CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

AI-Based Crash Severity Modeling Using United States Data for Potential Applications in Pakistan

by

Muhammad Usman

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Engineering
Department of Civil Engineering

2026

CERTIFICATE OF APPROVAL

AI-Based Crash Severity Modeling Using United States Data for Potential Applications in Pakistan

by

Muhammad Usman

Registration No: (MCE223001)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Ghulam Yaseen	MY University, Islamabad
(b)	Internal Examiner	Dr. Manzar Masud	CUST, Islamabad

Dr. M. Usman Farooqi
Thesis Supervisor
September, 2026

Dr. Majid Ali
Head
Dept. of Civil Engineering
September, 2026

Dr. Imtiaz Ahmad Taj
Dean
Faculty of Engineering
September, 2026

Author's Declaration

I, **Muhammad Usman ,** hereby state that my MS thesis titled "**AI-Based Crash Severity Modeling Using United States Data for Potential Applications in Pakistan**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Muhammad Usman)

Registration No: (MCE223001)

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**AI-Based Crash Severity Modeling Using United States Data for Potential Applications in Pakistan**" is exclusively my research work with no remarkable contribution from any other individual. Small contribution/help wherever taken has been acknowledged and that complete thesis has been written by me.

I understand the zero-tolerance policy of the Higher Education Commission and CUST towards plagiarism. Therefore, I as an author of the above titled thesis declare that no part of my thesis has been plagiarized and any material used as reference is properly cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Muhammad Usman)

Registration No: (MCE223001)

Acknowledgement

First and above all, I praise **ALLAH**, the Almighty, for providing me the opportunity and granting me the capability to proceed successfully. In addition, to **Muhammad (PBUH)**, the Divine servant Leader, who has changed my life.

My sincere and heartfelt gratitude goes out to my mentor **Engr. Dr. Muhammad Usman Farooqi**, for he not only his unwavering support, direction and inspiration during this research, but also for shaping my aptitude and laying solid foundations of Transportation Engineering. It is his genius that I was able to think 'Transportation Policy' and deep unrooting the inherent and intrinsic challenges in transportation engineering in general and challenges and obstacles for Pakistan. He has went beyond classroom mentorship and has actually provided sound guidance whenever sought. This gratitude is not 'formal'. It is my lifelong honor to have worked under his direction.

I am grateful to my **mother**, who worked to nurture me and to provide me and support my goals in the darkest of hours. There are simply no words to carry the weight of her greatness.

I am thankful to critics who gave critical insights and helped me to look with their eyes.

Muhammad Usman

Abstract

With the rapid expansion of transportation networks, predicting road traffic accident severity remains a critical challenge for highway engineers. Classical statistical methods have historically been applied to understand crash dynamics but these methods are limited in processing complex multidimensional patterns. This study employs machine learning to proactively predict crash outcomes and enable scenario-based infrastructure analysis. The research utilizes data from the United States Fatality Analysis Reporting System spanning the years 2005-2009 yielding a final analytical sample size of about 180,000 records. Because this database operates as a fatal crash sampling frame, the data is inherently skewed toward severe outcomes. The exact response variable evaluated in this study is the maximum driver injury severity which is structured across the five class ordinal KABCO scale.

To develop robust predictive frameworks without data leakage, the methodology utilized a strict 80/20 train and test split procedure where the Synthetic Minority Oversampling Technique was applied exclusively within the training folds. Three ensemble algorithms including Random Forest, XGBoost, and LightGBM were trained and subsequently evaluated using principal evaluation metrics including weighted averaged precision, weighted averaged recall, and the F1 score on the untouched test set. Following a rigorous evaluation of the test set consistency, the Random Forest model emerged as the most successful architecture achieving a weighted F1 score of 0.7805 followed by LightGBM and XGBoost. Because these algorithms function as opaque systems, Explainable Artificial Intelligence was incorporated through SHapley Additive exPlanations to interpret their internal logic. By evaluating feature contributions to a probability weighted expected severity score, the models revealed that driver behavior, including negligence and poor lane discipline, acts as a primary catalyst for severe outcomes. Furthermore, the presence of pedestrians, highway alignment characteristics, the extent of vehicle deformation, and the physical manner of collision were identified as the most highly significant predictors in determining the ultimate severity of a crash.

Keywords: Crash Severity, AI, Random Forest, XGBoost, LightGBM, FARS, Pakistan

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Transportation infrastructure serves as the fundamental backbone of modern economic development, societal connectivity, and global trade. Efficient mobility networks facilitate the movement of goods and populations, directly driving national growth and urbanization. However, the operational efficiency and public safety of these vital networks are continuously undermined by a persistent, systemic nemesis: Road Traffic Accidents (RTAs). In transportation engineering, diagnosing this problem requires establishing precise boundaries between the physical event and its consequences. An RTA, or crash, refers strictly to the dynamic collision event. The human toll resulting from a crash is defined as casualties, which must be strictly subdivided into non-fatal physical trauma (injuries) and the loss of human life (fatalities). The global scale of these outcomes is critical; according to the World Health Organization, traffic crashes cause approximately 1.19 million fatalities annually and leave an additional 20 to 50 million individuals with non-fatal injuries [1]. Beyond the profound human tragedy, these events generate severe economic losses. Current authoritative metrics indicate that medical treatments, infrastructure damage, and lost workforce productivity stemming from RTAs cost most nations approximately 3% of their annual Gross Domestic Product (GDP) [2].

Rapid urban sprawl generates more traffic and transportation demand, RTAs are now considered a major public health problem, in Italy the RTA caused fatalities in 2022 surpassed those in 2021 along with injuries and property damages [3]. Rapid urbanization has also been found to cause RTAs and injuries in Kingdom of Saudi Arabia where the number of RTAs surpassed as much as 30 times than the pre-urbanization model adaptation [4] the same case as with Malaysia [5] and Nigeria [6]. In the first world countries, the national level costs of RTAs have been estimated to be around 0.4% to 4.1% of the GDP, which is too high to go unnoticed. While the crash-cost per vehicle kilometer travel (VKT) have also been estimated as 0.04$/VKT for rural roads and 0.02$/VKT for the urban roads [7]. The RTA fatalities are more in low income and developing countries due to slow response and lack of research and upgradation of policies [8]. Low and middle socio-demographic index (SDI) regions experience more severity than the high income regions, which puts more economic burden on the economy [9]. Road Crash Fatality (RTF) victim's dependent families have been observed to have a decreased quality of life [10] and significant psychological effect on children and mothers involved in road crashes [11]. However, RTA have mostly been studied using classical statistical methods. There have been fewer studies incorporating the Artificial Intelligence and of those studies, most have considered a narrower dataset [12]. In view of above it is evident that the RTAs cause severe damages to human life and property, therefore and systematic mechanism to mitigate RTAs needs to be developed [13].

1.2 Research Motivation and Problem Statement

Achieving safer roads with fewer and less severe crashes is an important goal for sustainable transportation, particularly in developing countries such as Pakistan. A better understanding of why crashes result in different levels of severity can contribute to improved knowledge of road safety risks and support more informed safety decisions. The growing availability of detailed crash databases offers an opportunity to gain meaningful insights from the complex information contained

in these records. Accordingly, this study focuses on examining crash severity patterns and the factors associated with different crash outcomes, contributing to a better understanding of road safety risks and supporting more informed road safety research and practice.

"Road crashes continue to cause serious injuries and losses, but the severity of their consequences can vary considerably from one crash to another. Understanding the factors that contribute to these differences is challenging because crash outcomes are influenced by a complex combination of driver, vehicle, roadway, environmental, and crash-related conditions. In Pakistan, limitations in the availability, consistency, and accessibility of detailed crash data further constrain data-driven crash-severity analysis. There is therefore a need for reliable and interpretable approaches to classify injury severity and identify the factors associated with severe outcomes. In this context, the U.S. Fatality Analysis Reporting System (FARS) provides a comprehensive source of crash data for developing and assessing AI-based severity modelling approaches, with potential relevance to road safety analysis in Pakistan."

1.2.1 Research Questions

The problem statement above necessitates answering the following questions for resolution:

1. Which pre and post-crash features exhibit the strongest mathematical association with maximum driver injury severity on the KABCO scale, as determined by global Explainable AI feature attribution?

2. What are the direction and magnitude of each feature's marginal contribution to the model's severity predictions, as quantified through SHAP (SHapley Additive exPlanations) analysis?

3. How do the three ensemble models (Random Forest, XGBoost, and LightGBM) compare in predictive classification on unseen test data when evaluated using Accuracy alongside weighted Precision, Recall, and F1-Score?

1.3 Overall Goal and Specific Aim of This Study

This study aims to employ Artificial Intelligence by training multiple machine learning models, to help understand the complex interplay of various highway, traffic and human behavior related features related to the RTAs by demystifying the black-box nature of machine learning models by applying the methods of XAI. It seeks to aid the transportation engineers and policymakers in reducing the life and property risks associated with the road traffic by providing a mechanism of simulating the traffic and highway conditions and assessing the associated risks and predicting the severity of the crashes. It also aims to aid the emergency responders for early preparations in case of certain traffic and environmental conditions and in case of reported crashes.

1.4 Scope of Work and Study Limitations

This study relies on the U.S. Fatality Analysis Reporting System (FARS), a nationally representative database maintained under stringent federal standards (NHTSA, 2011). Its scope is intentionally restricted to ex-post severity classification using pre- and post-crash variables from FARS 2005-2009, and the resulting models are evaluated solely within that sampling frame. While the methodological approach of ensemble modelling combined with SHAP-based feature attribution is designed to be portable, several contextual factors caution against the direct transfer of the numerical findings to a setting such as Pakistan.

Crash reporting protocols differ markedly. FARS records are compiled by trained analysts using standardized vehicle, roadway, and injury codes, whereas Pakistani crash data are often sourced from police first-information reports, which exhibit substantial underreporting, especially for non-fatal and vulnerable-road-user injuries (World Health Organization, 2018). Road design standards, vehicle fleet composition, and speed measurement units also diverge: U.S. highways follow AASHTO design guidelines and express speeds in miles per hour, while Pakistan's road network includes a heterogeneous mix of motorways, arterials, and unpaved

roads, with a vehicle fleet dominated by motorcycles, rickshaws, and overloaded trucks, and speeds measured in kilometers per hour (Ahmed et al., 2019). Enforcement regimes, emergency medical services, and road-user composition including a high proportion of non-motorized and mixed traffic further differentiate the crash-generation and post-crash trauma environment.

The injury severity outcome in this study is coded on the U.S.-specific KABCO scale (FHWA, 2006), which may not align with the injury classification schemes used in Pakistani hospital or police records. Moreover, the model's feature space incorporates several variables that are endogenous to the U.S. context, such as the Federal Highway Administration's roadway functional classification and certain vehicle body-type categories. It is important to recognize that simply removing these country-specific variables does not by itself establish geographical transferability. Transferring a model to a new context demands systematic recalibration, local data collection, and re-evaluation of feature relevance to capture the distinct crash mechanisms at play.

These limitations do not, however, negate the study's value. The research offers a transparent, replicable framework for extracting actionable insights from high-dimensional crash data a framework that can be adapted once comparable, high-quality datasets become available in Pakistan or other low- and middle-income countries. The contribution lies in demonstrating how explainable ensemble methods can bridge the gap between predictive power and engineering interpretability, thereby laying the methodological groundwork for future region-specific applications. Several departments in Pakistan have their own crash data record system, however the data access is extremely limited and obtaining unbiased data is difficult. Therefore the following study limitations exist:

i. The data comes from the secondary and is not an outcome of primary data collection.

ii. The data is not fetched to the system in real time and requires human input or computer vision for automation.

1.4.1 Rationale Behind Variable Selection

The systematic selection of variables from a comprehensive data repository like the Fatality Analysis Reporting System (FARS) is critical to ensuring that a model trained on international open-source records remains structurally and practically relevant to the localized traffic ecosystem of Pakistan. The primary challenge in utilizing a secondary cross-border dataset is the presence of region-specific or administrative variablessuch as distinct legal blood-alcohol thresholds, specific vehicular registration categories, or state-specific legislative codeswhich hold no functional equivalent or engineering utility within the Pakistani transportation framework. Therefore, an unrefined ingestion of all available parameters would not only introduce computational noise and overfit the machine learning models but would also severely diminish the real-world interpretability and policy application of the resulting Explainable Artificial Intelligence (XAI) metrics.

To circumvent this, a strict, dual-layered filtering protocol was established to isolate a highly generalizable subset of parameters. The first layer targeted environmental and physical invariants. Parameters governing atmospheric conditions (e.g., rain, fog, clear visibility), surface moisture conditions (e.g., dry, wet, water accumulation), and lighting variations (e.g., daylight, dark-lighted, dark-unlighted) represent fundamental physical interactions that dictate vehicular braking distance and driver visibility uniformly, regardless of geographical boundary.

The second layer evaluated structural and behavioral cross-correlations. Traffic patterns in Pakistan are heavily characterized by a heterogeneous vehicle mix, varying roadway architectures, and widespread gaps in passive safety compliance. Consequently, specific features were isolated based on their direct physical and behavioral proxy value:

i. Total number of travel lanes, alignment characteristics (straight vs. curved profiles), and type of harmful event or impact configuration (e.g., rear-end collisions, head-on impacts, rollovers) map directly to structural safety challenges encountered on major Pakistani arterials and motorways.

ii. Total occupant count (`N_PERSONS`), involvement of non-motorized road users (`PEDS`), and passive safety compliance metrics such as seat-belt usage (`REST_USE`) and airbag deployment configurations represent crucial behavioral indices. These elements directly mirror the high-severity risk factors prevalent across both urban and rural highway networks in Pakistan.

By restricting the feature matrix to these highly interactive variables, the study aims to reduce explicit administrative data biases. However, excluding direct geographic identifiers does not inherently render the resulting models geography-independent, as latent structural and behavioral differences remain embedded within the training data. Consequently, the SHAP-based feature importance profiles are not presented as universal mechanical and human risk relationships. Instead, they establish a data-driven hypothesis regarding collision dynamics and severity escalation. The geographical transferability of these risk profiles to developing transportation networks remains a critical task for future validation. Once validated and recalibrated against localized datasets, this methodological framework is intended to assist Pakistani transportation planners, highway engineers, and emergency responders in formulating targeted safety interventions.

1.5 Research Impact on Industry

1.5.1 Industrial and Academic Collaboration Roles

This research bridges the gap between academic data modeling and real-world industrial deployment by defining clear roles for stakeholders across both sectors. Academically, this study provides universities and research institutions with a transparent, verifiable framework for applying Explainable AI (XAI) to high-dimensional public safety datasets, moving beyond traditional "black-box" machine learning.

Industrially, the trained predictive architectures and SHAP feature profiles serve as an open-access foundation for software developers, automotive manufacturers,

and municipal authorities. By collaborating with transport corporations and highway management authorities, academia provides the optimized predictive engines, while industry supplies the operational infrastructuresuch as fleet management systems and digital dash displaysto test these models in real-world environments, accelerating the translation of academic code into commercial safety applications.

1.5.2 Research Novelty - Uniqueness

While extensive research has utilized transportation databases to predict injury severity, the independent selection of advanced algorithms does not inherently guarantee operational utility or methodological rigor. To address the requirement for a structured comparison with previous literature, it is necessary to contrast typical baselines with the specific framework developed in this thesis. Many broad studies aggregate all vehicular occupants which blurs the biomechanical lines between child passengers, unbelted rear occupants, and drivers. In contrast, this study strictly isolates the target variable to the maximum severity of the driver to methodologically standardize the physical seating position, active engagement, and proximity to primary safety systems. Furthermore, whereas previous research often conflates pre-crash prediction with post-crash outcomes, this framework is explicitly defined as an ex-post diagnostic classification.

The novelty of this work is further defined by its strict prevention of methodological leakage and its practical application. A pervasive limitation in existing imbalanced crash-severity literature is the misapplication of synthetic data generation globally prior to the train-test split, which invariably leads to artificially inflated performance metrics. This research implements a rigorous pipeline where oversampling is applied exclusively within a cross-validation framework on the training data to ensure the isolated test set is evaluated strictly on actual support weights. Finally, rather than using explainable artificial intelligence merely for general feature ranking within a well-documented highway network, this study explicitly applies these techniques to extract structural safety hypotheses. This establishes a practical diagnostic template designed to be transported to data-scarce developing regions where localized and granular crash datasets are currently unavailable.

1.5.3 Research Significance and Benefits

This reach focuses on the key questions asked by the transportation engineers, highway network designers, traffic control authorities, emergency response management and the research professionals. It helps in deep understanding of vehicle crashes and their severity levels in relation to multi factor deep patterns which go undetected from the traditional statistical tools.

1.5.4 Practical Implementation

Transportation and highway engineers will find this research helpful as it helps simulate the highway network for crash detection and severity predictions. It aids in benchmarking the highway network and helps the administration in updating the localized rules to improve the highway network safety. Similarly, the driving assistance apps can improve their functionality using hence trained models. These models can be used in active-sync with the vehicle to predict the driving risks.

Apart from direct use in transportation engineering, this research is very significant in Insurance & Forensics. The AI models can help the crash forensics to understand the crash dynamics, possible reasons and culprits, the involvement of human behavior and affect of factors not in control of the driver. As computer gaming is becoming as multi-billion dollar industry and it is now a separate industry in the entertainment world, modern games require AI to simulate real world like traffic and human behavior. This research helps developers design the game world and program the NPCs (Non Playable Characters) for a better open or arcade world.

1.5.5 National and Global Impact with SDGs Relevance

This research directly contributes to national safety frameworks and aligns with global sustainability initiatives. By converting historical accident repositories into actionable predictive insights, this study establishes a localized benchmark for proactive infrastructure management in Pakistan. On a global scale, this methodology directly supports United Nations Sustainable Development Goal 3 (Good

Health and Well-being), specifically Target 3.6, which aims to halve the number of global deaths and injuries from road traffic accidents.

Furthermore, by optimizing traffic risk evaluation and helping urban planners design safer arterial networks, the outcomes of this study support SDG 11 (Sustainable Cities and Communities) by making transport systems more inclusive, safe, resilient, and sustainable.

1.5.6 Research Challenges

The primary challenge encountered during this study relates to data disparity and the absence of a centralized, digitized road accident logging infrastructure within Pakistan. Because regional data is often fragmented, incomplete, or logged manually in incompatible formats, directly training high-dimensional machine learning frameworks on localized data was unfeasible. To overcome this, adapting an international open-source repository required meticulous feature filtering to isolate factors that are fundamentally geography-independent, such as environmental parameters, impact configurations, and human constraints, ensuring the models remain relevant to the Pakistani socio-demographic context.

1.5.7 Ethical and Management Considerations Including Risk Management

While mathematical transparency through explainable artificial intelligence is necessary, it does not inherently establish ethical acceptability for civil engineering deployment. The predictive framework must acknowledge demographic biases within the training data and recognize that excluding vulnerable road users limits its holistic application. Furthermore, transferring models across diverse geographic environments introduces severe domain shift risks which could misallocate critical safety resources. Because the cost of misclassifying severe trauma is asymmetrical, risk protocols must strictly prioritize minority class recall to prevent the under triage of critical crashes. Finally, to ensure accountability and data privacy, this

model must function exclusively as a decision support tool requiring the continuous oversight of licensed professionals before any operational use.

1.5.8 Research Deliverables

The core technical and structural deliverables generated from this research program include:

i. A systematically cleaned, preprocessed, and structurally filtered target dataset optimized specifically for predictive machine learning operations under geographic and behavioral proxy constraints.

ii. Three distinct, fully trained, and hyperparameter-tuned predictive model architecturescomprising Random Forest, XGBoost, and LightGBM frameworksconfigured for multiclass KABCO crash severity indexing.

iii. A comprehensive Explainable AI (XAI) deployment profiling matrix charting global and local SHAP (SHapley Additive exPlanations) values to isolate non-linear interactions of environmental and highway geometric variables.

1.6 Brief Methodology

The study draws on the U.S. Fatality Analysis Reporting System (FARS), a census of all motor vehicle crashes on public roads that result in at least one fatality within 30 days, rather than a sample or a general crash database. This sampling frame is appropriate for the research question because the target outcome of maximum driver injury severity on the five level KABCO scale is examined within events already severe enough to involve a death, enabling a rich contrast among drivers who sustain varying injury levels under extreme collision conditions. The analysis begins by subsetting FARS to a five year window (2005-2009) selected for its relevance to Pakistani traffic conditions, and further restricts the feature set to variables that are demographically and geographically portable, such as temporal factors, lighting, road surface moisture, number of lanes, impact type, and

seat belt use. Following exploratory data analysis and descriptive statistics to assess distributions and impute missing values, multiple ensemble machine learning models are trained, and SHAP based explainable artificial intelligence is applied to render the decision logic transparent. A key interpretive boundary must be acknowledged: because FARS conditions on the presence of at least one fatality, the sample consists of drivers who were involved in a fatal crash, and the occurrence of property damage only or minor injury outcomes within the KABCO distribution represents drivers who survived such an event, not a representative cross section of all road crashes. Consequently, the findings describe injury severity patterns within the restricted universe of fatal crash involved drivers and should not be extrapolated to non fatal collisions or to the general crash population, a limitation inherent to the sampling design that defines the scope of the study's conclusions.

