveals that fatal crashes involve slightly younger drivers on average, while severe injury crashes show wider age distribution patterns.

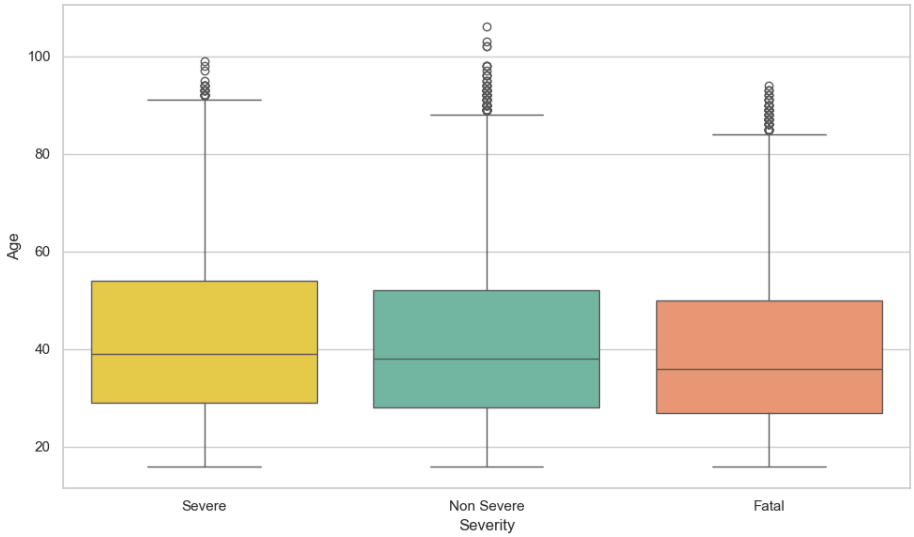

Figure 2: Age Distribution by Crash Severity

Temporal Variables (eight features): Time-related features including daylight presence, time of day categories (Early Morning, Morning, Afternoon, Evening, Night), weekday/weekend indicators, and seasonal variables capture temporal patterns in severity risk. The analysis reveals that weekend crashes demonstrate higher fatal proportions compared to weekday incidents, as illustrated below.

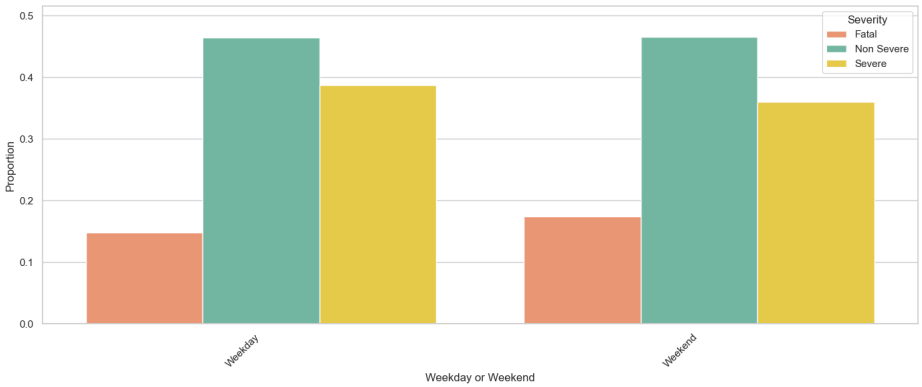

Figure 3: Fatal Crash Proportion by Weekday vs Weekend

Environmental Variables (10 features): Weather-related features including temperature, humidity, precipitation, wind speed, and categorical weather conditions (Cloudy, Fair, Rain, Light Rain, Heavy Rain, Snow, Ice, Fog) provide critical environmental context for crash severity prediction. The figure below depicts distinct severity patterns across different weather conditions, with certain weather types associated with higher fatal proportions.

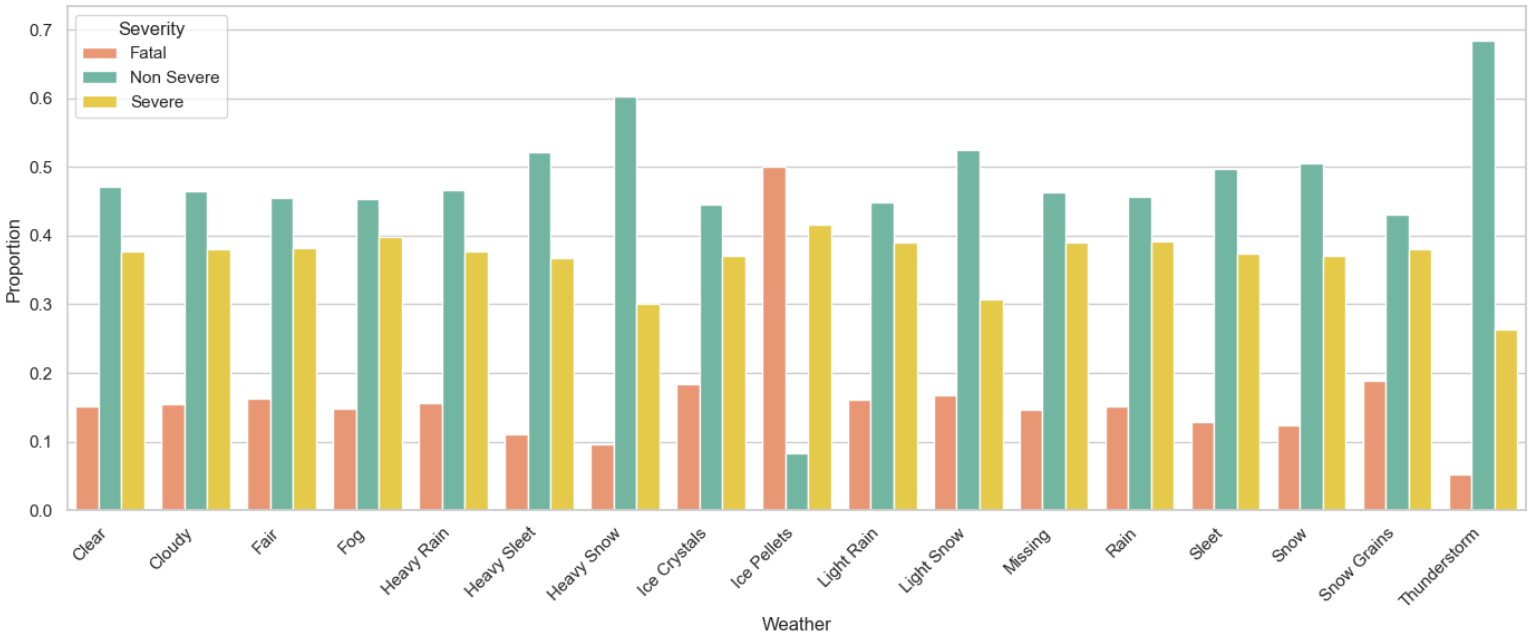

Figure 4: Severity Distribution by Weather Conditions

Geographic Variables (five features): Borough indicators for Brooklyn, Bronx, Manhattan, Queens, and Staten Island capture spatial patterns and infrastructure differences across NYC's diverse urban environments. Analysis reveals clear spatial patterns with Queens and Brooklyn demonstrating highest risk concentrations. To enable meaningful comparison across severity levels, economic cost estimates are applied: Non-Severe ($5,700), Severe ($80,000), and Fatal ($1,800,000) derived from NSC, FMCSA, and USDOT data, providing a common scale for unified risk mapping.