1.7 Thesis Outline

The thesis is organized into five chapters as follows:

Chapter 1 provides the foundational context for the research, including the background of road traffic accidents (RTAs), research motivation, problem statement, and specific research questions. It outlines the overall goal and specific aims of the study, the scope of work and study limitations, the rationale behind variable selection, and the research impact on the industry. A brief summary of the methodology and the corresponding research deliverables are also presented.

The literature review in Chapter 2 outlines the nature, causes, and consequences of road traffic accidents and their severe global and localized socio-economic impacts. It presents a comprehensive overview of the existing research and knowledge base, reviewing historical trends in traffic engineering and delineating previous publications that bridge transportation safety with artificial intelligence frameworks. This chapter critically discusses the limitations of conventional statistical modeling methods and identifies the prominent research gaps that provide the structural basis for the research conducted herein.

The research methodology is detailed in Chapter 3, which lays down the roadmap for the research program in a systematic and categorical way. It describes data acquisition using the Fatality Analysis Reporting System (FARS) dataset and provides the engineering rationale for subsetting geography-independent behavioral and environmental parameters fit for the Pakistani traffic context. It further covers data preprocessing, the treatment of missing values, exploratory data analysis (EDA), and descriptive statistical testing. Finally, it explains the train-test splitting protocols, hyperparameter tuning, and the configuration of the comparative machine learning models along with Explainable AI (XAI) verification pathways.

Chapter 4 focuses on the results, analysis, and discussions of the predictive modeling frameworks. This includes reporting the statistical significance of the engineered features and evaluating model performance using multiclass metrics, where the Random Forest framework achieved an F1 Score of 0.82, followed by LightGBM at 0.79 and XGBoost at 0.78. This chapter extensively utilizes Explainable AI protocols to demystify model predictions, analyzing global and local feature impacts using SHAP bar plots and beeswarm plots, and contextualizes these behaviors with a practical case study extracted directly from the target dataset.

Chapter 5 presents the key conclusions drawn from the study and offers actionable engineering recommendations based on the findings. It outlines localized policy interventions, provides a structural roadmap for integrating predictive models into industry pipelines, and highlights the research's alignment with national safety frameworks and United Nations Sustainable Development Goals (SDGs). Lastly, it suggests avenues for future research to expand the evidence base across regional systems. References are provided at the end of the thesis.

Chapter 2

Literature Review

2.1 Background

Transportation is an integral part of a nation's economy, human activity and serves as a bedrock for socioeconomic interactions. Thus, lack and lag in transportation can severely recess the economic development [14]. The world has become a global village, by virtue of transportation [15]. Increasing city structures [16] and evolving cities and urban sprawl have changed the commuting dynamics. The routes have also evolved from radial routes to complex patterns just as complex land use [17]. The road traffic safety also evolved from road traffic safety and engineering to "capacity problem" to complex pattern problem [18]. This formed an insight into planning the transportation models with prior analysis and safety considerations. Thus, transportation engineers concerned themselves with road geometrics to ensure traffic safety along with the durability of the road structure. While Pakistan is heading towards safe cities, surveillance systems and intelligent traffic system in urban areas, as programs like Punjab Intermediate Cities Improvement Program (PICIP), Punjab Cities Program (PCP) and in the now, the Punjab Development Program (PDP) are underway, along with the Green Line Program which necessitates the predictive traffic engineering, it is now more imperative to study the traffic patterns with applicability on a general scale and considerations specific to

Pakistan. The local governments have incorporated engineers as Municipal Officers Services, charged with the responsibility to ensure traffic safety and response control [19].

2.2 Road Traffic Accidents

Pakistan has been experiencing the challenge of high frequency of RTAs where RTAs cause high fatality rates [20] due to reckless driving and narcotics [21, 22]. RTAs cause about 40,000 fatalities in Pakistan each year. [23]. It is pertinent to note that the lack of RTA registration and data provision forces the researchers to resort to the correlation between the Pakistani and the other countries' RTAs. There are two prevalent severity scales, namely the KABCO Scale and Abbreviated Injury Scale (AIS). KABCO is used by the traffic engineers and the first responders to assess the injury level and does not require a professional medical qualification, while AIS is used by the medical officials to ascertain the situation. Both the reporting officers and the traffic engineers use KABCO for crash severity related studies and decisions. For a transportation policy maker, it is important to know and mitigate the severity, therefore, the transportation engineers assess the road safety of designed highways to on KABCO. KABCO is briefly explained as below:

TABLE 2.1: The KABCO Injury Scale [24].

Category	**Description**	**Common example**
K	Fatal Injury	Death at scene or shortly after crash
A	Suspected Serious Injury	Fractures, unconsciousness and other signs of incapacitation
B	Suspected Minor Injury	Minor cuts and abrasions
C	Possible Injury	No visible injury, yet complaints of pain
O	Only property was damaged	No injury or pain, only property damage

2.3 Artificial Intelligence

Artificial Intelligence (AI) is branch of computer science which attempts to simulate human reasoning, while this is an ideal and hypothetical state (General AI), the real-world AI is termed as Narrow AI where models specific to problems are trained and merged. This is a shift from Rule-Based Systems (Expert Systems)

where a lead of experts sets rules to determine the outcome, e.g., "IF Speed > 80km/h AND Seatbelt = No, THEN Severity = Incapacitating to Learning Based Systems, where a sophisticated algorithm attempts to find complex patterns and uses them for predictions [24]. While the Expert Systems are rigid and cannot readily incorporate variable dimensionality of variables, the AI systems can easily adapt. Random Forst is one such example of machine learning models, it employs ensemble voting algorithms [25]. Random forest is a versatile and mature algorithm used in various industries and disciplines [26]. XGBoost on the other hand, uses gradient boosting, shrinkage, tree boosting and parallel processing [27]. These models are termed as ensemble models because they consist of multiple internal mechanisms working together for a final result and usually provide superior quality of results and are industry tested [28]. The difference between human intervened and artificial intelligence processing along with transformation of AI training is elucidated in Figure 2.1.

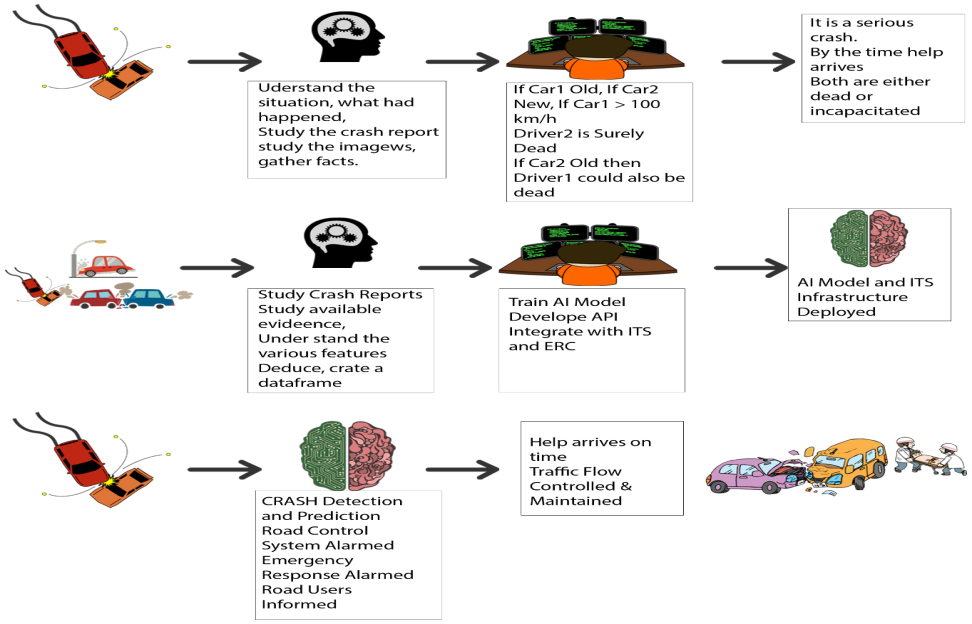

FIGURE 2.1: Understanding the Need of Artificial Intelligence in Crash Severity Prediction

2.3.1 Road Traffic Accidents and Artificial Intelligence

Over a century of RTA studies employing traditional statistical methods has led the transportation engineers to a scale of severity and traffic management, yet with

advancement in technology, both in construction of highways and control systems as well as in vehicles. This has led to a complex feature-interaction which cannot be readily processed with traditional statistical methods. Where, usually, the models rely on regression, which seeks to find linear and non-linear relationships among the events and the causes of events. In recent years, AI modeling technique, called decision trees has been found effective in determining the complex patterns. Later, the Gradient Descent method was improved in non-parallel algorithms. This was then revolutionized by the eXtreme Gradient Boosting or XGBoost.

The AI models are not generally human readable, they act as a black-box, which hides the underlying logic of what contributes, how contributes in a prediction and why a certain prediction is made in a certain way. Therefore, a modern research must address this problem by incorporating XAI. This study employs both Random Forest and XGBoost to train the respective models and compare with LightGBM. The models are trained on carefully selected parameters drawn from the Fatality Accidents Reporting System (FARS), U.S., relevant to the road and traffic conditions in Pakistan. This study employs SHAP (SHapely Additive exPlanations) as an eXplainable Artificial Intelligence (XAI) technique, to understand the influence of each parameter on the crash severity. SHAP can help improve on the model openness and reliability and discover previously unknown patterns.

The NHTSA is a US federal agency under the federal government and the department of transportation. NHTSA was created under the Highway Safety Act of 1970 to oversee safety programs that had earlier been managed by the National Highway Safety Bureau. The agency directs highway safety and consumer initiatives established by the National Traffic and Motor Vehicle Safety Act of 1966, the Highway Safety Act of 1966, the 1972 Motor Vehicle Information and Cost Savings Act, and subsequent amendments [29]. Like NHTSA, the NHA is the federal government agency in Pakistan, constituted under the federal ministry of communications in 1991. The National Highway Authority is charged with designing and maintaining the highways. It is responsible for implementing the geometric design standards, executing the safety features of the (barriers, signage, medians, emergency hotlines, safety isles), coordinating with the law enforcement and emergency

response operatives and conducting the Road Safety Audits. Both NHTSA and NHA work on transportation engineering principles.

2.3.2 Severity Taxonomy

To establish a rigorous foundation for this review, it is necessary to clearly delineate the operational definitions of severity utilized across the literature. Crash-level severity classifies the overall collision event based on the maximum injury sustained by any individual involved regardless of their seating position. Driver-injury severity explicitly isolates the physical trauma sustained by the operator of the vehicle which provides a direct link between pre-crash driver behavior and the resulting kinetic forces. Occupant-injury severity broadens this individual focus to include passengers by accounting for interior vehicle dynamics and restraint systems. Distinguishing these target variables is critical because predictive accuracy values from different studies cannot be compared directly when their underlying populations and outcome definitions fundamentally differ.

2.4 RTA Severity Studies Prior to Artificial Intelligence

Road traffic accidents are one of the three leading causes of unintentional deaths and cause millions of dollars in losses [30, 31]. Statistical tools can be used to understand fundamentals of RTAs [32]. However, RTAs are increasing with time [33], and urbanization. Urbanization closely reflects the transportation markers [34]. The increase in traffic volume and road accidents is also in direct proportion. Thus, urbanization, traffic demand and RTAs are inter-related [35]. For example, it has been studied that during the decade, 2002 till 2012, the roadside fatalities increased in China, especially in the rural areas, the fatality rate is more pronounced [36], with over 100,000 fatalities are reported due to road accidents each year [36]. Careless driving, speeding and going wrong lane and side by the drivers accounted for 92% of the major crashes in China, as [37] showed. The United

States experienced an alarming rate of 90% teen and male fatality due to on road all-terrain vehicles like motorcycles as the study [38] showed. It was found that RTAs increased from 2005 onwards with an increase of 16,000 fatalities from 2001 to 2007 due to cell phone use while driving [39]. A study over the period 1994 to 2012 showed that about 16 % of total fatalities on average were due to the inclement weather especially rain [40] which is an important revelation for studies to be conducted with a Pakistani Perspective.

The perceived safety has increased if compared with 1970, yet in reality it has not and there is immense pressure on transportation engineers to improve road safety [41]. It was found that marginalized communities and teens are more vulnerable to the road accidents in the United States [42]. This closely reflects the situation in Pakistan, as we have a large proportion of people living in deprived conditions, on foot and on road. In another study from 2006, in which FARS dataset was used, concluded that the road conditions were the prime cause of the crash and change in driving or vehicle capabilities may change the crash outcome [43]. In Pakistan the road conditions closely reflect the studied periods (1986), as the rural US was going under considerable development [44]. The road accident severity has seen more focus in the last two decades. Researchers have been applying approaches from statistical methods to understand and classify the road accidents. These approaches are limited in recognizing the cause and effect cycle of RTAs [45].

Dong et al. compared NGBoost, CatBoost, LightGBM, and AdaBoost using crash data from Pakistan's N-5 highway and found LightGBM to provide the strongest overall predictive performance, while SHAP identified temporal, driver, crash-cause, and collision-type variables as important predictors; however, the analysis was geographically restricted to a single national highway corridor, limiting broader spatial generalizability [46]. Sakib et al. compared seven tree-based ensemble methods for highway and non-highway crashes in Bangladesh, finding GBM superior in terms of G-mean and AUC and using SHAP to reveal differences in influential factors between road classes; nevertheless, reliance on police-reported records introduces potential underreporting bias, while the dataset lacked broader demographic, spatial, and infrastructure variables [47]. Chen et al. developed

an MSCPO-optimized XGBoost model using 4,287 Chinese crash records and reported improved performance over SVM, Random Forest, neural-network, and CNN baselines, supplemented by SHAP-based interpretation; despite the methodological sophistication, the relatively limited sample size, aggregation of the two highest severity categories, and absence of external temporal or cross-national validation constrain the demonstrated generalizability of the model [48]. Budzyski and Czerepicki analysed more than seven million road users in Poland using a harmonised five-level injury-severity target, XGBoost, and SHAP, identifying participant role, licence status, vehicle type, lighting, and road geometry as influential factors; however, the study concentrated primarily on one modelling architecture and one national administrative context, leaving comparative algorithm performance and geographical transferability largely untested [49].

Amiri et al. compared logistic regression, SVM, Decision Tree, and XGBoost together with SMOTE, ADASYN, random oversampling, cost-sensitive learning, and threshold moving, demonstrating that decision-threshold adjustment can outperform synthetic oversampling for minority-class prediction; however, the aggregated hospital dataset lacked detailed speed, environmental, infrastructure, and crash-context variables and could not be linked with police crash records [50]. Finally, Alobidan et al. synthesized the global literature across statistical, machine-learning, deep-learning, hybrid-AI, explainable-AI, and real-time prediction approaches, identifying persistent deficiencies in reproducibility, imbalance treatment, fairness, benchmark standardization, and cross-regional validation; however, the review adopted qualitative thematic synthesis rather than meta-analysis, meaning that reported performance levels from heterogeneous datasets, severity definitions, metrics, and validation protocols cannot be directly pooled or interpreted as evidence of universal algorithmic superiority [51].

2.5 Studies on Road Traffic Accidents using AI

Pakistan experiences a highly concerning frequency of road traffic accidents resulting in significant fatality rates. This is frequently attributed to reckless driving

and substance abuse. These collisions cause tens of thousands of fatalities annually. Due to the historical scarcity of comprehensive local collision registries, researchers often rely on established international scales to quantify trauma. The KABCO scale is universally utilized by traffic engineers to assess injury levels without requiring professional medical qualifications, categorizing outcomes from fatal injuries down to property damage only. In the United States, safety initiatives are governed by the National Highway Traffic Safety Administration. Similarly, the National Highway Authority in Pakistan executes geometric design standards and coordinates safety audits.

Historically, engineers relied on traditional statistical methods and expert rule-based systems to analyze collisions. However, the increasing complexity of vehicle interactions and highway features necessitates advanced approaches. Machine learning algorithms, particularly ensemble models like Random Forest and XGBoost, simulate complex reasoning to discover non-linear patterns in multi-dimensional data. Because these advanced models operate as opaque algorithms, modern research must incorporate Explainable Artificial Intelligence to understand how parameters influence predictions.

2.5.1 Video Based Crash Severity Detection

In a three stage study [52], YOLO technique was used for crash detection in videos achieving 89% accuracy but it required precise video recording. Thus, another limited study on improving computer vision using YOLO V3 was applied, which still requires active recording, feeding and analysis [53]. A study using Mask R-CNN for improved object detection followed by a centroid based surveillance and tracking algorithm found the probability of a road traffic accident dependent of speed and trajectory abnormalities. However, this method does not predict the vehicular collision [54].

A collaborative study shows that the main constraints for the computer vision in vehicles are the processing power and limited hardware. The image processing power inside a camera or a car is significantly lower than a low-end laptop, let

alone a full-scale AI system. This limits the applicability of the computer vision in the cars and requires upgradation of each car on the road [55].

In yet another study [56], aimed at collision detection, presents a framework relying on computer vision for detecting road traffic accidents (RTAs) through surveillance/CCTV cameras. The framework is comprised of five interconnected modules. This information is then communicated to the emergency services using the onboard GSM module. However, the framework comes short when it comes to predict the crash severity in general. It only detects the crash once it has already occurred. Therefore, it cannot be used for scenario tests and predictions.

These models require video stream or recorded video. These cannot be used for simulation purposes, analysis and design of highway systems. This also requires an active camera and a computer if to be used inside a vehicle, which requires hardware and design level upgrades in vehicles. Furthermore, even if installed inside a vehicle or a system the reaction time is too small to be practically beneficial, as it requires, video input, processing and prediction, which is data intensive and puts more load on bandwidth.

2.5.2 Data-Based Crash Severity Prediction

Some studies have evaluated SVM and ANN for an accuracy of 73% [57] for crash severity detection [58]. While another study with twenty most important variables, employed ensemble of Random Forest and CNN [45] to assess the performance of one over another in predicting the crash severity. The studies were limited to a narrow variable set.

According to the article [59], published by Oxford University Press, predicting traffic accident severity plays a pivotal role in optimizing dynamic traffic safety management. The models were trained, achieving 78.9% precision and 88.4% accuracy in forecasting accident severity. However, the study is limited to the mountain freeways and does not seek application on the other types of terrain. In their research, "Classification and Prediction of Driving Behavior at a Traffic Intersection Using SVM and KNN" [60], the researchers express their concern that

the road traffic accidents keep occurring at the traffic signals despite constant traffic patrolling and awareness campaigns. In it, SVM and KNN are employed to validate the classification of driving behavior as either safe or unsafe stopping at signalized junctions when the yellow light turns on. However, the study is limited to junctions and does not incorporate generalization. Moreover, the study is limited to the behavior being safe or unsafe, rather than predicting the crash or severity.

Ibrahim Khalil Umar from the University of Technology Babura, Nigeria obtained the dataset through surveying the truck drivers and gathering data from 248 drivers, in his article, [61]. He then trained models using Nave Bayes, support vector machine, and K-nearest neighbor methods of machine learning. The primary goal of the study was to predict truck drivers' involvement in road traffic accidents. The study shows that the in 52% of the crashes, the driver was directly involved, and the most significant factor remained the sleeping on the wheels. The study however lacks the generalization and uses extremely limited dataset.

In another study [62] focusing Brazilian traffic signs, the research team aimed to propose a traffic signs recognition system, employing the techniques of digital image processing (DIP), and temporal coherence. The team stated the focused scope of only detecting the speed limit signs, not all signs collectively. The system simply ignores the no overtaking and other signs. Their work attained the accuracy of 82% using SVM. However, this research is incapable of preemptive prediction of crashes and severity.

Yet another effort to contain the road traffic accident problem, the article [63] titled, "System for Predicting Traffic Violations Using Machine Learning" sheds light on the traffic violations and control. The study uses Logistic Regression, Random Forest and SVM for making behavioral predictions. The researchers developed a system to predict traffic violation by the driver. They noted that road accidents continue to rise due to neglecting traffic rules by the drivers. Though the study reflects the problems found in the US and Pakistan, the scope is limited and the study overlooks the contributing causes like the weather and environmental conditions. Yet another article [64], titled "Monitoring Physiological State of

Drivers Using In-Vehicle Sensing of Non-Invasive Signal" the authors offer an in-vehicle heartmonitoring system. Multiple models were trained. While this study aims to reduce the crash severity, it is limited to the one health condition only. It is not predictive and it does not apply to all age groups and requires bulky hardware and software installation. Just as the research mentioned above focused on singular health condition and presented a focused study, another article suggests that we should understand and keep track of driver's posture. The said research, [65], titled, "Estimation of driver's posture using pressure distribution sensors in driving simulator and on-road experiment" presented the reported that the sensor sheet-based approach yields highly accurate predictions of driver pose. While effective and focused, the research is limited to one function only. It does not take other factors of accident into account. Furthermore, the research is not suitable for Pakistan as the probability of drivers buying and using such equipment is dim. A LightGBM based study, [66] focused on the fatigued drivers to detect mental states of drivers quickly and accurately. Although the results confirm that LightGBM delivers realtime classification accuracy that is both more efficient and more accurate than existing methods, the research addresses only a particular facet of the whole. In the real world, road traffic accidents are caused by a myriad of factors, most importantly the road conditions, weather conditions and socio-psychological factors.

In another finding made by the authors of the article [67], "Road Accident Detection using SVM and Learning: A Comparative Study" it is recognized that the road traffic accidents claim many lives worldwide, especially among children and young adults aged five to thirty years old. The results show that both CNN and SVM achieve strong performance while R CNN lags behind in this setting. While this paper exhibits the need for detecting crashes and minimizing severity, the models are not severity predictors and are limited to the availability of CCTVs and proper footage. Concerned with the crash severity in Ethiopia, the article [68] published with title, "Predicting Car Accident Severity in Northwest Ethiopia: A Machine Learning Approach Leveraging Driver, Environmental, and Road Conditions" discusses the car accidents in Northwest Ethiopia. The study confirms that a robust machinelearning model can support roadsafety planning

in the region. However, the study uses limited variables, and a small dataset of only 2000 crashes of which most are taken from the city administration and cover only the urban areas. The importance of road signs is paramount and the need for their maintenance is rightly understood, studied and reported by the authors in their article [69], "Towards Smart Maintenance Machine-Learning Based Prediction of Retro reflectivity and Color of Road Traffic Signs". The study proposes an endtoend predictive maintenance system that keeps road traffic signs visible for a longer duration, thereby seeking to reduce accidents, waste, and material consumption. It uses data from Sweden, Denmark, and Croatia. The study builds regression and classification models using Random Forest, SVM, and ANN with the aim to forecast a sign's retro-reflectivity and color decay in addition to its overall longevity. In this study, age emerges as a key predictor, with chromaticity increasing linearly over time except for yellow signs that show nonlinear patterns between eight and twentytwo years. While the study aims to enhance the safety environment, it is limited to traffic signs only. It does not take into account the various environmental effects and does not predict any crashes or severity.

A study based on Sri Lanka, titled, "Integrated analysis of skid resistance, texture, and geometric factors for enhanced road safety on the Southern expressway" focuses on the road surface and investigates how skid resistance, texture, and geometric variables influence accident severity on Sri Lanka's Southern Expressway. Using data collected from 2012 to 2022 by the Expressway Operations, Maintenance & Management Division. To isolate key contributors, the study applied Principal Component Analysis and Kmeans clustering before training predictive models based on Random Forest, logistic regression and support vector machines. The Random Forest model achieved an accuracy rate of 91.5%, confirming that road curvature, gradients and lighting are strong predictors of accident severity. The research advocates for integrating realtime traffic and weather data into GIS-based hotspot analysis, enabling timely interventions and better safety management on the expressway. Though the research [70], takes surface conditions into account, it is limited to Expressways only. Also, it does not take weather conditions into account nor does it consider the driver's behavior and demographics.

The research published by the World Health Organization, TRL and other organizations have identified the road crashes as the most concerting cause of deaths, particularly in the developing world. The awareness of seriousness and need of the crash study is reflected in the formation of an association contributed by the private, civil and government sectors, called Global Road Safety Partnership (GRSP), with aim to reduce road traffic accidents worldwide. The GRSP is aided by TRL's study objectives as below:

i. To study and estimate the road crash fatalities, on global and local scale.

ii. To determine the crash costs with respect to the Gross National Product (GNP)

iii. To understand the regional trends of road traffic accidents and related fatalities.

iv. To identify the current road traffic accident and fatality rates along with the risk rates (deaths per 10,000 vehicles and deaths per 100,000 population respectively) in addition to the severity trends by age, sex and road user types [71].

While most of the population and economic growth is occurring in the developing countries, there is no comprehensive strategy in place to address the serious concerns of roadside safety and efficiency of the urban transport system in these countries [35].

2.6 Methodological Synthesis of Contemporary Studies

Road traffic accidents represent a leading cause of unintentional mortality worldwide. While urbanization and traffic demand proportionally increase collision rates, specific regional factors such as rural road infrastructure, demographic vulnerabilities, driver distraction, and inclement weather heavily influence severity

Table 2.2: Summary of Literature Review on Crash Severity Modeling

Reference	**Geography**	**Dataset and Sample Size**	**Severity Target**	**Preprocessing and Imbalance Strategy**	**Validation Design**	**Key Performance Metric**	**Principal Limitations**
Dong et al.	Pakistan	N-5 Highway	Crash-Level	Not explicitly detailed	Unspecified	LightGBM Superiority	Geographically restricted to a single corridor
Sakib et al.	Bangladesh	Police Records	Crash-Level	Tree-based splitting	Hold-out Set	GBM AUC Superiority	Potential underreporting bias and limited infrastructure variables
Chen et al.	China	Regional Data	Occupant-Level	MSCPO Optimization	Hold-out Set	Improved over baselines	Small sample size and aggregation of highest severity classes
Budzyski et al.	Poland	National Registry	Harmonized 5-Level	XGBoost internal weighting	Unspecified	Identified influential factors	Lacks comparative algorithmic performance
Amiri et al.	Unspecified	Hospital Registry	Unspecified	SMOTE and Threshold Moving	Cross-validation	Threshold outperformed SMOTE	Lacked detailed infrastructure and crash context variables
Alobidan et al.	Global	Literature Synthesis	Variable	Qualitative Review	Not Applicable	Not Applicable	Heterogeneous datasets prevent pooled meta-analysis

outcomes. Early severity studies relied heavily on statistical regression, but contemporary research has shifted toward advanced machine learning paradigms.

While various studies have explored video-based detection systems and specialized physiological sensors, these technologies detect events concurrently rather than predicting severity preemptively. Data-driven algorithmic studies offer superior predictive utility for transportation planners. To systematically evaluate the state of the art in severity prediction, a synthesis of the most relevant algorithmic studies is presented in the table below. This structured comparison demonstrates why performance metrics cannot be compared directly when methodologies fundamentally diverge.

2.7 Summary

Various studies have been performed on RTAs, [52–54], have incorporated video streaming for vehicle and collision detection, but, these studies are limited to video only, these require a constant or intermittent stream and specialized hardware as in case of [55, 56]. Furthermore, these methods do not predict crashes or crash severity before these happen and cannot be used for scenario based simulation or analysis of road networks. Data driven studies using various machine learning models have been performed but some are limited to a specific terrain [59] or employ a very narrow set of variables [57, 58]. Similarly, Support Vector Machine (SVM) and KNN have been used but do not incorporate generalization or predict only safe or unsafe, instead of predicting crash severity [59–62]. Although [63] has identified some variables in SVM, Logistic Regression and Random Forest, their study does not entertain weather and environmental conditions just as [64] does not incorporate all age groups. Similarly, various other studies exist but are limited in their scope, variable selection or terrain. Therefore, a research incorporating a complex mix of relevant variables targeting Pakistani road and traffic conditions is necessary to predict crash severity. Similarly, some researches have considered only one aspect e.g., driver's posture [65], or driver's fatigue [66]. Other researches have identified the need for predicting crash severity with a more

wider range of features. Therefore, a study encompassing various highway traffic, road geometrics, lighting conditions, presence of pedestrians and other features is highly awaited in industry.

Chapter 3

Methodology

3.1 Background

Highway and transportation engineering require a comprehensive understanding of complex safety and traffic models. In addition to road geometrics and structural design, factors such as human behavior, weather conditions, traffic mix, and lighting must be thoroughly analyzed to evaluate highway networks. Developing such robust models requires meticulously recorded crash data collected over many years. Due to the unavailability of highly granular historical crash datasets in developing regions, this study utilizes the reliably collected Fatality Analysis Reporting System data from the United States. While the feature space was engineered to retain variables such as temporal factors, road functional classes, and the number of lanes, it is critical to maintain a strict conceptual distinction between model development and geographic transferability. Feature deletion and selective engineering do not inherently transform a foreign dataset into a representative sample of Pakistani traffic conditions. The selected data provides a robust mathematical environment for developing diagnostic machine learning models and extracting structural safety hypotheses, but the resulting numerical weights remain fundamentally tied to the infrastructural and behavioral context of the original data source. Therefore, the application of this framework to Pakistan is presented as a methodological blueprint rather than an immediate predictive solution. To

successfully support the transfer and operational deployment of this model in a new geographic setting, empirical evidence in the form of localized crash data is strictly required. Future research must collect domestic data encompassing local geometric designs, vehicle fleet characteristics, and post crash medical response variables. This localized evidence is necessary to recalibrate the model parameters, adjust the feature weights, and formally validate the algorithmic risk profiles against actual regional outcomes.

With this framing in place, the data preparation pipeline was designed to ensure that no information from the test set leaks into model training. The study draws on FARS, a census of fatal crashes, because it provides a large, standardized, and publicly available dataset with detailed crash, vehicle, roadway, and occupant variables. A five year window (2005-2009) was selected to approximate the vehicle fleet characteristics still common on Pakistani roads, and feature engineering was guided by a deliberate effort to retain variables that are either inherently region invariant or closely relatable to Pakistani highway conditions, such as temporal factors, lighting, road surface moisture, number of lanes, and impact type. Before any splitting, each FARS variable was examined against the official data dictionary to interpret codes for "unknown," "not reported," and "not applicable." Rather than automatically treating these codes as ordinary missing values, they were assessed individually: where an "unknown" code represented a distinct and potentially informative state (for example, seat belt use unknown), it was preserved as a separate category in the analysis; only values that were genuinely missing or that, according to the dictionary, carried no analytical meaning were flagged for imputation later in the pipeline. After this code interpretation and the subsequent removal of features that could not be meaningfully mapped to the target context, the cleaned dataset was split into training and testing sets. All remaining steps that involve estimating parameters from the data were applied exclusively within the training set or within each training fold. Imputation of the small number of genuinely missing values was performed on the training set, and the learned imputation parameters were then used to transform the test set. Categorical variables were encoded using the same training set based approach, and no resampling was applied to the test data. Model hyperparameters were tuned

via repeated stratified k fold cross validation on the training set, further insulating the test set from tuning decisions. This pipeline ensures that the Random Forest, XGBoost, and LightGBM models were trained, evaluated, and compared strictly on unseen data. After model training, SHAP (SHapley Additive exPlanations) was employed to interpret the decision boundaries, and the results were compiled for discussion, analysis, and conclusions. Exploratory data analysis was performed on each variable to gain insight into data quality and distributions; chi square tests and Kruskal Wallis H tests were used for exploratory characterization of associations before modeling, consistent with the preventive measures against data leakage.

3.2 Toolset Selection

3.2.1 Python, Pandas and Sci-kit Learn

Python is a general purpose computer programming language, optimized for quality, productivity, portability and integration [66]. It is an interpreted, object oriented and high level language [67], developed in the early 1990s. Python is an open source project [68]. Python has interpreters for various platforms, namely, CPython, Jython and IronPython where CPython is the most widely used python implementation [61], written in the C programming language [62]. Therefore, python with its implementation CPython was selected. Pandas is also an open source project [69], the name comes from Panel Data [70]. It supports various file types and data structures e.g., xlsx (Excel), comma separated values (csv) and more [71]. It has a zero-based index [72] most suitable for array handling in programming languages. Pandas is more suitable for machine learning than its alternatives like Apache Arrow [73]. Scikit Learn (SK-Learn) is also a free and open-source python library, designed to interoperate with Numpy and SciPy seamlessly [74]. It provides an Application Programming Interface (API) with a consistent user interface. Therefore, Python, Pandas, and Scikit Learn were selected for this study.

3.3 Dataset

Understanding datasets is as much important as the data itself. In order to make data-informed decisions, such as highway design implementation, highway geometric design and facility design the comprehension of data is paramount. The data comes from the official U.S. data sources available publicly for anyone's use. This specific dataset is termed Fatality Analysis Reporting System (FARS), it covers the whole of the United States, Columbia, Puerto Rico. Also it can be used as benchmark study for any landscape with full or partial similarities [75]. FARS is the most referenced vehicle crash data outside the United States [76]. Pakistan exhibits highly heterogeneous roadway environments, including motorways, national highways, urban arterials, rural roads, and mountainous terrain. A crash database intended for surrogate modelling should therefore encompass a similarly broad spectrum of roadway and geographic environments. FARS was selected because it is a nationwide census of fatal crashes occurring on all public trafficways across the United States, providing comprehensive coverage of diverse roadway, environmental, and regional conditions. Unlike sampled datasets, FARS supports analyses across these varied landscapes without sampling uncertainty.

3.3.1 Crash Data Reports

The source of the data is the police reports. These reports are responses to the questionnaires. Each Year's data folder contains multiple data files of comma separated values (csv) format. Each file consists of data entries as made at different stages and offices and hence overlap. For example the Year, Weekday, Hour, Time and similar temporal features are in most data files because each serves as a record of a "response" filled by an official. The official FARS manual states, "Nineteen of these data files contain only one or two data elements and the analyst can code more than one response for these elements (i.e., "select all that apply"); thus, there is a record for each response." [77]. Since a vehicle crash involves multiple state agencies both in Pakistan and the United States, such as law enforcement (highway police, state police), emergency responders (highway responders/first responders

and medical care responders), insurance, department of statistics, judiciary, trade and transport and others, the data folder contains multiple files. The primary of these files are mentioned as under:

3.3.1.1 The Accident Data File

The ACCIDENT.CSV contains one row per crash, identified by ST_CASE, so in terms, while the PERSONS.csv contains one row per person, with a duplicate of ST_CASE, for example, if crash with ST_CASE 10024 involves one driver and one pedestrian, the PERSONS.CSV contains two records for 10024, one for driver and one for pedestrian. The ACCIDENT.CSV contains information of number of persons as PERSON, pedestrians as PED. In order to get the number of drivers in a crash, the PER_TYP was used to identify the driver. The final dataset contains exactly one row per crash, the identity column ST_CASE was dropped. Therefore, no crash-level grouping was required during the train-test splitting.

3.3.1.2 The Vehicle Data File

It describes the vehicles involved in the crashes, both active and passive, i.e., whether the vehicle is involved in in transportation or active roadway or just parked around the corner. The reason for inclusion of the details of parked and off-road vehicles is judicial, financial and insurance related. Furthermore, the vehicle file contains both crash and legal features along with pre and post crash information. Therefore, the vehicle data file needs to be carefully understood, traversed and filtered before use in any study.

3.3.1.3 The Person Data File

Like vehicle data file, the person data file contains information about each of persons involved or linked to a crash, whether or not the person was at fault or victim. Thus so much of the information like temporal and environmental data is either duplicated due to the consistency or multiple and multi-stage reporting. Therefore, like vehicle data file, the person data file needs to be carefully audited before

inclusion in any kind of study. The feature extraction, selection, and engineering is discussed in detail in accompanying Jupyter Notebooks.

3.4 Feature Engineering

The feature engineering improves model performance and usability. It helps understand the underlying patterns, confirmation of patterns and identifying solutions to problems identified or under analysis. The feature engineering uses several methods, for example, feature selection, extraction, interaction terms, storing categorical variables, and dimensionality reduction techniques. It helps models learn better by finding the most important features and excluding the features that do not contribute in model prediction. This also cuts down on overfitting and makes model's decision calculations faster. In this study, the features irrelevant to Pakistan and classes not applicable to the Pakistani traffic were eliminated as described in the previous sections.

3.4.1 Selection of Time Span for Road Traffic Accident Data

The study uses a five year FARS window from 2005 to 2009. This period was chosen because it captures a U.S. vehicle fleet whose average characteristics, particularly the of older, less technologically advanced vehicles without modern active safety systems, approximate the composition of the current Pakistani fleet. Evidence for this approximation comes from Pakistan's vehicle import and registration data, which show that the median age of the passenger car fleet exceeds 15 years, with a large share of vehicles imported as used units from Japan and Europe already being several years old at the time of entry (PAMA, 2022). Consequently, the safety features, structural designs, and restraint systems present in the 2005-2009 U.S. fleet are broadly comparable to those still operating on Pakistani roads today. This temporal mapping is not intended to assert that all traffic conditions, driver behaviors, or infrastructure standards are equivalent; rather, it provides a

defensible basis for selecting a source dataset whose vehicle level variables reflect a technological era relevant to the target context.

3.4.2 Feature Selection and Correlation with Pakistan

The correlation does not end with the time span, for a better reflection of Pakistani conditions, only the features reflecting Pakistani situation are selected. For example, the feature of Drunk/Alcoholic driver is not selected. For any possible value of the feature which does not appear on the Pakistani road and traffic, the entire record/row is dropped. A comprehensive summary of each feature is presented here. The details of data manipulation and results are provided as part of Jupyter Notebook along with the data for reproducibility.

The MONTH column shows the enumerated Month of Crash. These features exist as duplicates in other files for sake of readily available information. These features were considered only once. In PARKWORK (Parked or Working Vehicles) data file, it is named as PMONTH and so on. However, the value of a particular crash are the same. The months are enumerated with 1 based array. Mapping 1 to January and 12 to December, the unknown values are marked with 99. This column holds temporal data and is not region dependent. Day of Crash is a repeated column. It represents the day of month on which the crash occurred, it is the day in the date, with valid values ranging from 1 to 31 and 99 for unknown values. This column holds temporal data and is not region dependent. Hour of Crash holds a valid range 0 to 23 (hours) and 99 for unknown hour. Like Month and Day, Hour is also a repeated column in all files. Day column holds temporal data and is not region dependent. Like hours, minutes are also numeric and represent the actual data (non-encoded). Minutes hold valid values from 0 to 59 and 99 for unknown/missing values. This column holds temporal data and is not region dependent. PERSONS column holds the number of "persons" involved in the accident. It is also a representation of actual date. Like PERSONS, PEDS are also the people who were involved in the crash, but PEDS column holds only those who were outside the vehicle in transport (moving). These exclude the people who were either driving or were passengers of the vehicle. In Pakistan, a number of road

traffic accidents involve the bystanders or pedestrians. This data closely relates to the situation in Pakistan.

NO_LANES (Number of Lanes on Highway) data column holds the number of lanes on the highway at time of crash.

Table 3.1: Categories under NO_LANES

CODE	LANES
0	Non Trafficway
1	One Lane
2	Two Lanes
3	Three Lanes
4	Four Lanes
5	Five Lanes
6	Six or More Lanes
7	Seven or More Lanes
8	Not Reported
9	Unknown

LGT_COND (Lightening Condition) data column describes the type/level of light that existed at the time of the crash as indicated in the case material.

Table 3.2: Categories under LGT_COND

CODE	LIGHTING CONDITION
1	Daylight
2	Dark
2	Dark - Not Lighted
3	Dark - Lighted
4	Dawn
5	Dusk
6	Dark - Unknown Lighting
7	Other
8	Not Reported
9	Unknown

As the name suggests, the SP LIMIT (Speed Limit) data column holds the designated speed limit on the highway section related to the crash, at the time of crash. The speed is given in miles per hour, however, Pakistan uses kilometers per hour, so, miles per hour were converted into km/h in the final dataset.

Rest of the features are explained in Appendix A.

TABLE 3.3: Categories under SP_LIMIT

CODE	SPEED LIMIT
0	No Statuary Speed Limit
1 to 152	Speed Limit in kmh
998	Not Reported
999	Unknown

3.4.3 Data Transformation

The data transformation is a critical stage, in which the data is transformed into a format fit for use analytically. It is done through normalization, scaling, aggregation and encoding. This process ensures consistency and organization of data, enabling it to be fit for statistical or machine learning models. In this case, the data is meticulously fine grained and encoded at the time of data entry. However, to make it suitable for use in machine learning, aggregation was employed.

3.5 Data Preprocessing

Data preprocessing holds great significance in both data analysis and machine learning processes. It involves transforming raw data into a cleaned, organized, and prepared format suitable for analysis. Real-world datasets often contain flaws like missing values, irregular formatting, redundant records, irrelevant data, and noise, which can undermine the reliability of statistical models and machine learning outcomes. By meticulously cleaning and standardizing the data during preprocessing, we eliminate these inconsistencies, ensuring that subsequent analyses are based on accurate and coherent information. This rigorous preparation step enhances the credibility of study results and ultimately contributes to more robust and reliable insights. The details of data preprocessing are presented in Appendix-B.

3.5.1 Handling Missing Values

In the FARS database, values that are unknown, not reported, or not applicable are recorded using explicit numeric codes defined in the data dictionary, rather than left blank. For the present study, each variable was first examined against the FARS manual to interpret these codes. Codes that carry potential information (e.g., "seat belt use unknown" coded as 98) were preserved as separate categories, whereas codes that definitively indicate the absence of usable information (e.g., "not reported" when the variable was simply not collected) were treated as genuinely missing and flagged for imputation. All subsequent imputation decisions were based solely on the training set. The test set was held completely untouched, and imputation parameters learned from the training data were applied to the test data without modification.

After the code interpretation step, three imputation strategies were selected on a per-variable basis, depending on the measurement scale and the observed distribution in the training data.

Mode imputation was applied to categorical variables with low cardinality and a strongly dominant category. For such a variable, the most frequent category (the mode) was used to replace all missing entries. This assumes that the missing values are missing at random conditional on the observed data and that the mode provides a reasonable point estimate.