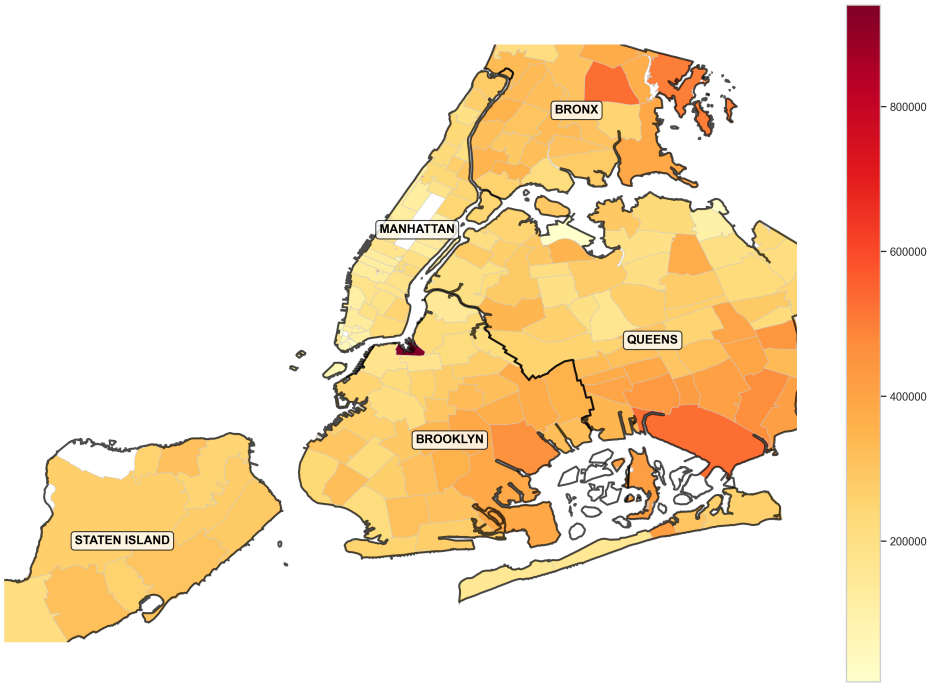

Figure 5: Estimated Cost per Accident by ZIP Code

Behavioral Variables (eight features): Driving behavior group classifications including Driver Distraction, Under Influence, Driver Health Issue, Rule Violation, Unsafe Maneuver, Environment Risk, Vehicle Defect, and Other Factors provide systematic categorization of crash causation patterns. Rule violations emerge as the highest risk category with nearly 20 percent fatal proportion, significantly exceeding other behavioral groups.

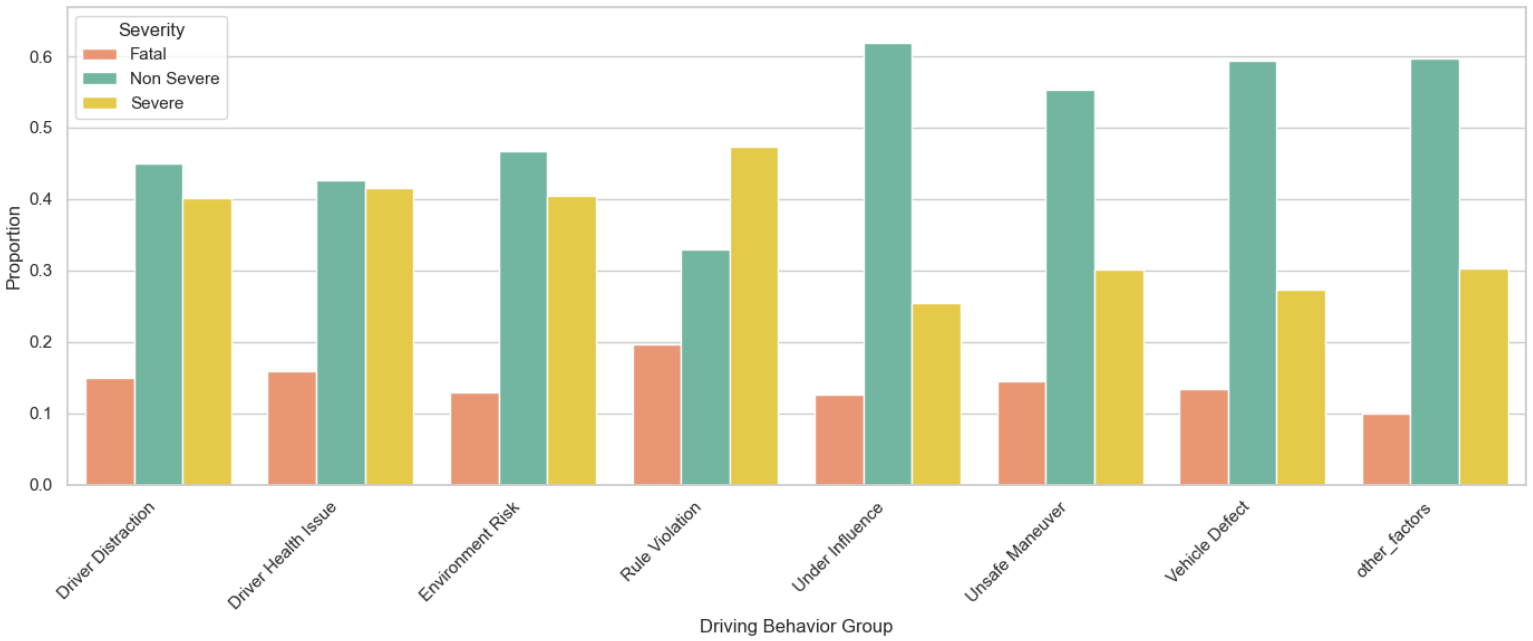

Figure 6: Fatal Proportion by Driving Behavior Group

Vehicle and Licensing Variables (12 features): Vehicle type groups (Small Passenger Vehicle, Large Passenger Vehicle, Two-Wheeler and Micro-Mobility, Light Truck or Utility), vehicle year, license status (Licensed, Unlicensed, Permit), and state registration indicators provide vehicle and driver qualification context. Two-wheel and micro-mobility vehicles demonstrate dramatically higher severe injury proportions compared to other vehicle types, which is logical given the lack of protective barriers and safety structures that passenger vehicles provide. Although these vehicles show lower fatal rates, the elevated severe injury risk reflects the inherent vulnerability of motorcycles, scooters, and bicycles in crash scenarios where riders are directly exposed to impact forces.

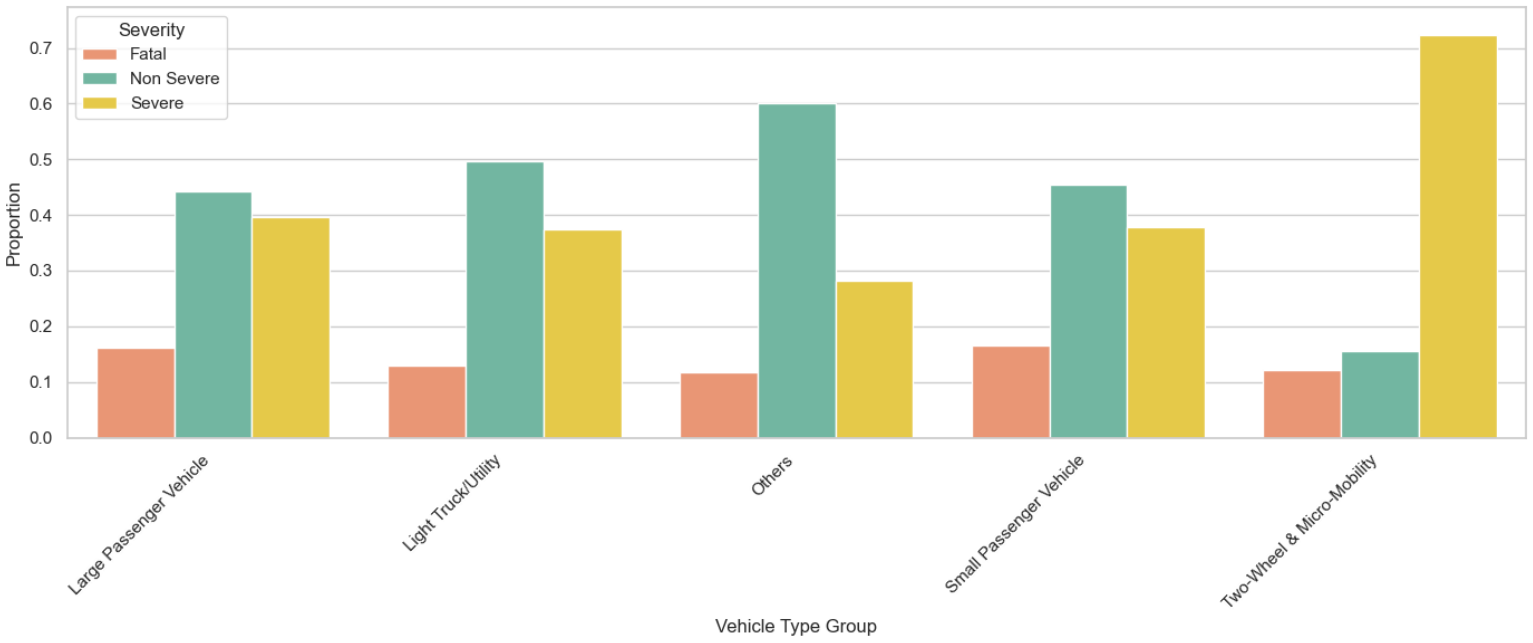

Figure 7: Fatal Proportion by Vehicle Type Group

The target variable maintains the three class structure with Non-Severe (46.4 percent), Severe (38.1 percent), and Fatal (15.5 percent) categories, providing balanced representation across severity outcomes. This final data structure represents a carefully engineered dataset optimized for machine learning analysis while maintaining interpretability and practical relevance for insurance and safety applications.