Mean imputation was used for continuous variables whose training set distribution was approximately symmetric and unimodal, under the same missing at random assumption. The missing entries were replaced with the arithmetic mean computed from the observed cases in the training set.

Ratio imputation was designed for variables with high dispersion, no clear central tendency, and a distribution that was nearly uniform or highly heterogeneous. Let X be a categorical variable with K observed categories in the training set, with counts $n_1, n_2, \ldots, n_K$ and total $\mathrm{N} = \sum n_K$

The ratio imputation procedure estimates the probability vector:

$$\mathbf{p} = (p_1, p_2, \ldots, p_K); p_k = \frac{n_k}{N} \tag{3.1}$$

which defines the empirical distribution of X among the observed cases. For each missing entry in X, a category k is drawn randomly from the multinomial distribution Multinomial$(1, \mathbf{p})$. When the target variable (driver injury severity, KABCO class) was used to inform the imputation, separate probability vectors $\mathbf{p}_y$ were estimated for each severity class y from the complete cases within that class, and a missing value was imputed by drawing from the distribution corresponding to the record's actual severity class. This class-conditional variant preserves the observed association between the feature and the outcome, under the assumption that the relationship estimated from complete cases also holds for incomplete ones.

Algorithmically, the procedure proceeds as follows. For each variable to be imputed via the ratio method, the training set frequencies (overall or class-conditional) are computed once and stored. For the training set itself, missing entries are filled by independent random draws from the stored multinomial distribution(s). For the test set, the identical stored probability vectors are used to impute any missing entries, ensuring no information leakage from the test data into the imputation model. In the present study, a single random draw was used per missing entry (single imputation). No multiple imputation was performed; therefore, the standard errors of the model estimates do not incorporate imputation-related variability. This limitation is acknowledged, and the stability of the downstream feature importance rankings was checked by comparing the SHAP results against a complete case analysis as a sensitivity check.

The choice of imputation strategy for each variable was guided by exploratory data analysis on the training set, including measures of dispersion such as Gini impurity and entropy. These measures were used only as descriptive summaries of the shape of the observed distributions (e.g., to identify variables with near-uniform distributions that would benefit from ratio imputation). They were not used to validate the imputation; the adequacy of the imputation was assessed indirectly through the consistency of the model evaluation metrics and the SHAP-based feature rankings under different imputation choices. A detailed per-variable

imputation plan, including the strategy applied and the training set statistics used, is provided in Appendix B.

3.5.2 Descriptive Statistical Analysis of Dataset

For a deeper understanding of the underlying distribution and central tendencies of any large-scale observational dataset, the descriptive statistical analysis serves as the foundational step. It systematically evaluates individual variables, such as the exact MINUTE of an occurrence. Researchers identify temporal patterns, anomalies and data quality before advancing to complex modeling. Thus, producing comprehensive frequency tables is a critical initial procedure. The statistical analysis tables report the frequency, which quantifies the raw absolute count of instances recorded for each specific minute, alongside the total percentage, which represents this count as a fraction of the entire dataset, including any missing or unrecorded entries. Importantly, the researcher must isolate the valid percentage, which is unlike the total percentage. The valid percentage adjusts the computational baseline by excluding the missing values or corrupted data points, hence providing a mathematically accurate reflection of the variable's true distribution within the observable data. Descriptive statistical analysis is performed after imputing the missing or unknown values to ensure the proper imputation and assess the data quality. The descriptive statistical analysis of each feature was performed and is presented feature-wise.

3.5.3 Feature Significance Testing

In research, it is very important to analyze the cross-classified categorical data with Karl Pearson's Chi-Square tests for investigating the associations or differences between categorical features. Thus, it is paramount to assess the statistical significance of each independent variable prior to and after modeling. To assess the statistical significance, two widely used hypothesis testing methods were used; for categorical variables, chi-squared and for numerical variables, Kruskal-Wallis

H test. These two tests serve as a guideline for feature selection, and ensure that only the features significant statistically are used in the modeling phase.

3.5.3.1 Chi-square Test for Categorial Variables

To determine the whether there is a significant association between two categorical variables, Chi-Square test of independence is a non-parametric statistical hypothesis test used. It is applied for feature selection, to analyze and assess whether the occurrence or frequency of a specific categorical feature is independent of the categorical target variable. Since, the features with high dependence are theoretically the most informative for classification tasks, the observed frequencies of target feature class combinations are compared against the frequencies if the feature class was independent, thus the test identifies the features that are highly dependent on the target class. Simply put, the Chi-square test assesses the significant correlation between two categorical variables. For example, between a categorical features (variable) Road Surface Condition (Dry, Wet) with Crash Severity Levels.

i. Chi-square Test Hypothesis

   a. Null Hypothesis (H_o): There exists no correlation between the two categorical variables.

   b. Alternative Hypothesis (H_1): A statistically significant connection exists.

The test statistic (X^2) is computed by taking the squared difference between the observed and expected frequencies, divided by the expected frequencies for all categories:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3.2}$$

Where:

r = Number of rows (categories in the feature).

c = Number of columns (categories in the target variable).

O_{ij} = The observed frequency in the $i\text{-}th$ row and $j\text{-}th$ column.

E_{ij} = The expected frequency in the *i-th* row and *j-th* column, calculated as:

$$E_{ij} = \frac{\text{Total of row } i \times \text{Total of column } j}{\text{Grand Total of all observations}} \tag{3.3}$$

The resulting X^2statistic is compared against a Chi-square distribution with (*r*-1)(*c*-1) degrees of freedom. Consistent with standard hypothesis testing practices, the significance threshold (alpha) was set at 0.05. Categorical features that returned a p-value <0.05 allowed us to reject the null hypothesis of independence. Consequently, these features were considered statistically significant and selected for inclusion in the final feature subset.

3.5.3.2 Kruskal-Wallis Test for Numerical Variables

To determine if there are statistically significant differences between two or more independent groups, the Kruskal-Wallis H test is a rank-based, non-parametric statistical test is performed. It is utilized as an alternative to the One-Way ANOVA when the assumption of normality is violated. The K-Wallis H test evaluates whether a continuous or ordinal independent feature is identically distributed across the different categorical classes of the target variable. If the distribution of a feature varies significantly across the target classes, it implies that the feature possesses strong discriminatory power and is valuable for predictive modeling.

The calculated H statistic approximates a Chi-square (X^2) distribution with *k*-1 degrees of freedom. For this analysis, a standard alpha level (α) of 0.05 was established as the threshold for statistical significance. Therefore, any continuous or ordinal feature yielding a p-value <0.05 was regarded to have a statistically significant relationship with the target variable and the feature was retained for machine learning model training.

i. Kruskal-Wallis Test Hypothesis

   a. Null Hypothesis (H0): numerical variable's distribution is constant across all severity categories.

   b. Alternative Hypothesis (H1): Significantly at least one group differs.

Kruskal-Wallis Test Formula:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i(R_i^2) - 3(N+1) \tag{3.4}$$

Where:

N= Total number of observations

n_i= number of observations in group i.

R_i= Sum of ranks for group i.

k= Number of groups (severity levels)

It can be concluded that the variable is essential for modeling and has significantly different distributions across severity levels if the p-value is less than 0.05.

The descriptive statistical analysis of each feature was performed and is presented feature-wise.

3.5.4 Encoding, Train-Test Split and Hyperparameter Tuning

Tree based models (Random Forest, XGBoost, and LightGBM) are inherently scale invariant. These models split nodes based on feature thresholds instead of distance calculations. Therefore, they do not strictly require scaling. Prior to model training, the raw dataset required systematic transformations to ensure compatibility with the selected machine learning algorithms. The Fatality Analysis Reporting System (FARS) dataset consists of a mix of categorical, ordinal, and continuous features. The dataset was split into training and test sets using an 80/20 stratified split on the target variable `HIGHEST_DRIVER_SEV` to preserve the class distribution. The split was performed with a fixed random seed 707 to ensure reproducibility, and no grouping by crash identifier was applied because the unit of analysis is the individual driver and the feature set includes driver-level as well as crash-level variables; multiple drivers involved in the same crash event are treated as separate observations. All categorical features were encoded using a training

set-based approach (label encoding or one-hot encoding, as appropriate), with the encoding mappings learned on the training data and applied to the test data without modification.

Hyperparameter tuning was carried out on the training set using repeated stratified 10-fold cross-validation with 5 repeats. The random seed for the cross-validation splits was set to 42. The search ranges and final selected hyperparameters for each model are summarized in Table 3.4.

TABLE 3.4: Hyperparameter Search Ranges and Selected Values

Model	**Hyperparameter**	**Search Range**	**Selected Value**
Random Forest	`n_estimators`	[100, 200, 300, 500]	250
	`max_depth`	[10, 20, 30, 35, 40]	35
	`min_samples_split`	[2, 3, 5, 10]	3
XGBoost	`n_estimators`	[100, 200, 300, 500]	200
	`learning_rate`	[0.01, 0.05, 0.1, 0.2]	0.05
	`max_depth`	[3, 5, 7, 9]	7
	`subsample`	[0.6, 0.8, 1.0]	0.6
LightGBM	`learning_rate`	[0.01, 0.05, 0.1, 0.13, 0.15, 0.17]	0.17
	`num_leaves`	[31, 50, 100, 200]	100
	`max_depth`	[3, 5, 7, 9]	5
	`subsample`	[0.6, 0.8, 1.0]	0.8

All models were evaluated on the held-out test set using the metrics described in Section 4.3, and no test data were used during hyperparameter tuning or feature selection.

The models were optimized using the hyperparameters. A Grid Search methodology was applied in conjunction with the cross-validation process. Grid Search exhaustively searches through a manually specified subset of the hyperparameter space e.g., maximum tree depth, learning rate, and number of estimators. By evaluating the cross-validated performance for every possible combination of these predefined parameters, the combination yielding the highest validation score was selected as the final, optimized model architecture prior to final testing.

3.5.5 Model Performance Metrics

The evaluation of predictive performance in this study addresses a five-class classification problem defined by the KABCO injury severity scale. Because the target variable contains five distinct outcomes rather than a binary setting, the foundational confusion matrix is interpreted through a class-wise one-versus-rest framework. For any specific severity class k evaluated against the remaining classes, the outcomes are defined as follows:

1. True Positives occur when the model correctly predicts that an instance belongs to severity class k denoted mathematically as TP_k.

2. True Negatives occur when the actual severity is not class k and the model correctly assigns the instance to any of the other four classes denoted as TN_k.

3. False Positives occur when the model incorrectly predicts class k for an instance that actually belongs to one of the other four classes denoted as FP_k.

4. False Negatives represent missed cases occurring when the actual severity is class k but the model incorrectly assigns it to a different class denoted as FN_k.

To assess model performance across these categories, several specific mathematical metrics are calculated for each severity class k.

1. Accuracy is the global ratio of total correct predictions across all classes divided by the total number of predictions made by the model.

$$\text{Accuracy} = \frac{\text{Total Correct Prediction}}{\text{Total Predictions}} \tag{3.5}$$

2. Precision for a specific class k is the ratio of True Positives to the sum of True Positives and False Positives. It measures the reliability of the model when it explicitly predicts that specific severity level.

$$\text{Precision}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k} \tag{3.6}$$

3. Recall for a specific class k is the ratio of True Positives to the sum of True Positives and False Negatives. Recall is the metric most directly affected by missed cases of a particular severity class, making it critical for identifying how often the model fails to detect instances of specific physical trauma.

$$\text{Recall}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k} \tag{3.7}$$

4. F1-score for a specific class k is the harmonic mean of precision and recall which balances the reliability of positive predictions against the prevention of missed cases.

$$\text{F1-Score}_k = 2 \times \frac{\text{Precision}_k \times \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \tag{3.8}$$

Summarizing these metrics in a multiclass environment requires distinct aggregation strategies.

1. Macro-averaging calculates the arithmetic mean of the metric across all five classes independently, treating each class equally regardless of its frequency in the dataset.

2. Micro-averaging pools the global totals of all True Positives, False Positives, and False Negatives across all classes before calculating the overall metric.

3. Weighted-averaging computes the metric for each class independently and calculates the final average by multiplying each result by the true support proportion of that class within the test dataset.

The selection of these averaging techniques is tied to the inherent class imbalance of the dataset. Because the Fatality Analysis Reporting System is a census of fatal crashes, the actual observed target distribution is structurally skewed toward fatalities and severe injuries, meaning property damage only and minor injuries naturally constitute minority classes within this sampling frame. To handle this distribution methodologically, the Synthetic Minority Oversampling Technique was applied strictly within the training folds to balance the classes during

the learning phase. No individual severity class was artificially prioritized through algorithmic class weights, targeted resampling of the test set, or manual decision threshold adjustments. The hold-out test set preserved the actual observed target distribution, and weighted-averaged metrics were subsequently utilized to provide a rigorous evaluation of model performance on operational data.

3.5.6 Training Data Preparation

To address the severe class imbalance inherent in the fatal-crash sampling frame, the Synthetic Minority Oversampling Technique was utilized. The oversampling algorithm was applied strictly and exclusively to the training data. At no point was the dataset oversampled prior to the train-test split. Furthermore, during model development, SMOTE was executed solely within the training folds of the cross-validation framework. The twenty percent hold-out test set remained entirely untouched by the oversampling procedure. This strict isolation ensures that all final model evaluation metrics reflect the true and unbalanced operational distribution of the actual testing data. The application of standard SMOTE to datasets containing categorical predictors requires specific methodological justification, as the algorithm interpolates linearly between data points and can theoretically generate invalid intermediate floating-point values for discrete categorical codes. In this study, the use of ordinary SMOTE is justified by its specific interaction with the machine learning algorithms selected for evaluation. All categorical variables were numerically encoded prior to the oversampling phase. While standard SMOTE generates continuous synthetic values along the feature space, the subsequent application of tree-based ensemble models including Random Forest, XGBoost, and LightGBM inherently mitigates the issue of fractional categories.

Unlike distance-based algorithms such as K-Nearest Neighbors or Support Vector Machines, tree-based classifiers do not rely on exact discrete matches or strict Euclidean distance metrics. Instead, they partition the feature space using continuous threshold inequalities. Consequently, an interpolated synthetic floating-point value resulting from SMOTE logically falls into the correct partitioned decision node, allowing the ensemble models to accurately learn the decision boundaries of

the minority classes without requiring exact discrete integers. While alternatives such as SMOTENC or algorithmic class weighting provide different mathematical approaches for categorical data, the integration of standard SMOTE with threshold-based tree algorithms provided a computationally robust methodology. The empirical metrics on the untouched test set demonstrate that this approach successfully improved minority class recall without corrupting the underlying logic of the categorical predictors.

3.6 Model Training

The model training phase was executed using the Python programming language utilizing the scikit-learn library for the Random Forest classifier and the native Python application programming interfaces for both XGBoost and LightGBM. To ensure complete reproducibility of the data splits and algorithmic initialization across all computational runs, a global random seed of 707 was established.

Addressing the mathematical optimization of these algorithms requires defining their specific objective functions for a five-class classification problem. The XGBoost model was configured utilizing the multiclass softmax objective function, while the LightGBM model was similarly optimized using its native multiclass objective. For the Random Forest algorithm, the objective was defined by its split criterion which was configured to minimize Gini impurity. Furthermore, because the structural class imbalance was methodologically resolved by applying the Synthetic Minority Oversampling Technique directly to the training folds, algorithmic class weights were intentionally set to uniform across all three models. This ensured that no redundant mathematical penalties were applied to the majority classes during the learning phase.

The final architectural hyperparameters for each ensemble model were established through rigorous cross-validation. The final Random Forest configuration consisted of 250 estimators with a maximum tree depth of 35 and a feature subsampling rate set to the square root of total features. The XGBoost model was optimized with a learning rate of 0.05, 200 boosting rounds, a maximum tree

depth of 7, a column subsampling rate by tree of 0.6, and an L2 regularization parameter of 1.0. The LightGBM configuration utilized a learning rate of 0.17, a maximum depth of 5, 100 number of leaves, and a sub sample rate of 0.8. To actively prevent overfitting during the iterative gradient boosting processes, early stopping was implemented for both XGBoost and LightGBM.

3.6.1 Random Forest

Random Forest is a voting based ensemble learning method that constructs a multiple decision trees during training. It operates on the principle of bootstrap aggregating called bagging, where each tree is trained on a random subset of the training data with replacement. The final prediction for an input x is determined by aggregating the predictions of all B individual trees. For a regression task, this is the average, and for classification, it is the majority vote:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} f_b(x) \tag{3.9}$$

Random Forest was selected for the high-dimensional tabular nature of this dataset because bagging inherently resists overfitting, handles non-linear feature interactions seamlessly, and is also highly robust to outliers.

3.6.2 XGBoost (Extreme Gradient Boosting)

XGBoost is also a tree-based model. It is an advanced, highly optimized implementation of gradient boosted decision trees. XGboost does not perform bagging, instead, boosting builds trees sequentially, where each new tree is designed to minimize the residual errors of the previously generated trees. XGBoost optimizes a regularized objective function L at step t:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{3.10}$$

where l is a differentiable convex loss function measuring the difference between the target y_i and the current prediction $\hat{y}_i^{(t-1)}$, and $\Omega(f_t)$ is a penalty term for model complexity. XGBoost is particularly suited for this dataset due to its internal

handling of missing values and its aggressive regularization (Ω), which prevents over-complexity and minimizes variance.

3.6.3 LightGBM (Light Gradient Boosting Machine)

LightGBM is also a gradient boosting framework that differs structurally from traditional boosting algorithms. Instead of growing trees level wise (depth first), LightGBM generates trees leaf wise by choosing the leaf with the maximum delta loss to expand. It uses Gradient based One Side Sampling (GOSS) to optimize speed. GOSS retains instances with large gradients while randomly sampling those with smaller gradients. LightGBM was chosen for its exclusive feature bundling and leaf wise growth strategy. This strategy makes it exceptionally memory efficient and fast. This is highly beneficial for broad (many features) and deep (hundreds of thousands of rows) datasets, allowing for rapid iterations during hyperparameter tuning.

3.7 Feature Importance

The Machine Learning (ML) models act like a "Black Box" therefore the model interpretation is very important. The suite of techniques which assigns a numerical score to the model's input features, based on each feature's contribution in predicting the target variable. In our study on vehicle crash analysis, feature importance serves as the primary driver of crash severity. The importance of a feature X_j can be formally defined as the decrease in a model's performance when X_j is removed or its information is perturbed:

$$I(X_j) = L(y, f(X)) - L(y, f(X^{\setminus j})) \tag{3.11}$$

Where:

L is the loss function (e.g., Log-Loss for severity classification).

f(X) is the model prediction with all features.

$f(X^{\setminus j})$ is the prediction excluding or permuting feature j.

3.7.1 Model-Specific Feature Importance (GINI)

Gini Feature Importance is primarily used in Tree-based models. This method measures the reduction in "entropy" also called "Gini Impurity" from each feature, during the splitting process. The Gini Impurity G for a splitting-node is calculated as:

$$G = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{3.12}$$

Where:

p_i is the probability of a crash belonging to severity class i.

The Gini importance considers the features as "purest" which contain only one class type and receive the highest importance score.

3.7.2 Model Agnostic Feature Importance (Permutation Feature Importance)

The fundamental principle of permutation feature importance is that, in order to evaluate a feature's importance, break the relationship between the feature and the target (y), and analyze the drop in model's performance. The magnitude of drop in performance is related to the feature importance.

The steps to determine the importance of a feature are as follows:

1. Baseline Score is the model's performance metrics (e.g., F1-Score or Accuracy) are calculated on the original test dataset D. Let this score be s.

2. Permutation for each feature $j \in \{1, \ldots, p\}$, column j in dataset D is randomly shuffled to create a corrupted version of the column, denoted as $\tilde{D}_{k,j}$.

   (a) Column j in dataset featuring $\tilde{D}_{k,j}$ is passed through the already-trained model to generate new predictions.

   (b) New Score: The performance metric $s_{k,j}$ is calculated based on the model's predictions on this corrupted data.

3. Importance Calculation: The importance of the feature is determined by evaluating the difference or ratio between the baseline score and the average corrupted score across multiple repetitions. The importance i_j is calculated as follows:

$$i_j = s - \frac{1}{K} \sum_{k=1}^{K} s_{k,j} \tag{3.13}$$

Where K is the number of repetitions executed to ensure statistical stability (usually ranging from 5 to 10).

3.7.3 SHAP (Shapley Additive Explanations)

To interpret the predictions generated by the ensemble algorithms, this study employs SHapley Additive exPlanations (SHAP). It is essential to establish the epistemological boundary of this technique: the application of SHAP does not definitively prove the physical causality of why a specific traffic collision escalated in severity. Instead, it provides a mathematically rigorous explanation of how the fitted machine learning model weighted and utilized the available input variables to arrive at its output. By deconstructing the model logic, this approach extracts learned structural hypotheses rather than dictating absolute physical laws.