9 Modeling Framework

9.1 Model Architecture

My modeling approach employs a multi-method strategy balancing interpretability requirements with predictive performance. The framework includes an ordered logit baseline model for interpretable coefficient analysis, multiple machine learning approaches for performance comparison, and an 80–20 train-test split with three-fold cross-validation maintaining severity class proportions. Model selection prioritizes AUC as the primary metric while evaluating

accuracy, precision, and recall across all severity classes.

To improve predictive performance, each machine learning model underwent hyperparameter tuning using randomized search with three-fold cross-validation. Randomized grid search was used to iteratively explore combinations of hyperparameters for each model while avoiding the computational cost of an exhaustive grid search. For Random Forest, the search spanned combinations of n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features. The Logistic Regression model was tuned over a range of C values (logarithmically spaced) and solvers (lbfgs, saga). The XGBoost model parameters included n_estimators, max_depth, learning_rate, subsample, and colsample_bytree, while LightGBM tuning involved n_estimators, max_depth, learning_rate, num_leaves, subsample, and colsample_bytree. The best hyperparameters for each model were selected based on the mean AUC score across the validation folds, ensuring robust generalization and fair model comparison.

9.2 Baseline Model Results (Ordered Logit)

9.2 Theory

The Ordered Logit Model serves as the theoretical baseline for predicting crash injury severity levels, which follow a natural hierarchical ordering: Non-Severe (0) < Severe (1) < Fatal (2). This specification respects the ordinal structure of severity outcomes while capturing the influence of multiple risk factors simultaneously, providing essential interpretability before implementing more complex machine learning approaches.

The model employs a latent variable formulation where an unobserved continuous severity score drives the observed categorical outcome. The latent severity score Y_n^* for crash n is expressed as: $Y_n^* = \alpha + \beta X_n + U_n$ where Y_n^* represents latent injury severity for crash n, X_n captures observed crash-level features (for example temperature, vehicle age, sex), β represents feature effect coefficients, and $U_n \sim Logistic(0, 1)$ serves as the random error term

following a logistic distribution.

The observed severity level Y_n emerges through threshold-based categorization, where the latent severity crosses estimated threshold parameters: $Y_n = \begin{cases} 0 & \text{if } Y_n^* \leq \tau_1 \\ 1 & \text{if } \tau_1 < Y_n^* \leq \tau_2 \\ 2 & \text{if } Y_n^* > \tau_2 \end{cases}$ These thresholds τ_1, τ_2 are learned from the data during model training, enabling precise calibration to the severity distribution observed in NYC crash data.

The model estimates the cumulative probability of falling in or below category j using the cumulative logit formulation: $\text{logit}(Pr(Y \leq j)) = \theta_j - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_k X_k$ where j represents the index of injury category (for example 0, 1), θ_j denotes the estimated threshold or cut point separating category j and $j+1$, and β_k captures the effect of predictor X_k (for example age, weather). This interpretable baseline helps understand directional influence of predictors, provides interpretable class probabilities, and serves as a benchmark for machine learning extensions (for example LightGBM, SHAP). Model coefficients quantify how each feature shifts severity likelihood on the latent scale, while estimated thresholds enable precise mapping to ordered outcomes.

9.2 Insights

The ordered logit model demonstrated acceptable performance, with the logarithm of the likelihood function equal to -121,120, an AIC of 242,300, and a BIC of 242,800. These metrics indicate a reasonable fit and provide essential insights into data patterns. The model also yields standardized coefficients, enabling direct comparisons across variables and supporting interpretable, data-driven conclusions.

Key coefficient insights reveal Vehicle Type Groups demonstrate the strongest associations with severity, with Small and Large Passenger Vehicles showing coefficients of 0.241 and 0.238 respectively. Driving Behavior categories provide actionable insights: Rule Violation

(0.168), Unsafe Maneuver (0.138), and Under Influence (0.115) all show substantial severity increases. Geographic effects show Manhattan with coefficient 0.124, while temporal patterns demonstrate Evening (0.092) and Night (0.075) periods associated with higher severity.

Table 1: Top 10 Variables from Logit Model

Variable	Coef.	P-value
Vehicle Type Group - Small Passenger Vehicle	0.241000	0.000000
Vehicle Type Group - Large Passenger Vehicle	0.238000	0.000000
Vehicle Type Group - Two-Wheel and Micro-Mobility	0.198000	0.000000
Driving Behavior Group - Rule Violation	0.168000	0.000000
Driving Behavior Group - Unsafe Maneuver	0.138000	0.000000
Borough - MANHATTAN	0.124000	0.000000
Driving Behavior Group - Under Influence	0.115000	0.000000
Time of Day - Evening	0.092000	0.000000
Time of Day - Night	0.075000	0.000000
Driving Behavior Group - Vehicle Defect	0.070000	0.000000

The model coefficients quantify how each feature shifts severity likelihood on the latent scale, with positive coefficients indicating increased severity propensity and negative coefficients suggesting protective effects. The estimated thresholds enable precise mapping between continuous severity propensity and discrete outcome categories, providing the interpretable foundation necessary for regulatory compliance while establishing benchmark performance for subsequent machine learning model evaluation.

10 Results and Analysis

10.1 Model Performance Comparison

Systematic comparison across machine learning algorithms validates model selection decisions and provides confidence in result robustness.

Table 2: Model Performance Results

Model	AUC	Accuracy	Precision	Recall
LightGBM	0.675	0.715	0.681	0.738
XGBoost	0.672	0.712	0.694	0.725
Random Forest	0.668	0.708	0.725	0.682
Logistic Regression	0.641	0.685	0.638	0.704

LightGBM achieves the highest AUC of 0.675, representing meaningful improvement over alternatives while demonstrating performance consistent with state-of-the-art results in crash severity prediction literature. The accuracy of 71.5 percent indicates the model correctly classifies nearly three-quarters of severity outcomes. Precision of 68.1 percent and recall of 73.8 percent demonstrate balanced performance across severity classes without systematic bias. XGBoost achieves comparable performance (AUC: 0.672), confirming gradient boosting approaches are well-suited to this domain.

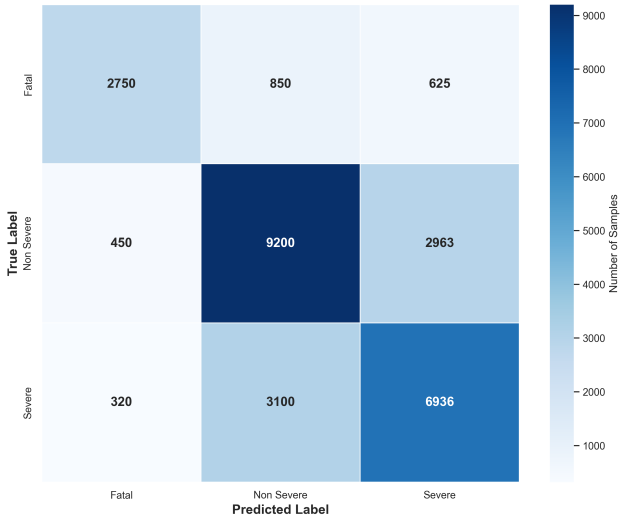

Figure 8: Confusion Matrix — LightGBM Model

The confusion matrix reveals strong diagonal performance with the model correctly predicting 9,200 non-severe cases, 6,936 severe cases, and 2,750 fatal cases. The model shows good discrimination between severity levels, with most misclassifications occurring between

adjacent severity categories rather than extreme misclassifications (for example predicting fatal as non-severe). The highest confusion occurs between severe and non-severe categories (2,963 severe cases predicted as non-severe), which is expected given the inherent difficulty in distinguishing these adjacent severity levels and represents a less critical error than misclassifying fatal crashes.

10.2 Feature Importance Analysis

LightGBM's feature importance analysis reveals relative contributions of different variable categories to severity prediction.

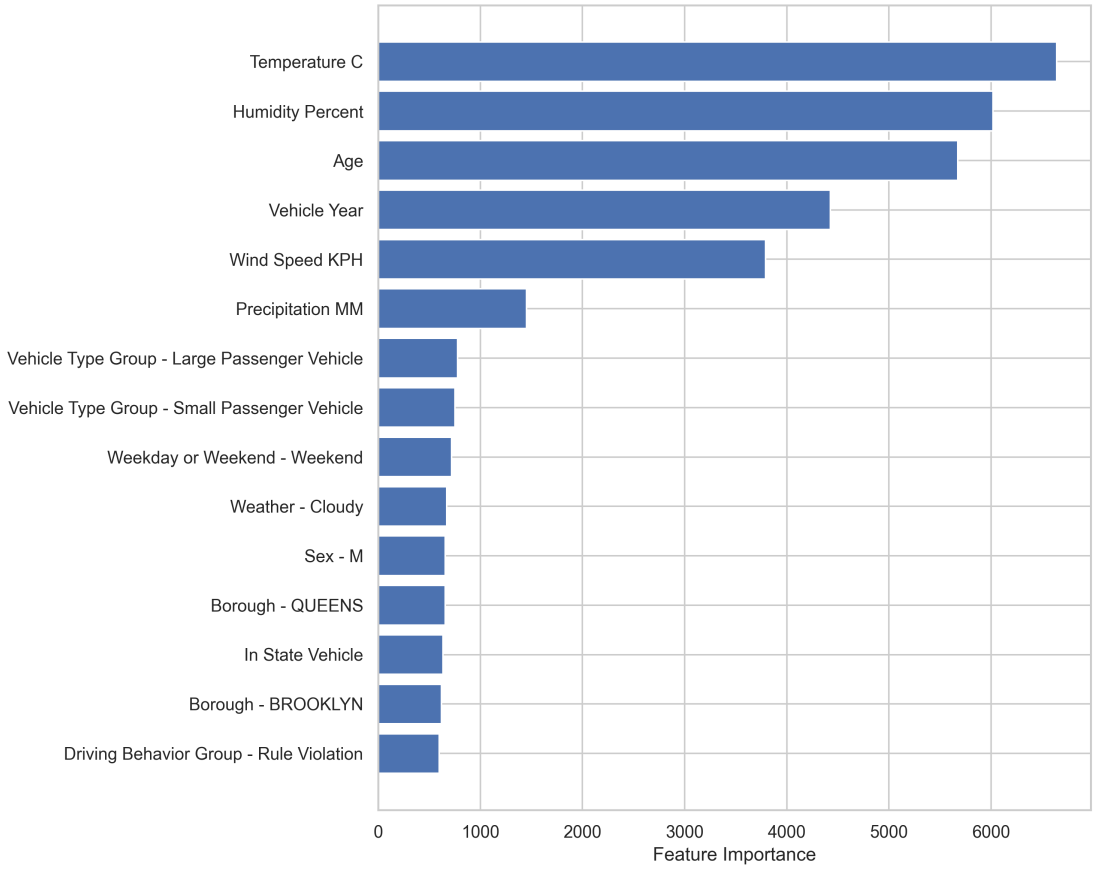

Figure 9: Feature Importance Rankings — LightGBM Model

Environmental variables dominate rankings, with Temperature and Humidity Percent occupying top positions, emphasizing the critical role of weather conditions in determining

crash severity. Age ranks third, confirming demographic importance while highlighting complex non-linear relationships. Vehicle Year appears fourth, indicating substantial roles of vehicle age and safety technology differences. The distributed importance across behavioral categories suggests different risky behaviors influence severity through distinct mechanisms requiring separate modeling.

10.3 SHAP Interpretability Analysis

SHAP analysis provides detailed insights into feature contributions enabling both global understanding and local interpretation.

10.3 Fatal Crash Risk Drivers

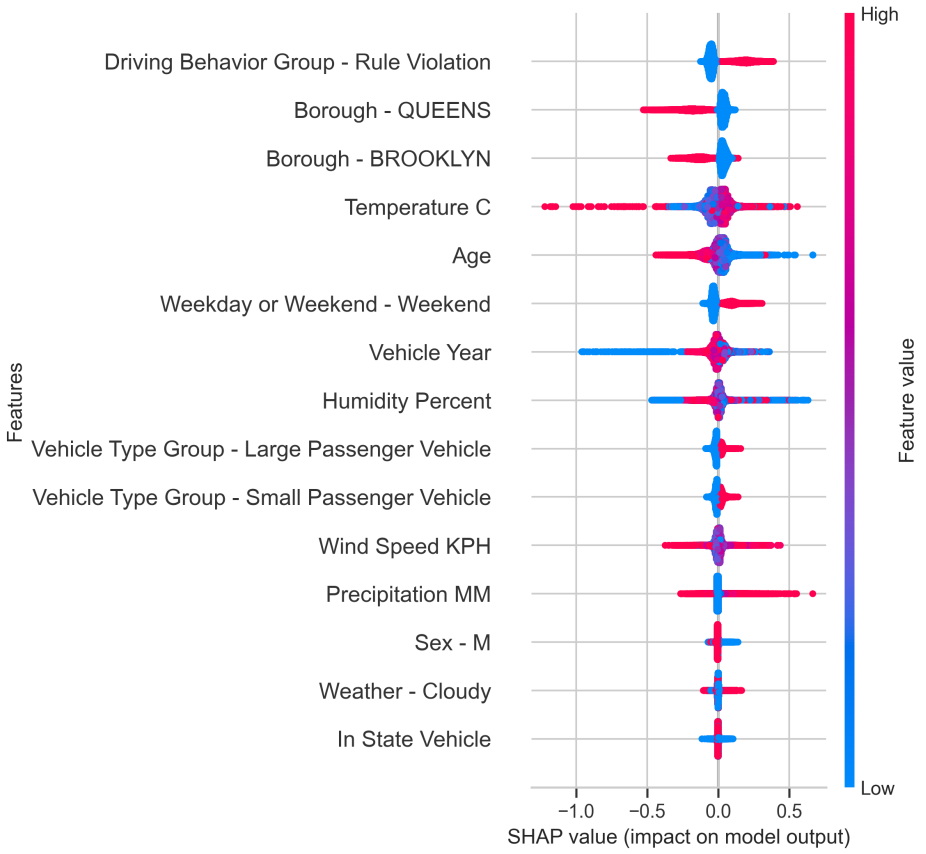

Figure 10: SHAP Summary — Key Drivers of Fatal Crash Risk

Rule Violation behaviors emerge as the strongest predictor with consistently positive SHAP values substantially increasing fatal crash probability. Geographic effects show Queens and Brooklyn with elevated fatal crash risk. Age demonstrates complex patterns: teenage drivers show elevated risk while elderly drivers show lower fatal risk than expected. Vehicle factors show protective effects for certain passenger vehicle types.