All models in this study are based on decision tree architectures, and accordingly the TreeExplainer algorithm was used to compute SHAP values. The model output being explained is a scalar expected severity score, defined as:

$$\mathbb{E}[Y \mid \mathbf{x}] = \sum_{k=0}^{4} P(Y = k \mid \mathbf{x}) \cdot k \tag{3.14}$$

where k is the numeric KABCO injury class (0 = Property Damage Only, , 4 = Fatal). This transformation collapses the five class probability vector into a single continuous value on the 0-4 scale, which respects the ordinal nature of the injury outcome and directly quantifies the expected level of trauma. Because the output is a single scalar, there is no need for per class decomposition or aggregation of

multiclass SHAP values; every SHAP value represents the contribution of a feature to a shift in expected severity relative to the baseline.

The background dataset, which defines the baseline expected value of the model output, consisted of a random sample of 100 observations drawn from the hold out test set. The same background sample was used for all three models to preserve comparability. For the global importance and beeswarm analyses, SHAP values were computed on a larger random sample of 500 test set observations, all distinct from the background sample. This setup ensures that the explanations reflect the model's behavior on unseen data, and that the baseline represents the average expected severity over a representative subset of that data.

When the research required targeted diagnostic insights related to fatal outcomes, the same scalar expected severity was interpreted with attention to the upper region of the scale: a high expected severity (close to 4) indicates that the model assigned a high probability to the fatal class. No isolation of a specific class index was necessary, because the ordinal score already encodes the severity hierarchy. The global feature importance ranking was obtained by averaging the absolute SHAP values over all explained observations, yielding a single importance metric per feature on the expected severity scale.

3.8 Summary

Several programming languages and data handling tools are available, usually C++ or Microsoft's .NET are used where intense mathematical calculations are required without changing the algorithm or underlying logic. Java is also a choice, but these language require extended time frame for production, moreover, these are limited in their applicability in scripting environments. The output, regardless of language chosen, remains the same and in the same format. Python was selected because of its wider use in industry. Similarly, though SQL databases have been there before Pandas, but SQL is not suitable for on the fly manipulation, matrix algebra and other operations. Because of its close compatibility with Python, Pandas was selected. The data source was chosen to be open source and public

domain to firstly follow the cyber ethics, and secondly to allow future researchers easily and readily access it. As with all raw data, FARS data set was not in a format ready to be used in machine learning, so feature engineering of data processing, sub setting, feature selection and imputation of unknown (missing) values was performed. Variable significance was tested before model training after which, three models, with various hyperparameters were trained namely, Random Forest, XGBoost and LightGBM. Out of these three, Random Forest outperformed. Feature Significance was again tested against the predicted values. The SHAPely method of eXplainable AI (XAI) was applied to understand the underlying mechanism of these models, which answered the "what", "why" and "how" of these models.

Chapter 4

Results and Analysis

4.1 Background

To achieve the research objective and to answer the research questions, the data was first analyzed under Exploratory Data Analysis (EDA), which shed light on each of the features. The EDA presented the distribution, skewness and Gini score for each feature, unknown (missing) values were identified, and based on each feature's statistics, the missing values were imputed as explained earlier, the results of EDA for selected features are presented for reference. Importantly, useful insights were gained from the EDA which are presented in later sections. Subsequently, after imputation, descriptive statistical analysis was performed for each of these features. The models were trained with Random Forest, XGBoost and LightGBM, and tuned using the hyperparameters. It was found that Random Forest performed best among the triad, followed by the LightGBM and XGBoost. The model performance scores were also recorded. After training the models, the significance testing was performed once more by generating the target column using predictions alone to verify the sustained significance of each feature against the model. Furthermore, the models were run to identify the feature importance, and SHAP was applied to gain a deeper insight into the model's internal working mechanism and to explain these models a comparative analysis of SHAP XAI was

also performed. The results of Model XAI and Comparative XAI are presented in the relevant sections.

4.2 Data Preprocessing Results and SHAP Explanations

Various operations were performed on the observational dataset to prepare it for model training. These operations include handling missing values, assessing the dataset through descriptive statistical analysis after imputation, and performing feature significance testing. The preprocessing steps involving the handling of missing values, descriptive statistical analysis, and feature significance testing are presented in this section for the MINUTE parameter, while the comprehensive details for all other variables are provided in Appendix B.

4.2.1 Explanatory Data Analysis

Because this research relies entirely on an observational dataset rather than a controlled experimental design, the handling of incomplete records requires strict methodological transparency. The collected dataset contains missing information that is systematically recorded using explicit numeric codes (e.g., 99 for "unknown") defined in the FARS data dictionary, rather than being left blank. Exploratory analysis quantified the amount and pattern of these coded missing values across all retained variables. In total, 1,312 records had a missing code for the MINUTE variable, representing approximately 0.9% of the observations. Other retained variables exhibited missingness rates ranging from less than 0.1% to roughly 1.5%, with the highest concentrations occurring in fields where a measurement was not applicable to the crash type or was not reported due to jurisdictional differences. The overall pattern indicated that missingness was largely unsystematic with respect to the injury severity outcome; detailed pervariable missingness counts and proportions are reported in Appendix B.

To prepare the dataset for robust analytical modeling, statistical imputation procedures were adopted rather than simply omitting the incomplete entries. It must

be explicitly stated that imputation does not restore data integrity. Instead, it estimates missing information under specific statistical assumptions. Following a comprehensive exploratory statistical analysis of the trainingset distributions, targeted imputation strategies were systematically applied, with all imputation parameters learned solely from the training data and then applied to the test set. For categorical variables where a single category was overwhelmingly dominant, mode imputation was used under the assumption that the most common value provides a reasonable estimate and that the missing values are missing at random conditional on the observed covariates. For continuous variables with approximately symmetric distributions, mean substitution was applied under the same missingatrandom assumption. For variables with high entropy and no clear central tendency, such as MINUTE, a ratio imputation strategy was employed: the missing entries were randomly drawn from the empirical multinomial distribution of the observed categories, under the assumption that the missing cases follow the same distribution as the observed ones. This procedure preserves the distributional shape of the variable and avoids introducing artificial peaks. To demonstrate this methodology in practice, the detailed preprocessing and imputation process executed for the MINUTE parameter is presented below for reference.

4.2.1.1 Minute of Crash Analysis

The distribution of MINUTE temporal feature of the dataset before the imputation is presented in Table 4.1. The distribution shows us the health of the MINUTE data. There are 1632 missing values coded 99. This frequency makes up to 0.9% of the total data under MINUTE. This is the least frequency of any known or unknown data point. Therefore, the imputation of the missing values does not pose a risk of imbalance and distortion

The following table (Table 4.1) shows the distribution for MINUTE.

As explained in the table above, the distribution of the temporal feature MINUTE contains a minute frequency of missing values. The frequency is 1632, which makes it 0.9% missing. The distribution is visualized below (Figure 4.1).

The distribution is visualized as below:

TABLE 4.1: Minute of Crash Distribution

Min	Freq.	%	Min	Freq.	%	Min	Freq.	%	Min	Freq.	%
0	1896	1.05	15	1767	0.98	30	1592	0.88	45	1798	1
1	1917	1.06	16	1973	1.09	31	1906	1.06	46	1878	1.04
2	1883	1.04	17	2207	1.22	32	1907	1.06	47	2131	1.18
3	1898	1.05	18	1985	1.10	33	1893	1.05	48	1838	1.02
4	5300	2.94	19	7128	3.95	34	5394	2.99	49	7118	3.95
5	1839	1.02	20	1745	0.97	35	1885	1.05	50	1778	0.99
6	2059	1.14	21	1976	1.10	36	1944	1.08	51	1996	1.11
7	2291	1.27	22	1979	1.10	37	2178	1.21	52	1977	1.10
8	1857	1.03	23	1957	1.09	38	1862	1.03	53	1964	1.09
9	6532	3.62	24	5587	3.10	39	6717	3.72	54	5398	2.99
10	1728	0.96	25	1848	1.02	40	1763	0.98	55	1846	1.02
11	2130	1.18	26	1946	1.08	41	2128	1.18	56	1997	1.11
12	1990	1.10	27	2210	1.23	42	2050	1.14	57	2215	1.23
13	2030	1.13	28	1818	1.01	43	1960	1.09	58	1845	1.02
14	7413	4.11	29	10422	5.78	44	7496	4.16	59	10939	6.07
									99	1632	0.90

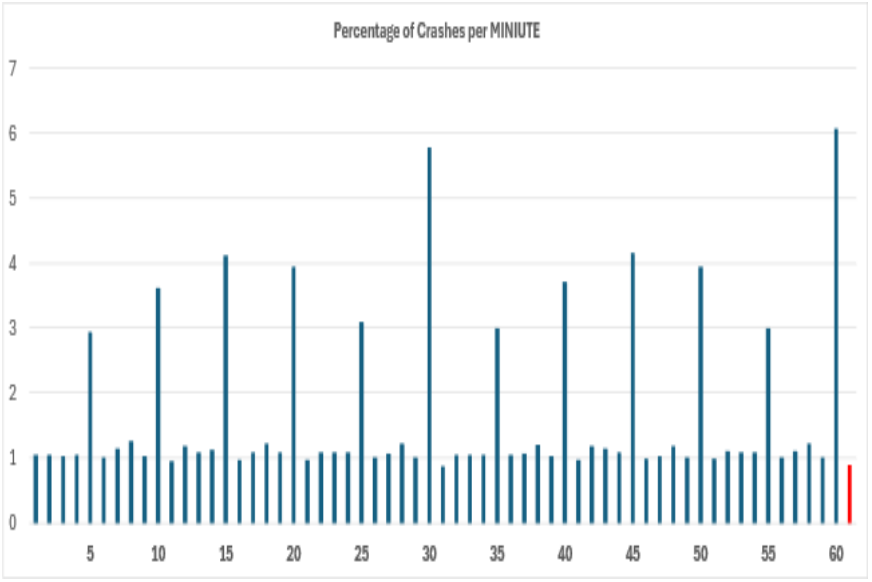

FIGURE 4.1: Distribution Chart for Minute of Crash

To impute the unknown values for MINUTE, statistical assessment was performed results of which are presented below (Table 4.2). The Mode Frequency of 0.0612 is a critical decision point. Also, the feature exhibited the high entropy of 5.6052 and Gini impurity 0.9741. If the missing values were filled with 0, it would be akin to assigning all crashes to the first minute that currently represents 6.12% of observed cases, thereby distorting the original form of distribution. With ratio based observed empirical proportions, assignment of missing values is performed probabilistically, according to the observed empirical proportions. This approach

preserves the marginal distribution of crash occurrence by minute in expectation while avoiding artificial concentration at any single temporal value. The imputation assessment metrics are given in the Table 4.2.

TABLE 4.2: Imputation Assessment: Minute of Crash

Metric	**Value**
Count	178704
Unique	60
Entropy	5.6052
Gini	0.9741
Mode	0.0
Mode_Freq	0.0612
Imputation Strategy	Ratio

The distribution after imputation validates the chosen imputation strategy, as the distribution remains undistorted. The minutes 5, 10, 15, 20 and 25 form first half of the distribution, where these spike up from the intermediate minutes, with local maxima at minute 15. The minute 30 experiences a high spike, while the rest of minutes return to lower frequencies than the minute 30, however, the minutes 35, 40, 45, 50 and 55 form the right half of distribution with spikes at these points with local maxima at the minute 45. The minute 30 has the highest bar in the range 1 to 59, while the last minute has the highest bar.

The distribution after imputation is presented below (Figure 4.2):

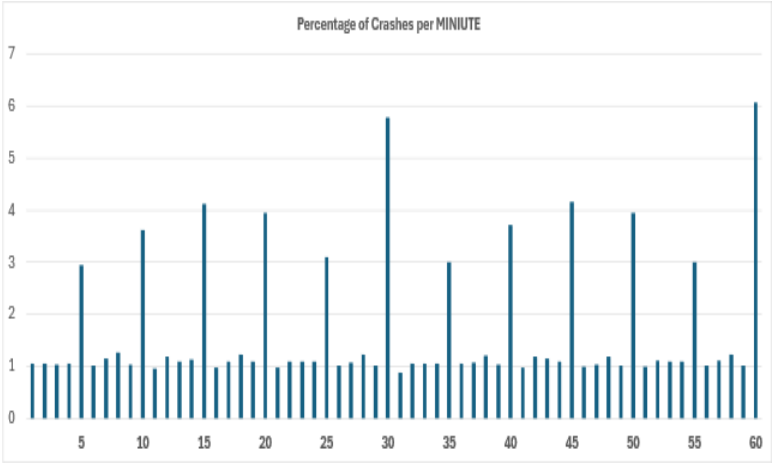

FIGURE 4.2: Post Imputation Distribution Chart

4.2.2 Descriptive Statistical Analysis of Dataset

For a deeper understanding of the underlying distribution and central tendencies of any large-scale observational dataset, the descriptive statistical analysis serves as the foundational step. It systematically evaluates individual variables, such as the exact MINUTE of an occurrence. Researchers identify temporal patterns, anomalies and data quality before advancing to complex modeling. Thus, producing comprehensive frequency tables is a critical initial procedure. The statistical analysis tables report the frequency, which quantifies the raw absolute count of instances recorded for each specific minute, alongside the total percentage, which represents this count as a fraction of the entire dataset, including any missing or unrecorded entries. Importantly, the researcher must isolate the valid percentage, which is unlike the total percentage. The valid percentage adjusts the computational baseline by excluding the missing values or corrupted data points, hence providing a mathematically accurate reflection of the variable's true distribution within the observable data. Descriptive statistical analysis is performed after imputing the missing or unknown values to ensure the proper imputation and assess the data quality. The descriptive statistical analysis of each feature was performed and is provided in Appendix-B. The descriptive statistical analysis of MINUTE is presented here for reference.

4.2.2.1 Descriptive Statistics of Minute of Crash

The FARS variable MINUTE records the minute of the crash using integer codes 0 through 59, with 99 designated for unknown. After ratio imputation of the 1,312 unknown records, all 142,624 observations contained a valid minute code in the 0-59 range. The post imputation distribution showed distinct concentrations at multiples of five minutes, consistent with digit preference in crash reporting. The highest frequency was observed at minute 0 (6.12%, representing the top of the hour), followed by minute 30 (5.83%). Other prominent peaks occurred at minutes 15 (4.14%) and 45 (4.20%), as well as at minutes 10, 20, 40, and 50 (each approximately 3.7-4.0%). Smaller peaks were present at minutes 5, 25, 35, and 55. In contrast, most non rounded minute values each accounted for roughly

1% of the observations. This pattern reflects temporal heaping, whereby crash times that were not known precisely were reported at convenient round intervals. The ratio imputation preserved this distributional structure, and the post imputation MINUTE variable therefore retained the characteristic rounding pattern of the original recorded data. Because all records possessed a valid minute after imputation, the total and valid percentages were identical for all categories. The descriptive summary confirmed that the imputed MINUTE variable remained representative of the original temporal reporting pattern and was suitable for inclusion in subsequent feature engineering and modelling stages. The complete frequency distribution is presented in Table 4.3.

Table 4.3: Minute of Crash Distribution

Minute	**Frequency**	**Total %**	**Valid %**	**Minute**	**Frequency**	**Total %**	**Valid %**
0	1908	1.06%	1.06%	30	1601	0.89%	0.89%
1	1937	1.07%	1.07%	31	1919	1.06%	1.06%
2	1891	1.05%	1.05%	32	1917	1.06%	1.06%
3	1919	1.06%	1.06%	33	1912	1.06%	1.06%
4	5342	2.96%	2.96%	34	5448	3.02%	3.02%
5	1856	1.03%	1.03%	35	1904	1.06%	1.06%
6	2079	1.15%	1.15%	36	1967	1.09%	1.09%
7	2318	1.29%	1.29%	37	2196	1.22%	1.22%
8	1883	1.04%	1.04%	38	1877	1.04%	1.04%
9	6600	3.66%	3.66%	39	6765	3.75%	3.75%
10	1743	0.97%	0.97%	40	1777	0.99%	0.99%
11	2154	1.19%	1.19%	41	2152	1.19%	1.19%
12	2004	1.11%	1.11%	42	2070	1.15%	1.15%
13	2048	1.14%	1.14%	43	1979	1.10%	1.10%
14	7465	4.14%	4.14%	44	7571	4.20%	4.20%
15	1785	0.99%	0.99%	45	1819	1.01%	1.01%
16	1994	1.11%	1.11%	46	1895	1.05%	1.05%
17	2233	1.24%	1.24%	47	2145	1.19%	1.19%
18	2011	1.12%	1.12%	48	1850	1.03%	1.03%
19	7195	3.99%	3.99%	49	7191	3.99%	3.99%
20	1762	0.98%	0.98%	50	1798	1.00%	1.00%
21	1994	1.11%	1.11%	51	2006	1.11%	1.11%
22	2001	1.11%	1.11%	52	1994	1.11%	1.11%
23	1973	1.09%	1.09%	53	1979	1.10%	1.10%
24	5650	3.13%	3.13%	54	5449	3.02%	3.02%
25	1865	1.03%	1.03%	55	1866	1.03%	1.03%
26	1964	1.09%	1.09%	56	2013	1.12%	1.12%
27	2227	1.23%	1.23%	57	2225	1.23%	1.23%
28	1835	1.02%	1.02%	58	1859	1.03%	1.03%
29	10516	5.83%	5.83%	59	11040	6.12%	6.12%

4.2.3 Feature Significance Testing

In research involving cross classified data, statistical hypothesis testing can help identify associations between predictor variables and the target outcome. In this study, two hypothesis testing methods were applied prior to modelling, selected according to the measurement scale and distributional properties of each variable. Pearson's chi square test of independence was used for categorical predictors to assess whether a statistically significant association existed with crash severity. For numerical variables, the Kruskal-Wallis H test was employed because it provides a non parametric comparison of distributions across the five severity groups without requiring the normality assumption of parametric ANOVA. However, with over 140,000 observations, these tests are extremely sensitive, and even trivial associations can yield p values well below conventional thresholds. Therefore, practical significance was evaluated through effect size measures reported alongside the p values. For each chi square test, Cramer's V was computed, with values below 0.10 interpreted as negligible regardless of statistical significance. For Kruskal-Wallis tests, epsilon squared () was calculated to indicate the proportion of variance in ranks explained by severity group membership, with values below 0.01 considered trivial. The results of these tests served only as a preliminary, exploratory guide to feature relevance and did not constitute a rigid selection criterion. Descriptive statistics were also computed for all features, and the detailed test outputs are provided in Appendix B.

4.2.3.1 Feature Significance of Minute of Crash

After ratio imputation of the 1,312 records originally coded as unknown (99), the MINUTE variable contained 142,624 complete observations with values in the range 0-59. A Pearson chi square test of independence was performed on this imputed dataset, yielding a statistic of 567.654 with 236 degrees of freedom and a p value below 0.001, which leads to rejection of the null hypothesis at conventional significance levels. However, with a sample of over 140,000 cases, the test is sensitive enough to detect even trivial departures from independence. To evaluate practical significance, Cramer's V was computed and found to be approximately

0.028, well below the usual threshold of 0.10 for a small effect. This indicates that the minute of crash explains a negligible proportion of the variance in driver injury severity. The likelihood ratio test gave a similar result ($p < 0.001$).

The output also included a linear by linear association statistic (6.200, $p = 0.0128$), which tests for a monotonic trend across ordered categories. This test is not appropriate for the MINUTE variable because minutes are cyclic rather than truly ordinal: minute 59 and minute 0 are adjacent in time but appear at opposite ends of a linear scale. Consequently, the linear by linear result is not interpreted. The observed peaks at multiples of five minutes correspond to well documented digit preference in crash reporting and do not represent behavioral patterns.

Given the negligible effect size, MINUTE is not considered a strong or practically meaningful predictor of injury severity. It was nonetheless retained for modeling to allow the ensemble methods to capture any subtle interactions with other features, but its standalone importance is minimal.

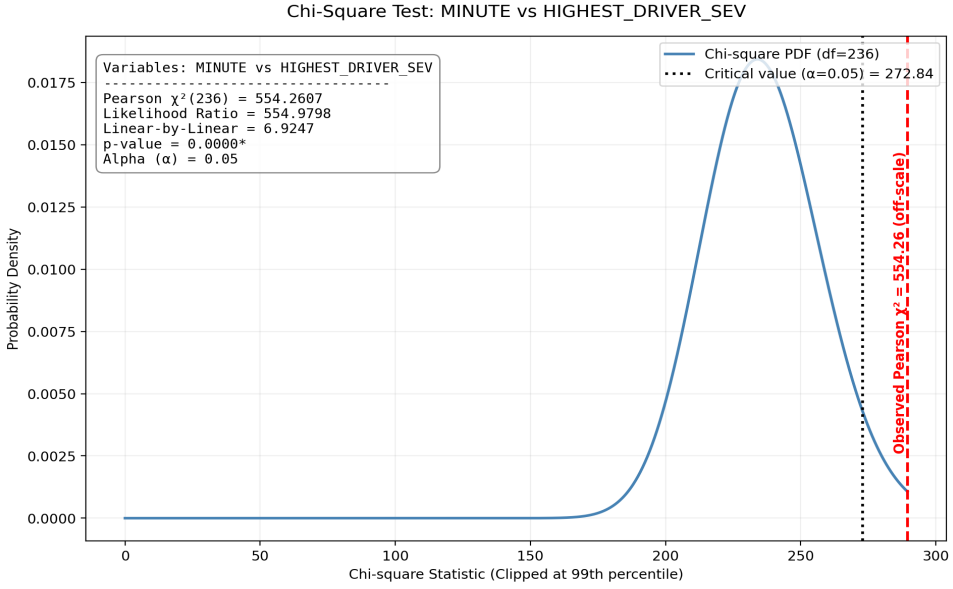

FIGURE 4.3: Chi-Square Test for MINUTE

4.2.4 SHAP for Random Forest

SHAP values were computed for the trained Random Forest model using a scalar expected severity score as the model output. The expected severity is defined as

$\mathbb{E}[Y \mid \mathbf{x}] = \sum_{k=0}^{4} P(Y = k \mid \mathbf{x}) \cdot k$, where k is the numeric KABCO injury class (0 = Property Damage Only, ..., 4 = Fatal).

Thus, each SHAP value represents the contribution of a feature to the shift in expected severity on the 0–4 scale, with positive values indicating an increase in the predicted severity level and negative values indicating a decrease. Because the output is a single continuous score, there is no decomposition into separate per-class explanations; all SHAP values are directly interpretable as effects on the expected severity. Global feature importance is obtained as the mean absolute SHAP value over all test set instances, also on the expected severity scale. No per-class averaging is involved, and the importance reflects the average magnitude of influence irrespective of direction.

The SHAP scores and the corresponding bar plot (Table 4.5 and Figure 4.5) present the global importance hierarchy. The length of each bar corresponds to the mean absolute SHAP value, indicating the average magnitude of contribution to the expected severity score. The feature Number of Persons (`N_PERSONS`) emerged as the top ranking predictor with a mean absolute SHAP value of 0.4015, followed by Pedestrian Involved (`PEDS`, binary) at 0.2309, while the count of pedestrians (`PEDESTRIAN_INVOLVED`) ranked sixth at 0.0780. However, because `N_PERSONS` and the pedestrian related features are mechanically related (a crash involving a pedestrian often has a low vehicle occupant count, and a crash with many occupants rarely involves a pedestrian), their SHAP values should not be interpreted as independent causal effects. The model partitions the shared predictive signal among these correlated variables, so their individual rankings reflect a joint association with severity rather than separable, independent influences. The combined importance of pedestrian related features, approximately 0.309, highlights the substantial role of vulnerable road user involvement, but this aggregate figure must be understood in the context of the interdependencies among these variables.

Vehicle damage and collision mechanics features also ranked prominently: Vehicle 1 Deformation (`VEH1_DEFORMED`) at 0.1271 (rank 3), Harmful Event (`HARM_EV`) at 0.0996 (rank 5), Driver 2 Impact Location (`DRIVER2_IMPACT1`) at 0.0734 (rank 7), Driver 1 Impact Location (`DRIVER1_IMPACT1`) at 0.0580 (rank 10), Vehicle

2 Deformation (VEH2_DEFORMED) at 0.0566 (rank 11), and Manner of Collision (MAN_COLL) at 0.0508 (rank 13). The prominence of deformation, impact location, and collision manner indicates that the model relies strongly on descriptors of both the crash configuration and its physical consequences.

Occupant protection and vehicle characteristics formed another important dimension. Driver 1 Restraint Use (DRIVER1_REST_USE) ranked fourth (0.1099), while Vehicle 1 Body Type (VEH1_BODY_TYP) and Vehicle 2 Body Type (VEH2_BODY_TYP) ranked eighth (0.0629) and twelfth (0.0519), respectively. The presence of an airbag in Vehicle 2 (VEH2_AIR_BAG) also appeared among the leading predictors (0.0373). Roadway and environmental variables, such as Road Functional Class and light conditions, ranked lower, but their collective contribution was non-negligible.

Throughout this section, numeric category codes have been replaced with descriptive labels to aid interpretation. For instance, "VEH1_BODY_TYP" is referred to as "Vehicle 1 Body Type," and "DRIVER1_REST_USE" as "Driver 1 Restraint Use." The full mapping of feature codes to descriptive names is provided in Appendix C, and the same labels were used when generating the SHAP figures.

The bar plot and Table 4.5 convey the magnitude of global influence but not the direction of effects. Directional information is provided by the beeswarm plot, which displays SHAP values on the expected severity scale, and by the waterfall plot for individual case interpretation. The beeswarm analysis, presented next, shows how high or low values of each feature push the expected severity upwards or downwards.

4.2.4.1 SHAP Explanations of Random Forest Model

The SHAP values used in this section were computed for the Random Forest model's expected severity score, defined as $\mathbb{E}[\text{severity}] = \sum_{k=0}^{4} P(Y = k) \cdot k$, which yields a continuous value between 0 and 4. The background dataset used for the explainer consisted of 100 randomly sampled observations from the test set (the same background was employed for all three models to preserve comparability). Over this background, the baseline expected severity was 3.05. Consequently, a

SHAP value of +0.2 for a feature means that the feature raised the expected severity by 0.2 units relative to the baseline. The global importance, beeswarm, and bar plots are all on this expected severity scale. For the individual case interpretation (waterfall), SHAP values were instead computed for the predicted probability of the fatal injury class (KABCO 4) to satisfy the requirement of a class-specific probability; the baseline here is the average probability of fatality in the background sample (0.31). All features are referred to by their descriptive labels, and the code-to-label mapping is supplied in Appendix A.

The beeswarm plot (Figure 4.6) arranges features by their global importance (mean absolute SHAP on expected severity). Each point represents one observation; the horizontal position indicates the SHAP contribution, while the color reflects the feature value. For truly numerical features (e.g., Driver Age, Number of Persons), the color gradient from blue (low) to red (high) is meaningful. For nominal categorical features (e.g., Vehicle 1 Body Type, Harmful Event), the color indicates distinct categories, and no ordinal or magnitude interpretation should be drawn from the color alone; the relevant category labels are described in the text where a specific effect is noted. The plot reveals that Number of Persons (`N_PERSONS`) and Pedestrian Involved (`PEDS`, binary) produce the largest spread of SHAP values. Higher counts of persons are predominantly associated with negative SHAP contributions (lower expected severity), while lower counts extend toward positive values, though the effect is asymmetric. Pedestrian involvement (`PEDS` = 1) tends to increase expected severity, but the magnitude varies greatly. It is important to reiterate that `N_PERSONS` and `PEDS` are mechanically linked, since a pedestrian crash often involves few vehicle occupants and their SHAP contributions should be viewed as a joint signal from the exposure dimension rather than independent causal effects. When interpreting the beeswarm, the combined pattern of these two features reflects the vulnerability of crashes involving pedestrians and low occupancy, not separate additive influences.

Several crash consequence variables show heterogeneous contributions. Vehicle 1 Deformation Extent (`VEH1_DEFORMED`) and Harmful Event (`HARM_EV`) exhibit both positive and negative effects; higher deformation categories are generally linked

to higher expected severity, but the relationship is not monotonic, as some lower deformation cases still receive positive SHAP values, possibly due to interactions with other crash features. Driver 1 Restraint Use (`DRIVER1_REST_USE`) shows a similar pattern: belted status (coded as a specific category) tends to push expected severity downward, while unbelted or unknown status is associated with upward contributions. These patterns align with the physical understanding that restraint systems mitigate injury, but the model also captures situations where other factors dominate.

Pre-crash variables such as Road Functional Class, Light Condition, and Road Surface Condition appear lower in the importance ranking but still show systematic patterns. For instance, unlit roads (a particular category of Light Condition) are linked to increased severity. These pre-crash features are directly relevant to the infrastructure risk prioritization use case, whereas the deformation and impact variables (post-crash) inform forensic structural analysis.

Driver Age exhibits a moderate effect: older drivers (red points) tend to appear on the positive SHAP side, consistent with greater physiological vulnerability, though the effect is small relative to the top predictors. Vehicle Model Year shows a slight trend where more recent model years are occasionally associated with higher severity contributions; however, the effect is inconsistent and may reflect the composition of the vehicle fleet in the FARS timeframe rather than a decline in vehicle safety. No causal claims about build quality are made; the association may be confounded by vehicle type, crash type, or other factors not fully isolated by the model.

The beeswarm plot demonstrates that the Random Forest's predictive structure is non-linear and context-dependent. The most influential features not only have large average effects but also exhibit wide dispersion, indicating that their impact varies substantially across different crash scenarios. This observation underscores why the mean absolute SHAP (importance) and the beeswarm distribution serve complementary roles.

To illustrate local explainability, a correctly classified case with a high predicted probability of fatality was selected. The waterfall plot (Figure 4.7) shows how each

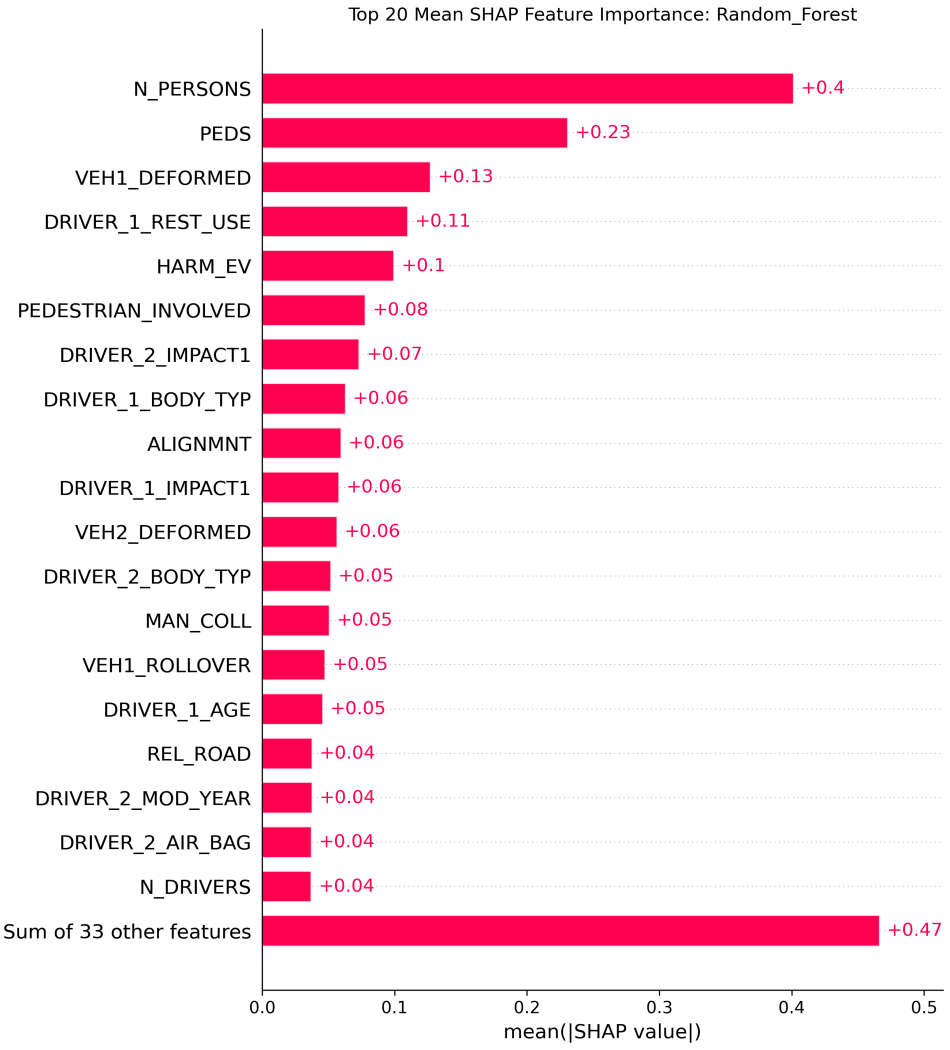

FIGURE 4.4: Bar Plot of SHAP Values from Random Forest Model

feature contributed to the model's output for the Fatal class (KABCO 4). The baseline probability of fatality was 0.31. The final predicted probability for this case was 0.89 (model output in log-odds transformed to probability). The features are displayed with descriptive labels. The largest upward push came from Vehicle 1 Body Type being 'Light Truck/SUV' (a category associated with higher severity in collisions with smaller vehicles), contributing +0.22 in log-odds. Harmful Event category 'Head-on collision' added +0.18, and Vehicle 2 Body Type also 'Light Truck/SUV' contributed +0.17. The fact that no pedestrian was involved (Pedestrian Involved = No) pushed the probability upward by +0.15, reflecting that in this occupant-only crash, other high-severity factors were dominant. Number of Persons being 3 provided a downward push of −0.18, likely because multiple

occupants can sometimes distribute crash forces. The impact location (Driver 1 Impact Area = Front) had a small downward effect (−0.05). The remaining 33 features collectively added a small upward increment of +0.04. The net effect moved the log-odds from the baseline to a high probability of fatality.

This case illustrates that the model reached a high severity prediction through the accumulation of several adverse crash characteristics, not a single dominant feature. It also exemplifies the distinction between global importance (where `N_PERSONS` ranked first) and local explanation (where vehicle body type and collision configuration were the primary upward contributors). The interpretation is strictly model-based and does not imply causation; it shows how the Random Forest combined the available information for this specific crash.

The same analytical structure is followed for XGBoost and LightGBM in the next sections, with identical output scales, background data, and descriptive labelling to permit consistent interpretation and cautious comparison of feature rankings.

Table 4.4: Top 20 SHAP scores for Random Forest Model

Rank	**Feature**	**Mean Absolute SHAP**
1	N_PERSONS	0.4014
2	PEDS	0.2309
3	VEH1_DEFORMED	0.1270
4	DRIVER_1_REST_USE	0.1098
5	HARM_EV	0.0996
6	PEDESTRIAN_INVOLVED	0.0780
7	DRIVER_2_IMPACT1	0.0733
8	VEH_1_BODY_TYP	0.0628
9	ALIGNMNT	0.0597
10	DRIVER_1_IMPACT1	0.0580
11	VEH2_DEFORMED	0.0565
12	VEH_2_BODY_TYP	0.0519
13	MAN_COLL	0.0507
14	VEH1_ROLLOVER	0.0474
15	DRIVER_1_AGE	0.0459
16	REL_ROAD	0.0377
17	VEH2_2_MOD_YEAR	0.0377
18	VEH_2_AIR_BAG	0.0372
19	N_DRIVERS	0.0371
20	DRIVER_2_AGE	0.0367

4.2.4.2 SHAP Explanation of Features

The SHAP beeswarm plot presents the distribution of feature contributions across the analyzed crash observations, as shown in Figure 4.6. Features are arranged vertically according to their global importance, while the horizontal axis represents the SHAP contribution to the model output. Points on the positive side increase the explained model output relative to its baseline, whereas points on the negative side decrease it. The color scale represents the numerical feature value, ranging from lower values in blue to higher values in red.

The Random Forest beeswarm plot showed that N_PERSONS and PEDS produced the greatest variation in model output. Higher values of both features were concentrated predominantly on the negative SHAP side, while lower values extended towards positive contributions, indicating strong but asymmetric relationships with model predictions. The prominence of human exposure and pedestrian related variables was consistent with previous studies that identified participant type and vulnerable road user involvement as important determinants of injury severity [47], [49]. PEDESTRIAN_INVOLVED also showed distinct separation between its binary states, although its effect was considered jointly with PEDS because both variables represented related dimensions of pedestrian exposure.

Collision mechanics and occupant protection variables showed more heterogeneous patterns. VEH1_DEFORMED, HARM_EV, DRIVER_1_REST_USE, impact location, vehicle body type, collision manner, and rollover status produced contributions on both sides of zero, indicating that their effects depended on the specific encoded categories and crash context rather than following simple linear relationships. Similar importance of collision type, vehicle characteristics, and occupant related factors had been reported in earlier SHAP based severity studies [46–48]. Road alignment showed a comparatively clearer directional pattern, while driver age exhibited moderate positive and negative effects with higher ages appearing more frequently on the positive side, consistent with previous findings on the contribution of age to severity prediction [46, 48, 50]. The contrasting patterns of

N_PERSONS and N_DRIVERS further indicated that the model treated total human exposure and driver involvement as distinct predictive dimensions rather than interchangeable measures.

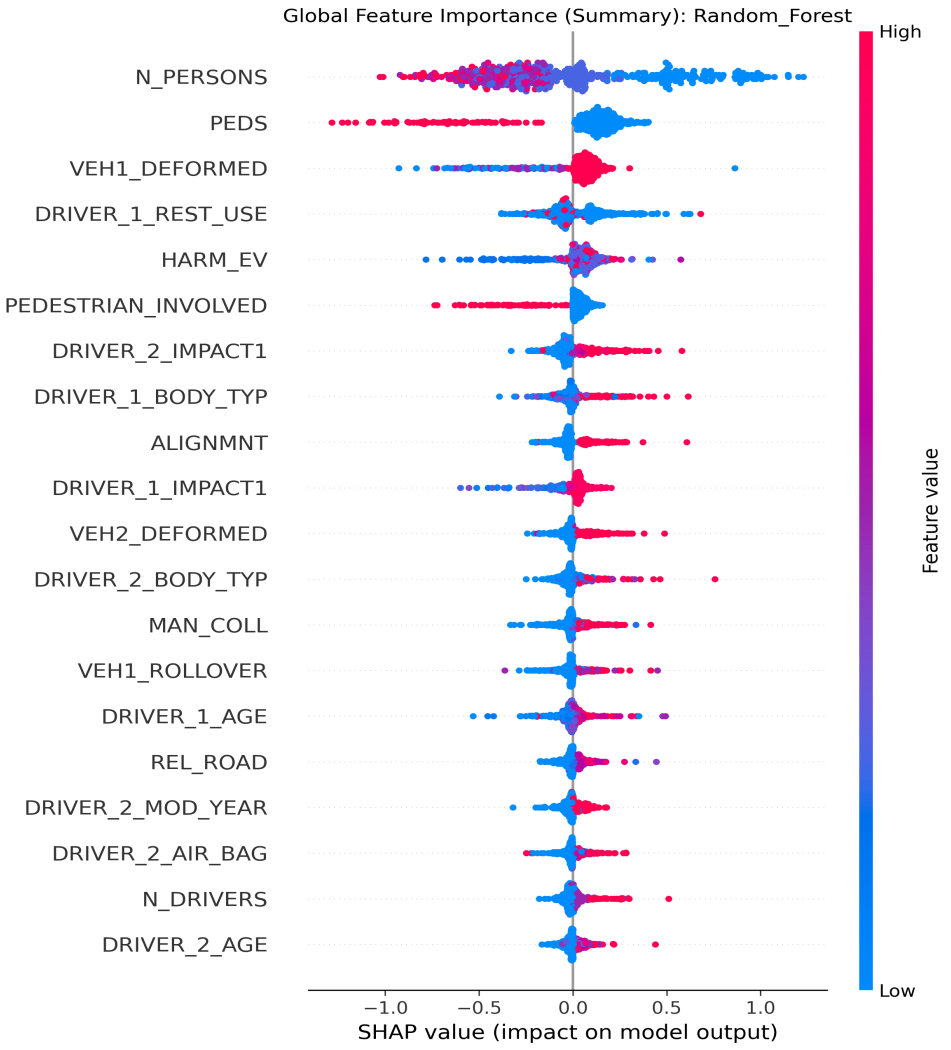

FIGURE 4.5: Beeswarm Plot of SHAP Scores from Random Forest Model

It was observed that the F01and F04 have "Heteroscedasticity of Influence" or non-linear impact because it showed blue trails on both sides of the vertical axis, also, a high confidence protective factor (red swarm on left) suggesting that with more people using the primary restraint, the severity will decrease, was seen. However, the heteroscedasticity explained the model, suggesting that when the number of personnel involved are low and lesser people are using seatbelts, the model uses other variables to determine the severity. Blue on the right, showed that when smaller numbers of people were involved in the crash with lower use of

seatbelts, they posed more risk of higher severity. Blue on left, showed that even when these features were absent, other factors kept the severity low. F01 and F04, thus, exhibited a distinct asymmetry in the summary distribution (Beeswam plot) in which high values are confined to the left (red are in negative space) acting as consistent deterrent, the low values (blue to the right and left) showed wide dispersion across both poles with a bi-model distribution.

Also the F03 (Vehicle Deformation) exhibited a strong high-risk contribution. The more the deformation, the more the energy transfer and the more the severity. The shape of the red dots (higher) values showed a consistent high-risk. The SHAP explains this feature behavior as concentrated positive cluster. Also, a low deformation showed lower severity.

In Pakistan too, the deformation, the use of seatbelts and number of people involved usually determines the severity. But it doesn't end there, this feature shows a protective trail with some red dots. This, with F02 the number of pedestrians, shows that even if a vehicle itself had low damage, it hit with a vulnerable person on road, and caused high severity or the driver attained some internal trauma (incapacitation). F02 i.e., the number of pedestrians, also exhibited a significant model intelligence, if a crowd of people (as common in Pakistan) are walking by, they increase the visibility of them being obstruction, the vehicle may slow down causing low severity crash. But with smaller number of pedestrians involved, the collision is usually fatal or serious.The features F07 onwards exhibited a serious trend of high value (density/frequency) leads to high severity. It can also be observed that The more the age of the driver the more the severity. However, an important finding is that with the time the build quality of the vehicles is decreasing, as the model year of the vehicle increases, meaning more recent years, the crash severity also increases.