10.3 Severe Crash Risk Drivers

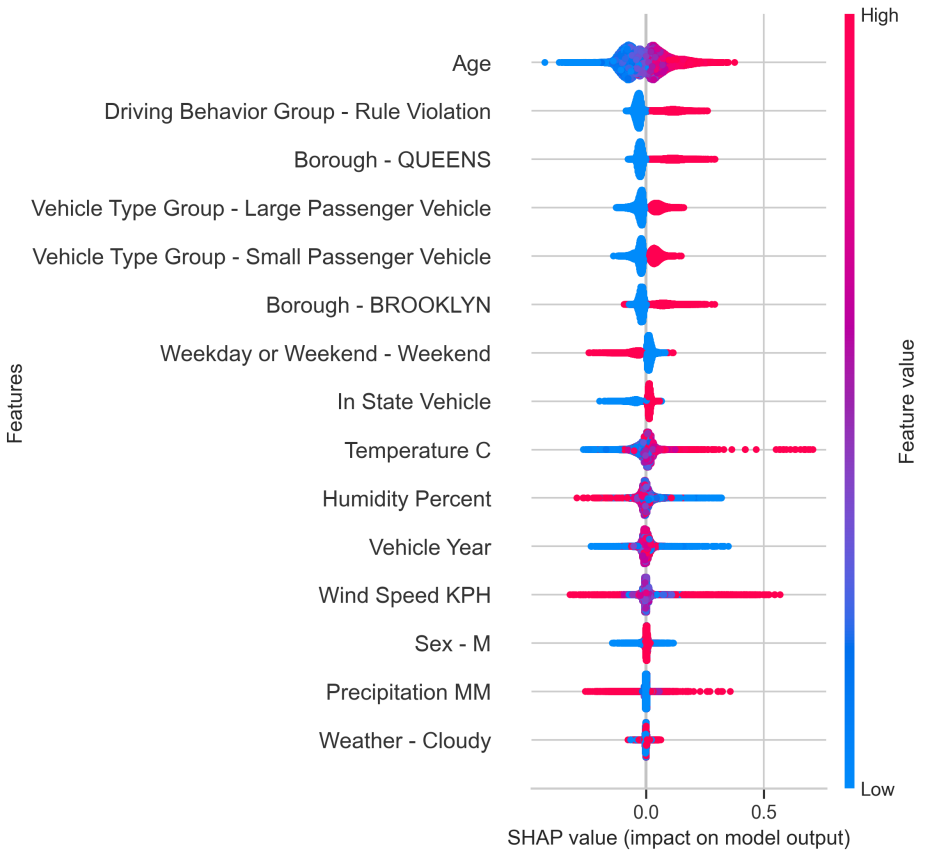

Figure 11: SHAP Summary — Key Drivers of Severe Risk

Age emerges as the primary factor for severe crashes with complex non-linear relationships. Rule Violations remain important but show less dominance compared to fatal outcomes. Geographic concentration continues in Queens and Brooklyn. Environmental factors show increased prominence with Temperature and Humidity demonstrating substantial SHAP

value ranges.

10.3 Age-Specific Risk Patterns

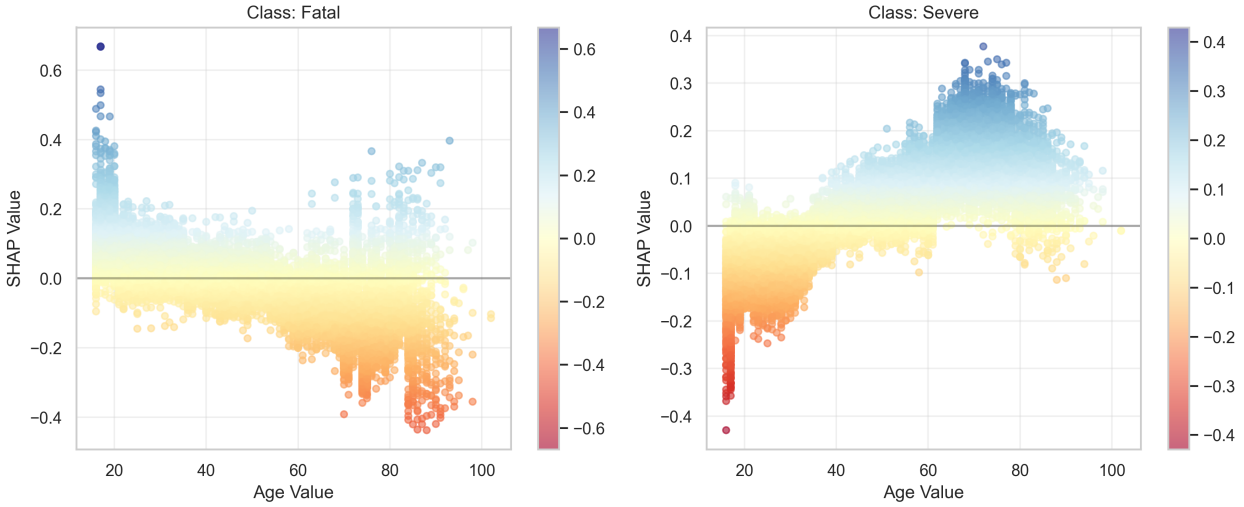

Figure 12: SHAP Analysis — Age and Injury Severity Patterns

Teenage drivers demonstrate high SHAP values for fatal crashes, confirming elevated fatality risk supporting age-based insurance practices. Middle-aged drivers show neutral effects representing baseline risk. Elderly drivers show lower fatal risk but elevated severe injury risk, suggesting age-specific vulnerabilities requiring differentiated approaches.

10.3 Vehicle Year Risk Patterns

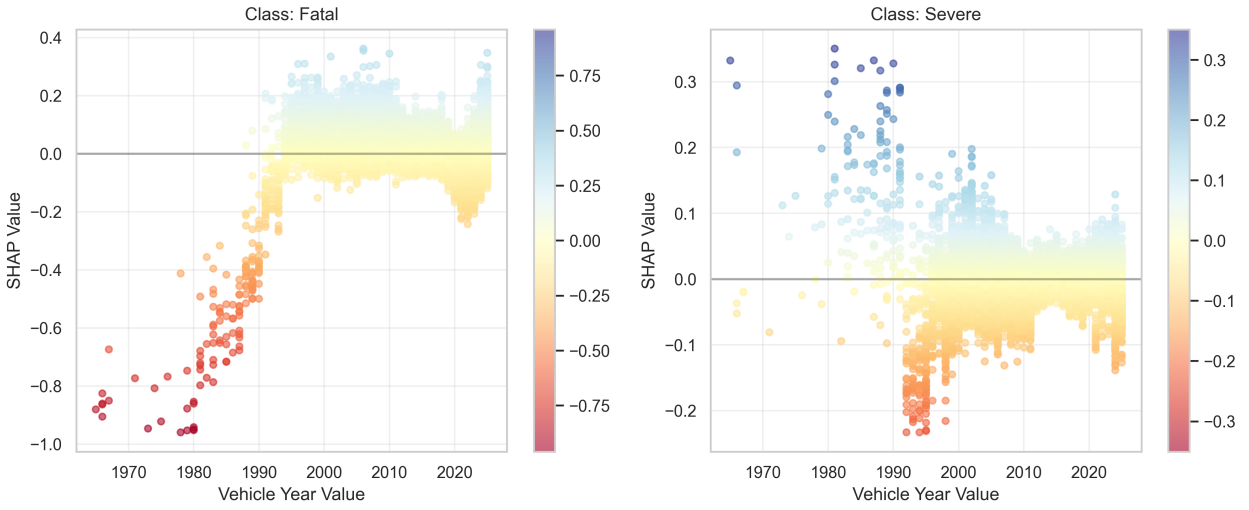

Figure 13: SHAP Analysis — Vehicle Year and Injury Severity Patterns

Vehicle year demonstrates counterintuitive risk patterns revealing safety trade-offs in modern design. Newer vehicles show protective effects for severe injuries (negative SHAP values) due to advanced safety technologies, but paradoxically show increased fatal crash risk (positive SHAP values) despite superior safety features. This likely reflects that newer vehicles enable higher-speed crashes through better performance, are often driven by younger drivers with riskier behaviors, and may encourage aggressive driving. Pre-1980 vehicles show highest risk for both outcomes due to absent safety technologies. These findings suggest modern safety technology effectively reduces injury severity but behavioral and performance characteristics of newer vehicles may offset safety gains in fatal scenarios, indicating insurance pricing should balance both protective effects and increased high-severity crash exposure.

11 Insurance Industry Applications

The SHAP analysis provides quantitative evidence for implementing differentiated pricing strategies based on empirically validated risk factors. Rule violations demonstrate the

strongest predictive power for fatal outcomes, with consistently positive SHAP values across all severity categories, supporting premium loadings for drivers with violation histories. Geographic risk stratification reveals systematic severity differences across NYC boroughs, with Queens and Brooklyn consistently showing elevated risk profiles, providing actuarial justification for location-based rating factors that reflect underlying infrastructure and traffic density differences.

Age-related risk patterns exhibit complexity requiring sophisticated pricing approaches, as the SHAP analysis reveals that teenage drivers show elevated fatal crash risk while drivers over 65 demonstrate increased severe injury risk but reduced fatal risk. This divergent pattern suggests that traditional linear age rating may be suboptimal, supporting segmented pricing strategies that account for distinct risk profiles across demographic categories. Vehicle characteristics present nuanced considerations, as the counterintuitive finding that newer vehicles show protective effects for severe injuries but increased fatal crash exposure suggests that simple vehicle age discounts may inadequately capture risk dynamics.

Environmental variables, particularly temperature and humidity, rank among the top predictive features, suggesting systematic integration of weather data into risk assessment processes for dynamic pricing adjustments and seasonal risk management strategies. The behavioral classification system enables evaluation of driver risk profiles beyond traditional demographic factors, allowing identification of drivers with histories of rule violations and unsafe maneuvers for enhanced risk assessment. SHAP interpretability provides regulatory compliance advantages by offering transparent, quantifiable explanations for pricing decisions, demonstrating empirical relationships between specific risk factors and severity outcomes while supporting actuarial justification requirements and maintaining customer transparency in rating factor applications.

12 Public Safety and Policy Applications

The geographic concentration of severe and fatal crashes in Queens and Brooklyn, validated across multiple analytical approaches, provides clear targeting criteria for infrastructure investment and enforcement resource allocation, enabling transportation authorities to prioritize interventions in areas where severity reduction potential is empirically demonstrated to be highest. The dominance of rule violations as fatal crash predictors suggests reorienting enforcement strategies from general traffic monitoring toward compliance-focused interventions, with speed enforcement, traffic signal compliance monitoring, and right-of-way violation prevention emerging as evidence-based priorities that directly address the strongest predictors of fatal outcomes identified in the analysis.

Environmental risk patterns support implementing weather-responsive safety protocols, as the prominence of temperature and humidity as severity predictors indicates that dynamic safety measures—such as enhanced enforcement during high-risk weather conditions or real-time driver advisories—could provide measurable severity reduction benefits. The finding that two-wheel and micro-mobility vehicles show dramatically higher severe injury proportions suggests prioritizing protected infrastructure for vulnerable road users, while the complex relationship between vehicle age and severity outcomes indicates that safety benefits from fleet modernization may be more nuanced than traditionally assumed.

The severity prediction framework enables proactive rather than reactive safety planning by identifying risk patterns before they manifest in increased crash frequencies, allowing identification of emerging risk concentrations based on environmental, behavioral, and demographic factors that precede crash occurrence. Age-specific risk patterns support differentiated safety interventions, with young driver fatal crash prevention programs emphasizing rule compliance and risk awareness, while elderly driver programs focus on severe injury prevention through enhanced infrastructure design and vehicle safety technology promotion. The behavioral classification system provides a foundation for targeted safety campaigns, as

understanding that different risky behaviors influence severity through distinct mechanisms enables developing specific interventions for distracted driving, impaired driving, and aggressive driving behaviors based on their empirically demonstrated severity impacts.

13 Model Limitations and Future Research

13.1 Current Limitations

The AUC of 0.675 is indicative of weak predictive capability—only slightly better than chance, as a coin toss would deliver 50 percent accuracy. Thus, improvements are required. Enhanced weather data could help capture local variations that influence crash severity beyond the borough-level weather data currently used. For example, New York City exhibits diverse microclimates and urban heat island effects that create meaningful within-city variation. Incorporating driver history and experience could also provide individual-level behavioral insights, complementing the incident-specific categories currently employed and enabling more precise risk assessment.

Road infrastructure and design features could capture systematic factors that influence severity outcomes beyond the geographic effects currently in the model. Information about road geometry, intersection design, traffic control systems, and infrastructure condition could explain the geographic patterns found in the analysis while providing additional predictive power for severity outcomes.

The analysis uses a three year window designed to avoid COVID-19 disruptions that created significant departures from normal traffic patterns during 2020-2021. While this approach ensures the model reflects normal operating conditions, longer historical periods could provide additional insights into trends and seasonal variations. Additionally, the analysis is limited to reported crashes meeting NYC's reporting requirements and focuses on at-fault drivers, which may exclude information about passenger and pedestrian severity patterns that could inform broader safety strategies.

13.2 Future Research Directions

Future research should examine longer historical periods to understand how crash severity patterns evolve over time and evaluate the effectiveness of safety interventions. Incorporating detailed infrastructure variables such as road design, intersection types, and traffic signal configurations could help explain why certain areas like Queens and Brooklyn consistently show higher severity risks. Additionally, tracking individual drivers over multiple years would provide insights into how age, experience, and behavioral changes affect crash outcomes, enabling more personalized risk assessment for insurance applications.

Testing this framework in other cities would help determine which findings are specific to New York versus universal patterns that apply broadly. The discrete choice theory foundation provides a solid basis for adaptation to different urban environments with varying traffic patterns and infrastructure characteristics. Developing practical applications of this research, such as real-time risk assessment tools that incorporate current weather conditions and traffic data, could benefit both insurance companies and transportation agencies in making data-driven decisions about pricing and safety interventions.

14 Conclusion

In this research, I have developed a comprehensive machine learning framework for predicting traffic accident injury severity in New York City. The theoretical foundation is the ordered logit model, which establishes baseline interpretability before implementing advanced machine learning approaches. This methodology builds on prior validation by Rifat et al. (2024) and Santos et al. (2021), who demonstrated the effectiveness of ensemble methods in crash severity prediction.

Building on literature insights from Aziz et al. (2013) regarding NYC's borough-specific effects and Park et al. (2020) on environmental factors, the analysis of 135,966 collision records reveals that LightGBM with SHAP interpretability provides optimal balance of predictive

performance (AUC: 0.675) and actionable insights. Key findings identify rule violations as the strongest fatal crash predictor, complex age-severity relationships, and geographic risk concentration in Queens and Brooklyn, consistent with theoretical expectations from discrete choice modeling.

The framework's practical impact extends to insurance pricing strategies through evidence-based premium adjustments and policy applications supporting targeted resource allocation. The successful integration of behavioral categorization, environmental data, and interpretable machine learning establishes severity prediction as an important complement to traditional frequency-based approaches, offering immediate applications for improving both insurance market efficiency and public safety outcomes while providing a foundation for extension to other urban contexts.

15 Acknowledgements

I am profoundly thankful to Dr. Harry J. Paarsch for his exceptional mentorship and intellectual guidance throughout this project. Widely regarded for his contributions to economics and econometric methodology, his deep expertise, sharp insights, and consistent encouragement played a key role in shaping this research.

I also wish to extend my appreciation to the faculty members of my master's program — Dr. Michael Tseng, Dr. Alexander Mantzaris, and Professor Majid Mahzoon — whose instruction and academic support have been greatly influential.

As a graduate teaching assistant, I had the opportunity to collaborate with Dr. Stacey Brook and Dr. Jordan Izenwasser, whose mentorship and constructive feedback enriched my academic development. I am also grateful to Professor Joshua Eubanks, Associate Chair, for his steady guidance and encouragement. Finally, I am thankful to my fellow students in the program for fostering an environment of mutual support, collaboration, and shared learning that made this journey truly rewarding.

16 Appendix

Data Sources

The data sources used in this analysis can be found at the following locations: (1) Motor Vehicle Collisions – Crashes: `https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions` `h9gi-nx95/about_data`, (2) Motor Vehicle Collisions – People: `https://data.cityofnewyork.` `us/Public-Safety/Motor-Vehicle-Collisions-Person/f55k-p6yu/about_data`, (3) Motor Vehicle Collisions – Vehicles: `https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Col` `bm4k-52h4/about_data`, and (4) NOAA Weather Data via API: `https://data.meteostat.` `net`.