The Beeswarm plot asserted that the Random Forest model exhibits a heterogeneous and nonlinear predictive structure. The most influential variables did not merely have larger average effects. They also showed substantially wider distributions of positive and negative contributions across individual crashes. The

model's strongest variation were associated with human exposure variables, particularly N_PERSONS and PEDS, followed by vehicle deformation, restraint use, harmful event characteristics, impact location, and vehicle body type. The figure also demonstrated why the SHAP importance table and Beeswarm plot serve different analytical purposes. The importance table identifies which variables matter the most on average, whereas the Beeswarm explains how their contributions vary across observations and whether different feature values tend to occupy different regions of the model response.

A case study of individual prediction is presented here. The Waterfall Plot of SHAP Scores (shown in Figure 4.7) exhibited that the model output was 3.05. The "Safety Story" of this crash is explained here. The waterfall plot presented the prediction pathway of an individual crash case and explained how the Random Forest model arrived at the final output of 4 from a baseline model output of 3.05. The case presented a cumulative risk pattern rather than a prediction dominated by a single variable. The strongest upward contribution came from the body type of the first vehicle represented by DRIVER_1_BODY_TYP, value being 80, which increased the model output by approximately 0.21. This was followed by the harmful event category represented by HARM_EV with value 12 and the body type of the second vehicle DRIVER_2_BODY_TYP value being 80, each contributing approximately 0.18. The absence of pedestrians or PEDS = 0 contributed a further 0.16 to the upward movement. These effects were reinforced by the manner of collision, airbag-status variables, impact-location characteristics, driver ages, restraint-use status, vehicle model year, speed limit, and vehicle travel speed. Individually, several of these contributions were moderate or small; collectively, however, they formed a consistent upward prediction pathway.

The upward movement was partially opposed by several downward contributions. The strongest counteracting influence was the presence of three persons in the crash or N_PERSONS with a value of 3, which reduced the model output by approximately 0.18. The first driver's impact-location category represented by DRIVER_1_IMPACT1 and with value of 1, provided a further downward contribution of approximately 0.05, while road functional class of 12 reduced the output

by approximately 0.02. The remaining 33 features collectively produced only a small downward contribution of approximately 0.04. Consequently, these counteracting effects were insufficient to offset the combined upward contributions of the vehicle, collision, and occupant-related characteristics, and the model output moved from the baseline of 3.05 to the final predicted class of 4.

The resulting safety story is therefore one of accumulated adverse crash characteristics. The model did not arrive at the final prediction because of one isolated factor. Rather, it concluded the prediction from the configuration of the involved vehicles, the harmful event, collision manner, impact characteristics, and occupant-related conditions collectively outweighed the comparatively limited downward influences. This case demonstrates how the Random Forest model combines multiple characteristics of a crash into an individual prediction and illustrates the distinction between global and local explainability. Although N_PERSONS was the most influential feature globally, in this individual case it acted as a downward contributor, whereas vehicle body type and harmful event characteristics became the principal upward contributors.

4.2.5 SHAP for XGBoost Model

The XGBoost SHAP analysis ranks the predictor variables according to their Mean Absolute SHAP values, representing the average magnitude of each feature's contribution to the model output across the analyzed observations. The SHAP was run for XGBoost model for 500 iterations. These yielded the most significant features, the bar plot, the beeswarm plot, waterfall plot and the force plot. These results are presented below.

4.2.5.1 SHAP Explanation of XGBoost Model

N_PERSONS emerged as the most influential feature, with a Mean Absolute SHAP value of 0.3671, followed by PEDS at 0.2873. A second group of influential predictors consisted of DRIVER_1_REST_USE with SHAP value 0.1090, VEH1_DEFORMED with 0.1057, N_DRIVERS with 0.1044, and HARM_EV with

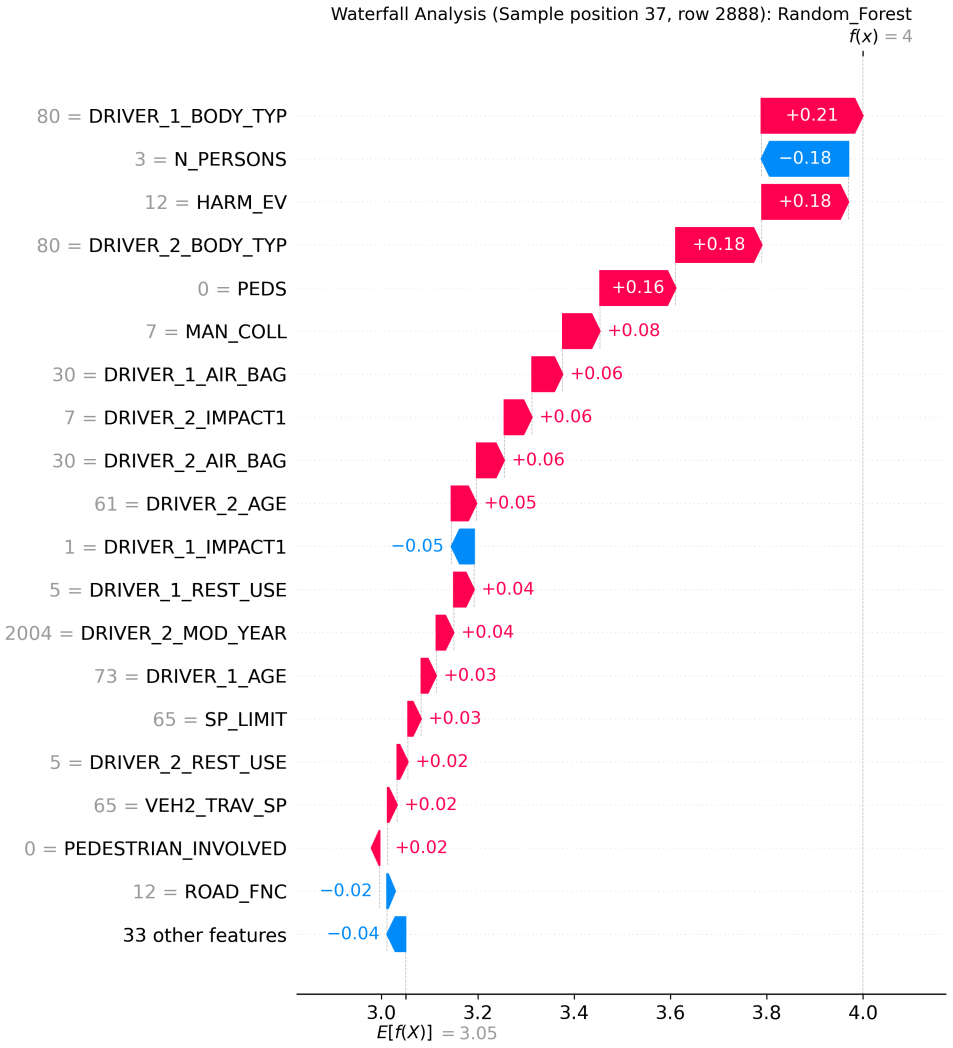

FIGURE 4.6: The Waterfall Plot of SHAP Scores for Random Forest Model

0.0996. The remaining features showed more gradual decline in importance. Beginning with DRIVER_2_IMPACT1 with SHAP value 0.0600 and extending to SP_LIMIT with SHAP value 0.0267 at rank 20. The XGBoost model places its greatest predictive emphasis on N_PERSONS and PEDS. N_PERSONS, with a SHAP importance of 0.3671, is the dominant predictor, but the difference between the first- and second ranked features is less pronounced than in the Random Forest model. PEDS reached 0.2873, approximately 78% of the importance magnitude of N_PERSONS. This indicated that pedestrian count occupies a particularly prominent position in the XGBoost predictive structure. The presence of N_DRIVERS at rank 5, with a SHAP importance of 0.1044, further indicated that the model

relied substantially on crash level human-exposure and participation characteristics. Taken together, N_PERSONS, PEDS, and N_DRIVERS showed that the composition of persons and road users involved in the crash forms a major predictive dimension of the XGBoost model. DRIVER_1_REST_USE is the third-ranked predictor, with a Mean Absolute SHAP value of 0.1090. Its high ranking indicates that restraint use information contributes substantially to the XGBoost prediction structure and is more influential globally than any individual impact location, driver's age, roadway, or speed related variable. DRIVER_1_AIR_BAG also enters the Top 20 at rank 14 with a SHAP importance of 0.0397. The presence of both restraint use and airbag related variables indicates that occupant-protection characteristics provide meaningful predictive information, although restraint use has substantially greater global influence. Collision-related characteristics constitute a major portion of the XGBoost Top-20 predictors.

VEH1_DEFORMED ranked fourth with a SHAP importance of 0.1057, while VEH2_DEFORMED ranks twelfth at 0.0444. HARM_EV ranks sixth at 0.0996, confirming the importance of the harmful event classification in the model's predictive structure. Impact location variables for both vehicle positions are also prominent. DRIVER_2_IMPACT1 ranks seventh and DRIVER_1_IMPACT1 ranks ninth, with nearly identical importance magnitudes of approximately 0.0600 and 0.0593. VEH1_ROLLOVER ranks thirteenth, while MAN_COLL ranked sixteenth. The repeated presence of deformation, impact location, harmful event, rollover, and manner of collision variables indicated that XGBoost relied on a broad representation of collision configuration and physical crash characteristics, rather than depending on a single crash mechanics variable. DRIVER_1_BODY_TYP ranks eighth with a SHAP importance of 0.0600, while DRIVER_2_BODY_TYP appeared at rank 19 with a lower importance of 0.0297. This difference indicated that the first vehicle's body-type information has substantially greater global influence within the XGBoost model than the corresponding second vehicle's variable. Driver's age showed a notably balanced pattern. DRIVER_2_AGE ranked tenth at 0.0485, while DRIVER_1_AGE ranked eleventh at 0.0483. The near identical importance magnitudes of the two age variables suggest that driver age contributes consistently across both driver positions. Roadway context is represented

by REL_ROAD at rank 15 and ALIGNMNT at rank 18. Their presence indicates that roadway position and alignment contribute to XGBoost predictions, although their global influence is lower than that of the leading human exposure, occupant protection, and collision characteristic variables. Speed related information is represented by VEH1_TRAV_SP at rank 17 and SP_LIMIT at rank 20. The inclusion of both actual travel speed and posted speed limit suggests that XGBoost extracts predictive information from both vehicle operating conditions and the roadway speed environment. The SHAPely values for XGBoost are presented in Table 4.6.

Table 4.5: Top 20 SHAP scores for XGBoost Model

Rank	**Feature**	**Mean Absolute SHAP**
1	N_PERSONS	0.3670
2	PEDS	0.2873
3	DRIVER_1_REST_USE	0.1089
4	VEH1_DEFORMED	0.1056
5	N_DRIVERS	0.1044
6	HARM_EV	0.0996
7	DRIVER_2_IMPACT1	0.0599
8	DRIVER_1_BODY_TYP	0.0599
9	DRIVER_1_IMPACT1	0.0592
10	DRIVER_2_AGE	0.0485
11	DRIVER_1_AGE	0.0483
12	VEH2_DEFORMED	0.0443
13	VEH1_ROLLOVER	0.0418
14	DRIVER_1_AIR_BAG	0.0396
15	REL_ROAD	0.0389
16	MAN_COLL	0.0361
17	VEH1_TRAV_SP	0.0307
18	ALIGNMNT	0.0297
19	DRIVER_2_BODY_TYP	0.0297
20	SP_LIMIT	0.0266

The bar plot of most significant features is presented in Figure 4.8. The SHAP explanations in the table above are represented in a bar plot. Therefore the explanation remains the same. The N_PERSONS is the most influential feature.

Followed by the number of pedestrians and deformation of the vehicle. Similarly, the use of primary restraint emerged as the fourth most important feature. The crash statistics of HARMFUL_EV also appeared at a top rank of five along with DRIVE_2_IMPACT1, DRIVER_1_IMPACT1, MAN_COLL and VEH1_ROLLOVER at ranks 7, 10, 13 and 14 respectively. Other factors contribute at varying but smaller ranks.

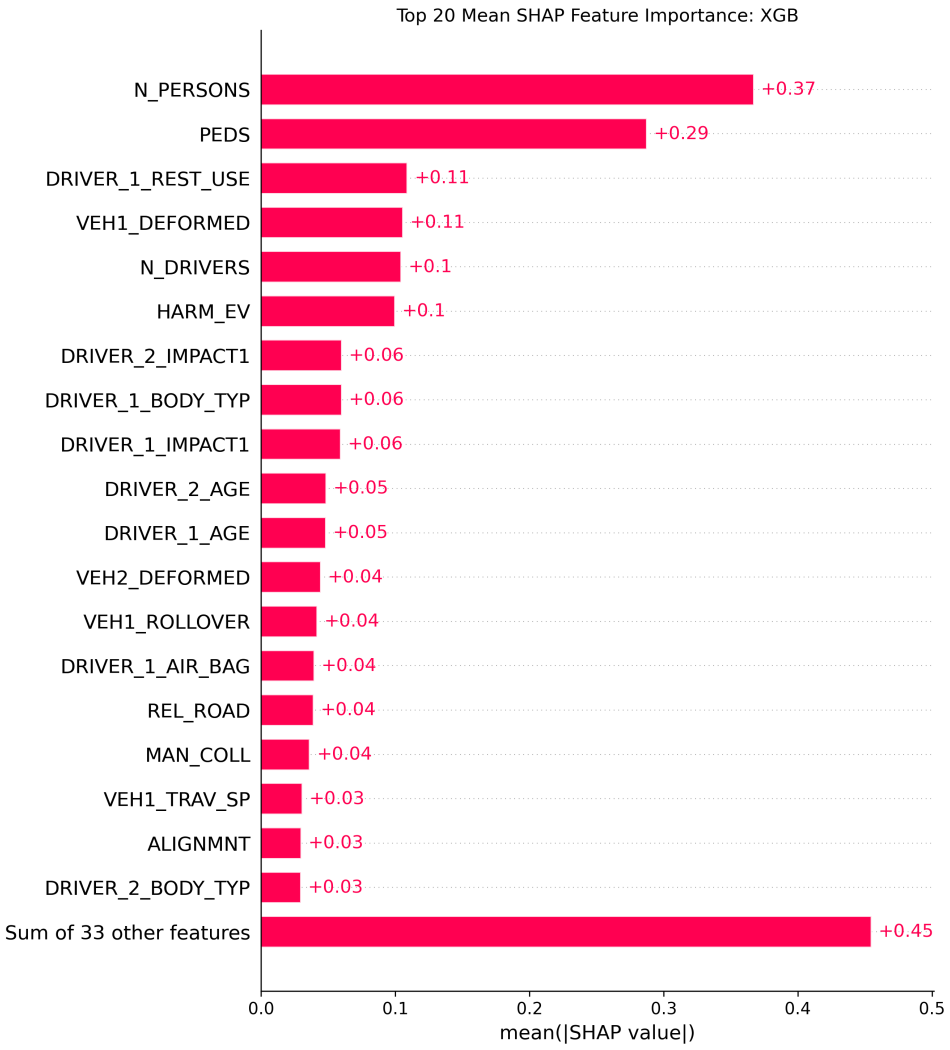

FIGURE 4.7: Bar Plot of SHAP Explanation of XGBoost

The XGBoost SHAP analysis identified N_PERSONS as the most influential feature with a mean absolute SHAP value of 0.3671, followed by PEDS at 0.2873. Their prominence, together with N_DRIVERS at 0.1044, indicated that human exposure and road user composition formed the dominant predictive dimension of the model. This finding was consistent with previous studies that identified

participant type, pedestrian involvement, and vulnerable road user characteristics as important determinants of crash severity [47,49]. Occupant protection also contributed substantially, as DRIVER_1_REST_USE ranked third at 0.1090, while DRIVER_1_AIR_BAG also appeared among the top 20 features. The importance of these variables agreed with previous findings concerning safety equipment and occupant related characteristics [49].

Collision mechanics formed another major component of the XGBoost predictive structure. VEH1_DEFORMED, HARM_EV, impact location variables for both vehicles, VEH2_DEFORMED, VEH1_ROLLOVER, and MAN_COLL collectively indicated that physical damage and collision configuration provided substantial predictive information. This finding corresponded with earlier studies that identified collision type, vehicle type, and vehicle condition among leading severity predictors [46–48]. Vehicle body type and driver age for both vehicle positions also appeared among the leading features, consistent with previous findings regarding vehicle characteristics and driver age [46, 48, 50]. Roadway position, alignment, travel speed, and speed limit contributed at lower levels, suggesting that XGBoost primarily relied on human exposure, occupant protection, and collision mechanics, while roadway and operating conditions provided supplementary predictive information.

4.2.5.2 SHAP Explanation of Features

For each feature, the Beeswarm plot shown in Figure 4.9 elucidates the effect of its either being "Protective Factor" or "Risk Factor". The color gradient from blue to red shows the value (lower to higher) and direction of right or left shows the direction of risk. The XGBoost Beeswarm plot indicated that N_PERSONS and PEDS had the widest distributions of SHAP values and therefore produced the greatest variation in model output. Higher values of N_PERSONS and PEDS were concentrated predominantly on the negative side, whereas lower values extended towards the positive side. N_DRIVERS showed the opposite tendency, as higher values contributed mainly towards positive model output. DRIVER_1_REST_USE, VEH1_DEFORMED, and HARM_EV also produced substantial variation across

crash observations, although their mixed distributions indicated heterogeneous effects among their encoded categories. The remaining features produced narrower but distinct contribution patterns. Higher coded values of DRIVER_2_IMPACT1, vehicle body type, VEH2_DEFORMED, VEH1_ROLLOVER, REL _ ROAD, MAN _ COLL, and ALIGNMNT were concentrated mainly towards positive SHAP values, while lower coded values generally occurred near zero or towards the negative side. Driver age variables showed comparatively moderate contributions, while travel speed and speed limit produced relatively compact distributions around zero with occasional larger individual effects. Overall, the plot indicated that XGBoost predictions had been driven principally by human exposure characteristics, followed by occupant protection, collision mechanics, vehicle characteristics, and roadway conditions, with substantial variation in feature influence across individual crash observations.

The XGBoost beeswarm plot illustrated the magnitude and direction of feature contributions across individual crash observations. Positive and negative SHAP values represented upward and downward shifts in model output, while the colour gradient represented lower and higher feature values. N_PERSONS and PEDS showed the widest SHAP distributions, with higher values concentrated mainly on the negative side and lower values extending towards positive contributions. In contrast, higher values of N_DRIVERS were associated predominantly with positive contributions. The strong influence of human exposure and vulnerable road users agreed with earlier studies that identified participant type and pedestrian related collision patterns as important determinants of severity [47, 49].

DRIVER_1_REST_USE, VEH1_DEFORMED, and HARM_EV showed substantial but heterogeneous contributions, reflecting the categorical and nonlinear relationships captured by XGBoost. Impact location, vehicle body type, deformation, rollover, collision manner, roadway position, and alignment also showed distinct directional patterns, supporting earlier findings regarding the importance of collision type, vehicle characteristics, and roadway conditions [46–48]. Driver age produced moderate contributions, while travel speed and speed limit remained comparatively concentrated around zero. Overall, the plot indicated that human

exposure had produced the strongest variation in predictions, followed by occupant protection, collision mechanics, vehicle characteristics, and roadway conditions.