Other SHAP Plots

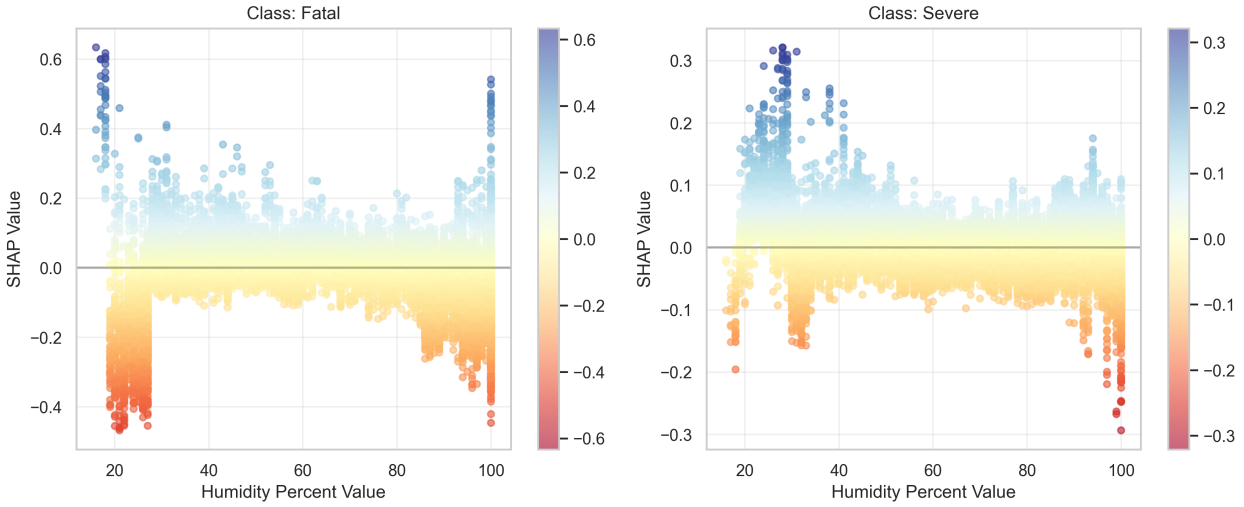

Figure 14: SHAP Analysis — Humidity and Injury Severity Patterns

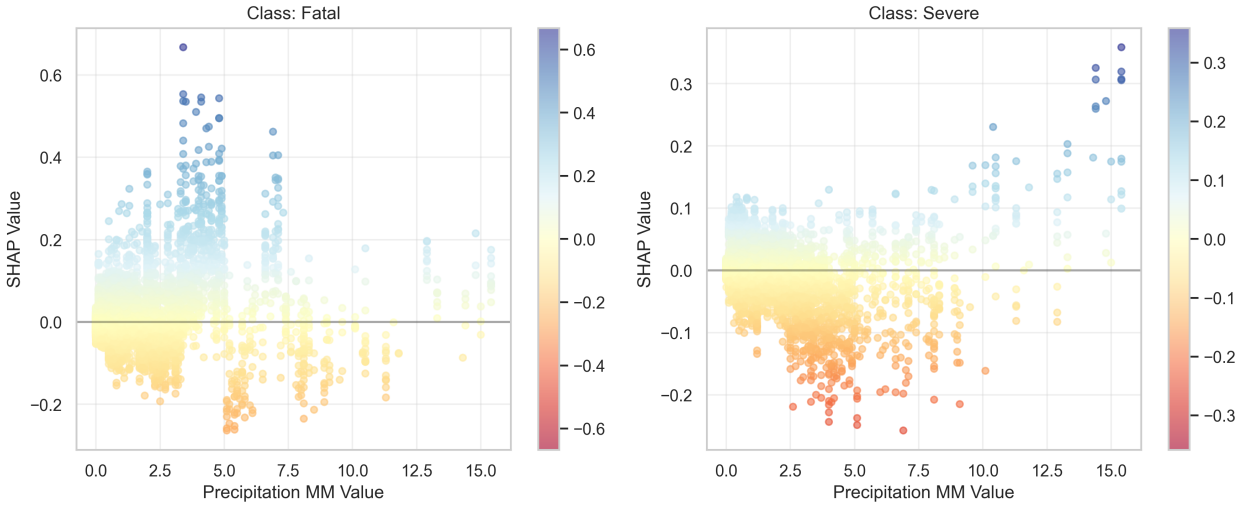

Figure 15: SHAP Analysis — Precipitation and Injury Severity Patterns

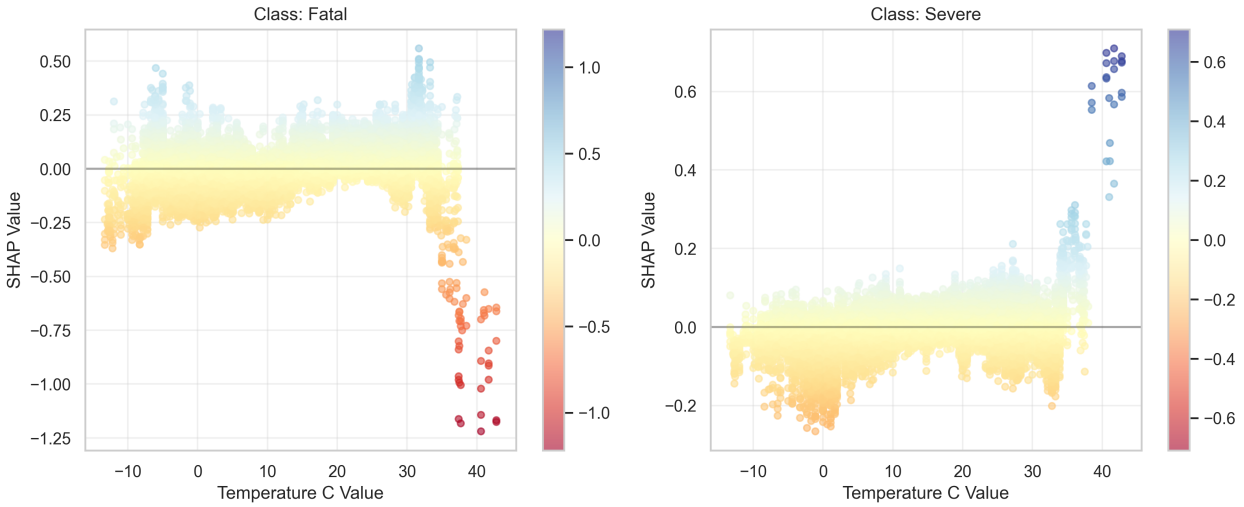

Figure 16: SHAP Analysis — Temperature and Injury Severity Patterns

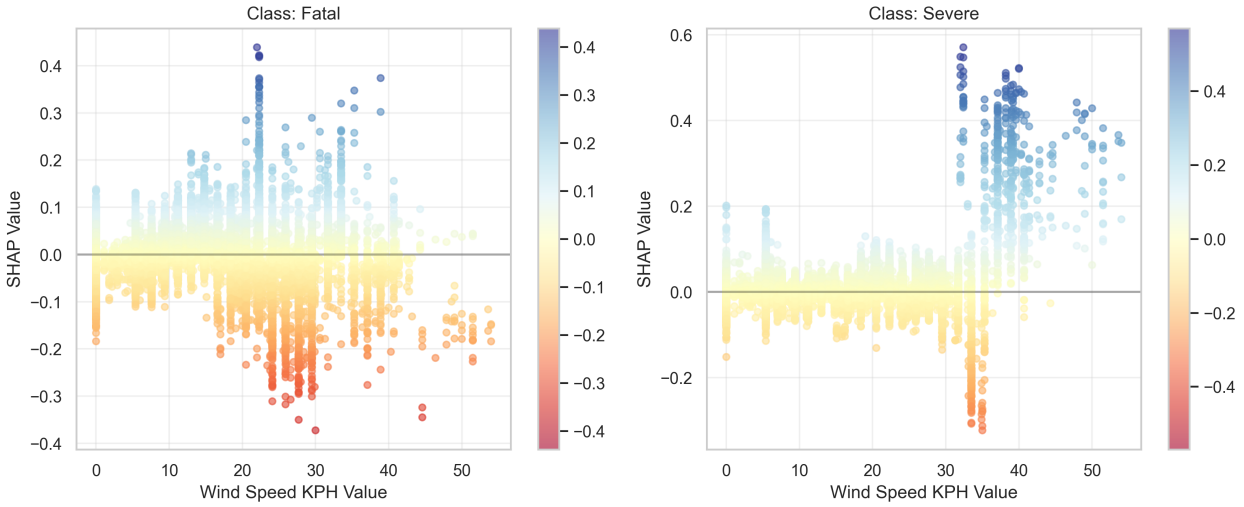

Figure 17: SHAP Analysis — Wind Speed and Injury Severity Patterns

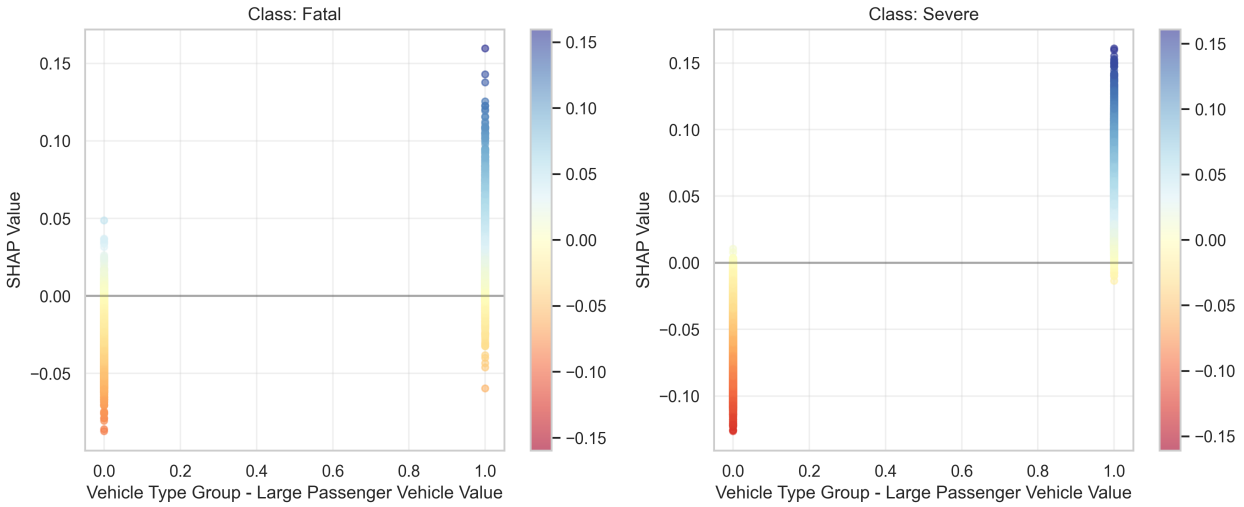

Figure 18: SHAP Analysis — Large Passenger Vehicle Type and Injury Severity Patterns

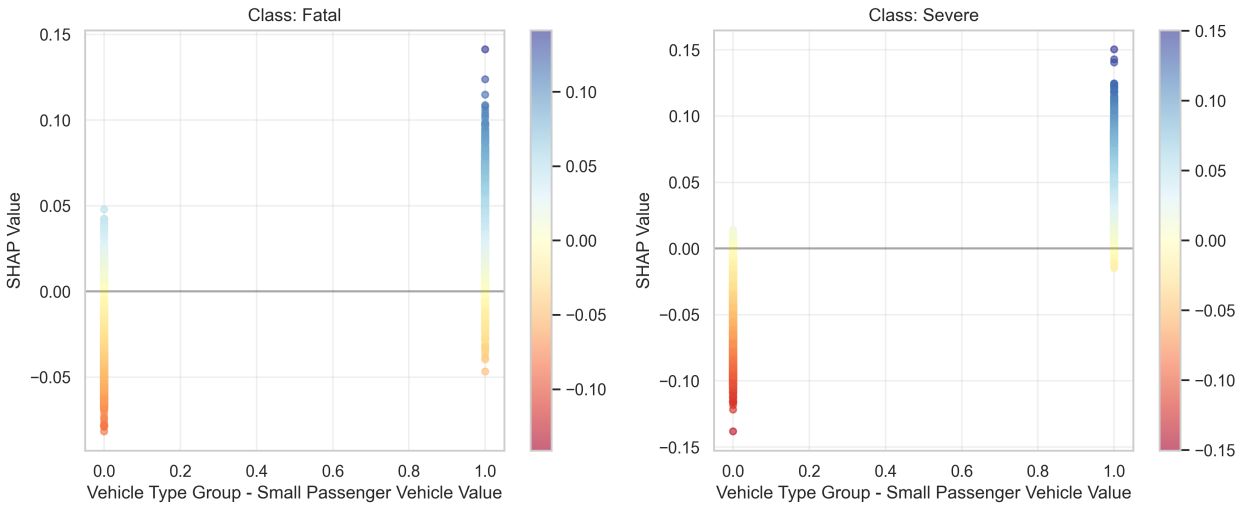

Figure 19: SHAP Analysis — Small Passenger Vehicle Type and Injury Severity Patterns

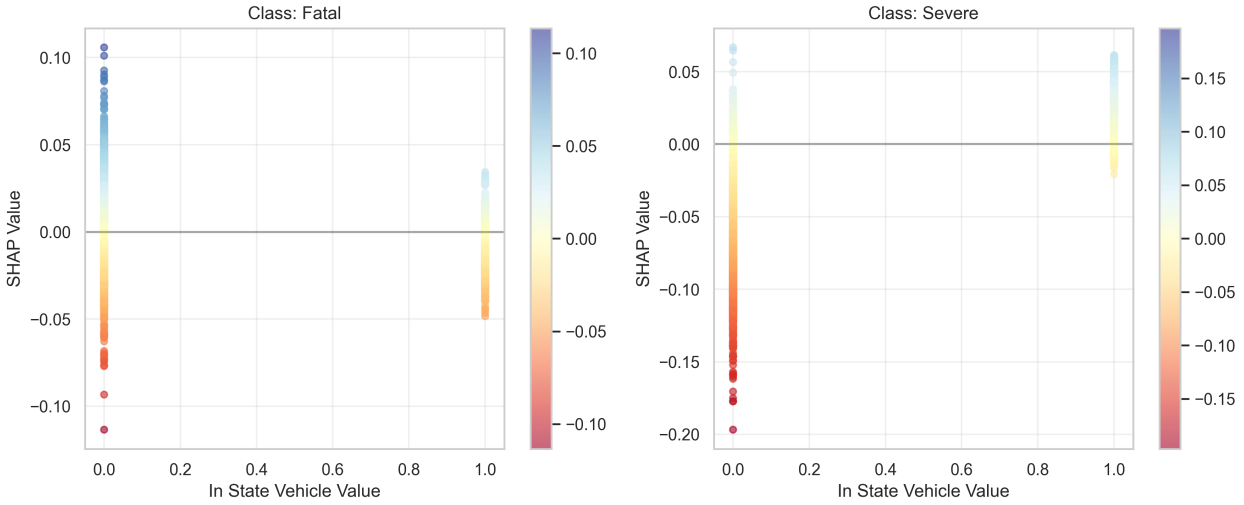

Figure 20: SHAP Analysis — In-State Vehicle Registration and Injury Severity Patterns

References

Technical report.

Aziz, H. M. A., S. V. Ukkusuri, S. Hasan, and S. Akter (2013). Exploring the Determinants

of Pedestrian Injury Severity in New York City. *Accident Analysis & Prevention 50*, 1298–1309.

Bankrate (2024). Average Cost of Car Insurance in 2024. Accessed: 2024.

Crash Intelligence Solutions (2024). Total Cost of Repair Industry Report. *Auto Industry Insights 15*(2), 45–62.

Dionne, G. and C. Rothschild (2014). Economic Foundations of Auto Insurance. In *Handbook of Insurance Economics*, pp. 35–75. Springer.

Einav, L., A. Finkelstein, and J. Levin (2022). The Impact of Monitoring Programs in U.S. Auto Insurance. NBER Working Paper 29096, National Bureau of Economic Research.

Forbes (2024). Auto Insurance Rates Rise in NYC. Accessed: 2024.

Gutiérrez-Osorio, C. and C. Pedraza (2024). Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of Machine Learning Techniques. *arXiv preprint*. arXiv:2406.13968.

Insurance, M. (2023). Industry Insights on Auto Risk Scoring. Accessed: 2023.

Investopedia. Adverse Selection. Accessed: July 26, 2025.

Obasi, I. C., C. Benson, and M. O. Okwu (2023). Evaluating the Effectiveness of Machine Learning Techniques in Forecasting the Severity of Traffic Accidents. *Heliyon 9*(8), e18812.

Park, S. H., S. M. Kim, and Y. G. Ha (2020). Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study. *Applied Sciences 10*(1), 129.

Rifat, M. A. K., B. Alam, and D. Saha (2024). An Interpretable Machine Learning Approach to Injury Severity Prediction. *Transportation Research Record 2678*(3), 312–323.

Santos, K., J. P. Dias, and C. Amado (2021). A Literature Review of Machine Learning Algorithms for Crash Injury Severity Prediction. *Journal of Safety Research 80*, 254–269.

U.S. Department of Transportation (2016). Crash Cost Estimates by Maximum Police-Reported Injury Severity. Technical report DOT HS 812 231, National Highway Traffic Safety Administration (NHTSA).

Wang, C., A. Mohammed, and R. S. Chauhan (2020). Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis. *Sustainability 15*(7), 5939.