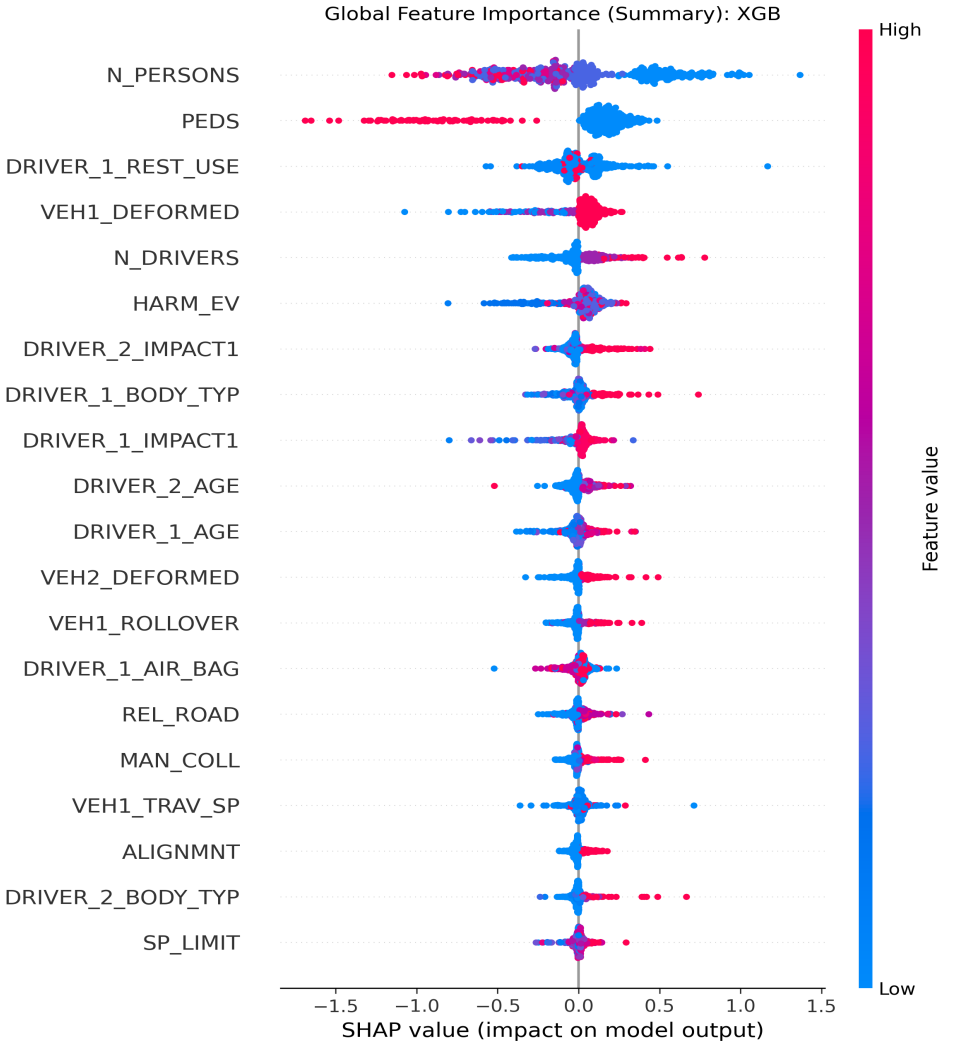

FIGURE 4.8: Beeswarm Plot for XGBoost

For a case study, the individual prediction logic is presented here to understand the internal decision making. A waterfall plot was plotted (shown in Figure 4.10) which is presented and discussed below. The XGBoost waterfall plot explained how the selected crash case moved from the baseline model output of 3.05 to the final output of 4. The strongest upward contributions came from the absence of pedestrians, which added 0.24, and the body type of the first vehicle, which added 0.21. Further upward contributions came from the manner of collision and the body type of the second vehicle, each adding 0.09, followed by restraint use, vehicle deformation, and the age of the second driver, each adding approximately

0.06. The harmful event, hit and run status, driver age, airbag status, travel speed, number of drivers, speed limit, and other smaller contributions further supported the upward movement.

The principal downward contribution came from the presence of three persons in the crash, which reduced the model output by 0.09. The impact location of the first vehicle reduced the output by 0.03, while lighting conditions contributed a further reduction of 0.02. These downward contributions were substantially outweighed by the accumulated upward contributions. The case therefore indicated that XGBoost reached the final output through the combined influence of pedestrian absence, vehicle configuration, collision manner, restraint status, vehicle deformation, harmful event, driver characteristics, and operating conditions. Compared with the global feature ranking, the case also demonstrated the local nature of SHAP explanations, since the globally important N_PERSONS reduced the model output in this observation while other features became the dominant upward contributors.

The XGBoost case study showed that the final model output of 4 resulted from the cumulative influence of multiple crash characteristics rather than a single dominant factor. The strongest upward contributions came from pedestrian absence and vehicle body type, supported by collision manner, restraint use, vehicle deformation, harmful event, driver characteristics, and operating conditions, while the number of persons, impact location, and lighting conditions provided limited downward contributions. The predominance of upward contributions moved the prediction from the baseline model output of 3.05 to the final output of 4, demonstrating how XGBoost combined human exposure, vehicle, collision, occupant, and operational characteristics to explain the individual prediction.

4.2.6 SHAP for LightGBM

The SHAP was run for XGBoost model for 500 iterations. These yielded the most significant features, the bar plot, the beeswarm plot, waterfall plot and the force plot. These results answer Research Questions 1 and 2. These results are presented below.

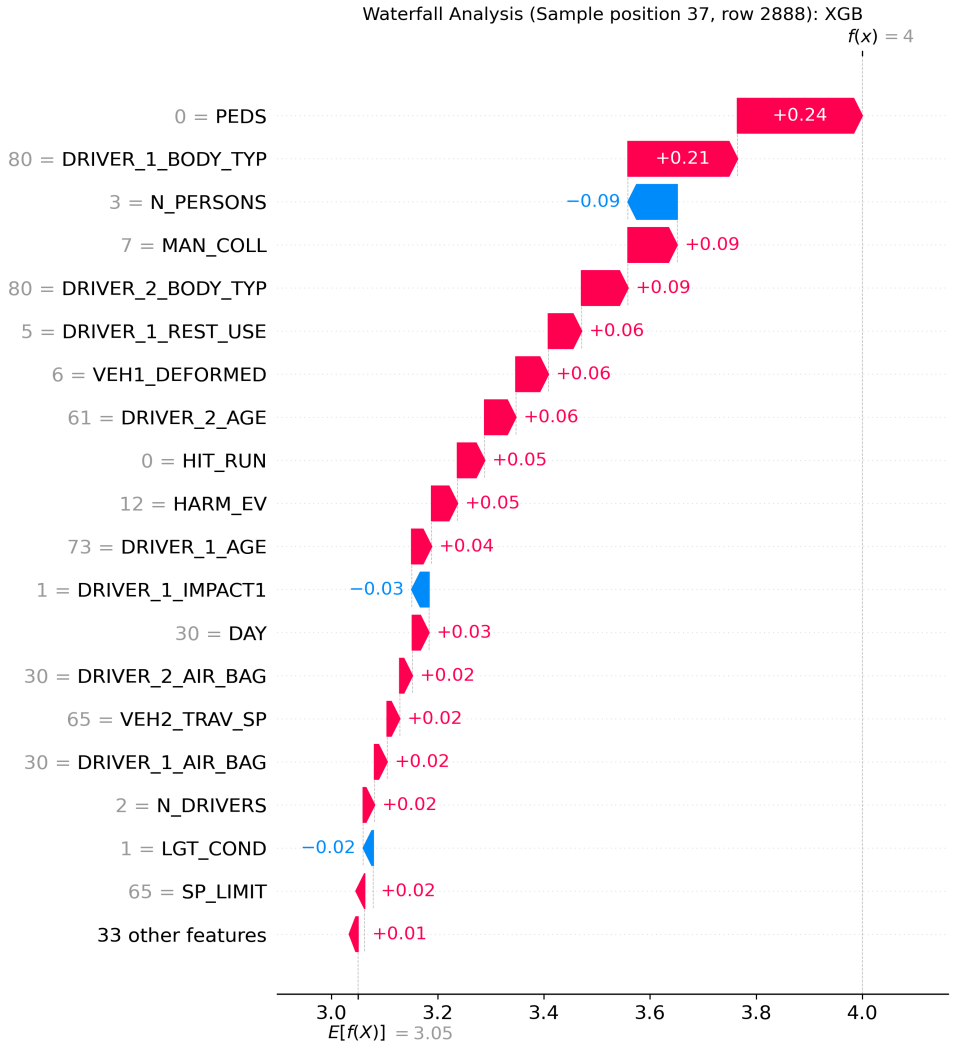

FIGURE 4.9: Waterfall Plot for SHAP Explanation for XGBoost

4.2.6.1 SHAP Explanation of LightGBM Model

The SHAP scores and Bar Plot show the most significant features which contribute to the crash severity predictions. The LightGBM SHAP analysis identified N_PERSONS as the most influential predictor with a Mean Absolute SHAP value of 0.4049, followed by PEDS at 0.2199. Collision and occupant related characteristics formed the next major group, including HARM_EV at 0.1218, DRIVER_1_REST_USE at 0.1136, N_DRIVERS at 0.0975, and VEH1 _DEFORMED at 0.0945. Driver characteristics also received substantial importance, particularly DRIVER_2_AGE, which ranked seventh with a value of 0.0876, followed by vehicle body type and impact location variables. The ranking therefore indicated that LightGBM relied principally on human exposure, collision circumstances,

occupant protection, vehicle deformation, and driver characteristics. The remaining features represented additional dimensions of collision mechanics, vehicle characteristics, roadway context, and operating conditions. VEH2_DEFORMED, VEH1_ROLLOVER, airbag status, travel speed, collision manner, speed limit, and relative road position all contributed to the predictive structure, although with lower average magnitudes than the leading variables. The distribution of importance across these feature groups indicated that LightGBM combined information from multiple components of the crash system, while the dominant influence remained concentrated in the number and type of road users involved, harmful event characteristics, occupant protection, and physical collision conditions. The SHAPely values for LightGBM are presented in Table 4.7.

Table 4.6: Top 20 SHAP scores for LightGBM Model

Rank	**Feature**	**Mean Absolute SHAP**
1	N_PERSONS	0.4048
2	PEDS	0.2199
3	HARM_EV	0.1217
4	DRIVER_1_REST_USE	0.1135
5	N_DRIVERS	0.0974
6	VEH1_DEFORMED	0.0944
7	DRIVER_2_AGE	0.0876
8	DRIVER_1_BODY_TYP	0.0693
9	DRIVER_2_IMPACT1	0.0647
10	DRIVER_1_IMPACT1	0.0590
11	VEH2_DEFORMED	0.0564
12	DRIVER_1_AGE	0.0528
13	VEH1_ROLLOVER	0.0471
14	DRIVER_1_AIR_BAG	0.0416
15	VEH1_TRAV_SP	0.0384
16	MAN_COLL	0.0379
17	SP_LIMIT	0.0352
18	REL_ROAD	0.0350
19	DRIVER_2_BODY_TYP	0.0339
20	HIT_RUN	0.0295

The LightGBM SHAP bar chart ranked the features according to their mean

absolute SHAP values and therefore represented the average magnitude of their influence rather than the direction of their effects. N_PERSONS showed the highest individual importance at approximately 0.40, followed by PEDS at 0.22. A second group of influential features included HARM_EV, DRIVER_1_REST_USE, N_DRIVERS, VEH1_DEFORMED, and DRIVER_2_AGE, with importance values ranging approximately from 0.09 to 0.12. The remaining leading features contributed progressively smaller magnitudes and represented vehicle body type, impact location, vehicle deformation, driver age, rollover, airbag status, travel speed, collision manner, speed limit, and relative road position.The combined SHAP importance of the remaining 33 features reached approximately 0.54, which exceeded the importance of any individual predictor. However, this aggregated value represented the cumulative contribution of numerous variables and should not have been interpreted as a single dominant feature. Overall, the chart indicated that LightGBM had concentrated substantial predictive influence in a limited group of leading variables, particularly human exposure, pedestrian involvement, harmful event, occupant protection, and collision characteristics, while the broader feature set had collectively provided considerable supplementary predictive information. The bar chart is presented in Figure 4.11.

The LightGBM SHAP analysis identified N_PERSONS as the most influential predictor with a mean absolute SHAP value of 0.4049, followed by PEDS at 0.2199. Their dominance, together with N_DRIVERS at 0.0975, indicated that human exposure and road user composition formed the principal predictive dimension of the model. This finding was consistent with previous studies that identified participant type, pedestrian involvement, and vulnerable road users as important determinants of injury severity [47,49]. Collision and occupant related characteristics formed the next influential group, including HARM_EV at 0.1218, DRIVER_1_REST_USE at 0.1136, and VEH1_DEFORMED at 0.0945. Similar importance of collision characteristics, vehicle conditions, and occupant related factors had been reported in previous SHAP based severity studies [46–48]. Driver and vehicle characteristics also contributed substantially, particularly DRIV- ER_2_AGE, vehicle body type, impact location, second vehicle deformation, rollover status, and airbag status. The importance of driver age and vehicle characteristics agreed with previous

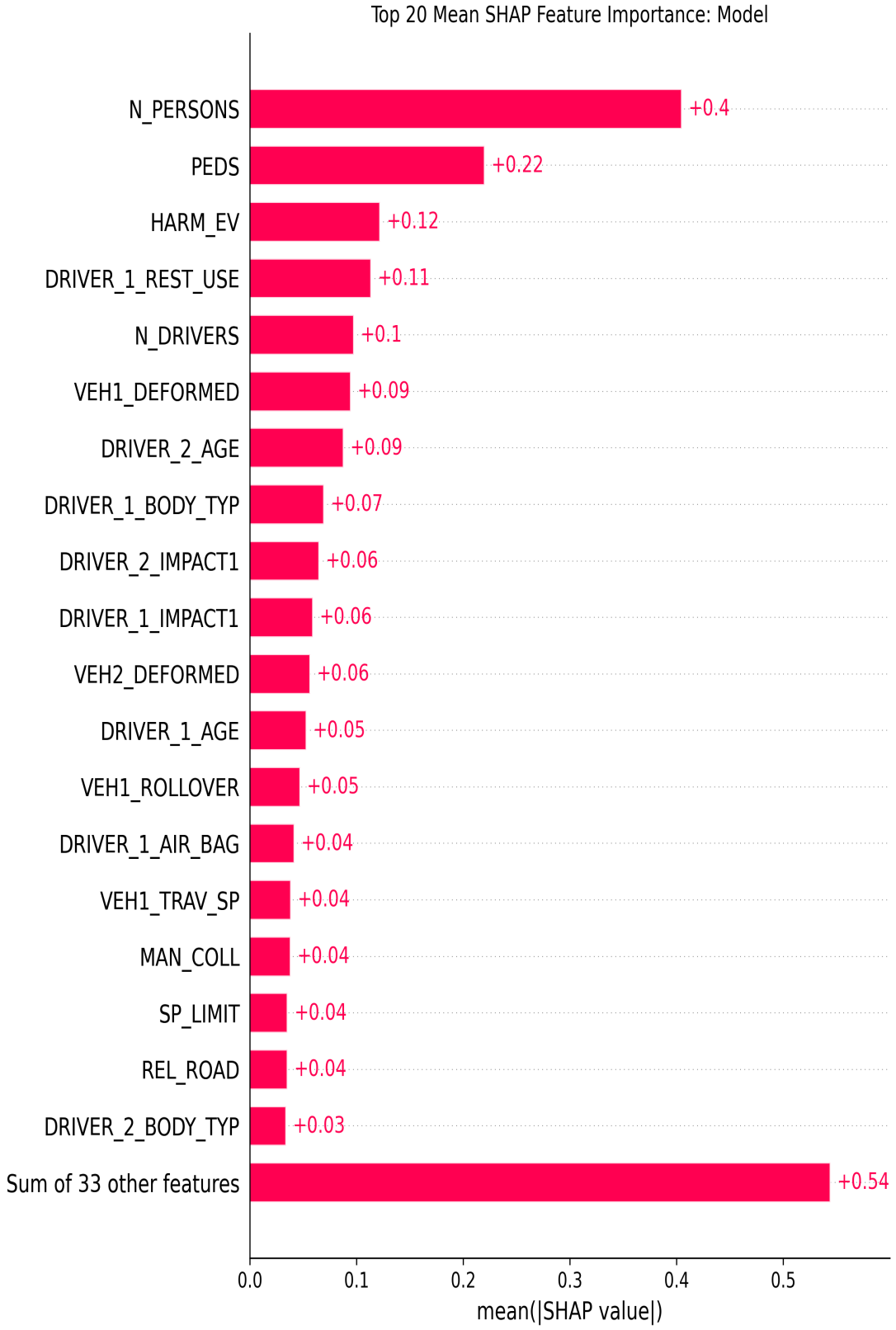

FIGURE 4.10: SHAP Bar Plot for LightGBM

findings in which age, vehicle type, and physical vehicle properties contributed significantly to severity prediction [46, 48, 50]. Travel speed, collision manner, speed

limit, and relative road position provided additional predictive information, consistent with earlier findings concerning speed environment, collision configuration, and roadway context [47,49]. Overall, LightGBM combined human exposure, collision mechanics, occupant protection, driver characteristics, vehicle attributes, and roadway conditions, while the greatest predictive influence remained concentrated in exposure related and physical crash characteristics.

4.2.6.2 SHAP Explanation of Features

The LightGBM beeswarm plot shown in Figure 4.12 indicated that N_PERSONS and PEDS had produced the widest variation in model output. Higher values of both features were concentrated predominantly on the negative side, whereas lower values extended towards the positive side. N_DRIVERS showed the opposite pattern, with higher values contributing mainly towards positive model output. HARM_EV, DRIVER_1_REST_USE, and VEH1_DEFORMED also produced substantial variation, while the mixed distribution of their coded values indicated heterogeneous effects across crash observations and categories.

Among the remaining predictors, higher values of driver age, vehicle body type, second vehicle impact location, second vehicle deformation, rollover status, collision manner, speed limit, and relative road position generally appeared more frequently on the positive side. HIT_RUN showed a particularly distinct pattern, with higher coded values concentrated strongly on the negative side. Overall, the LightGBM model had relied on a combination of human exposure, collision circumstances, occupant protection, vehicle characteristics, driver attributes, speed, and roadway conditions, while the broad spread and mixed distributions of several features had reflected nonlinear and observation specific effects within the model.

The Beeswarm Plot (WHY part of SHAP) reinforced the logic adopted by XGBoost and Random Forest in that the number of personnel and pedestrians involved in a crash are strong predictors of severity, it also explained that the relationship between either of these with severity is not linear. If a large number of people are on the road and they happen to be involved in a crash, most likely the

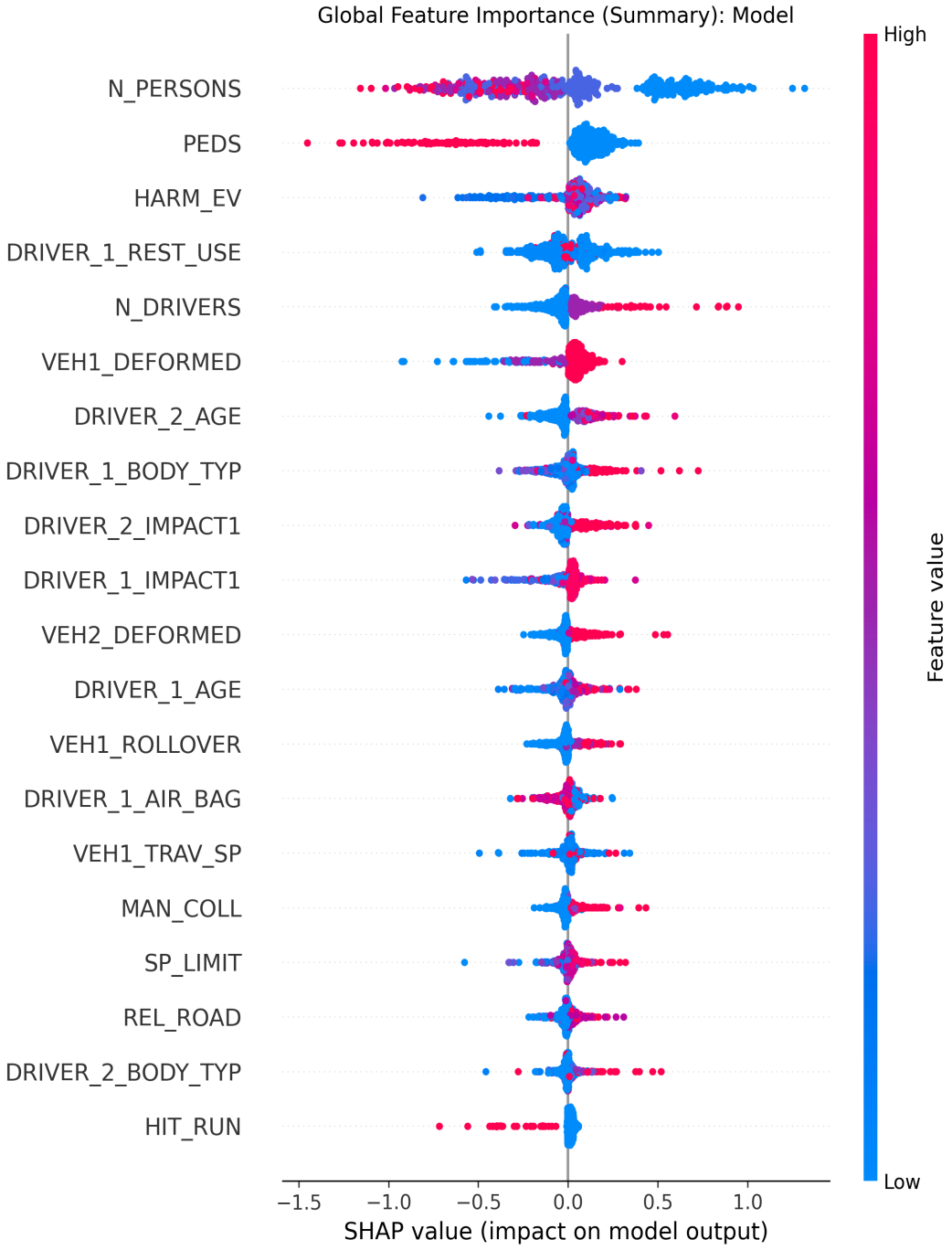

FIGURE 4.11: Beesswarm Plot for LightGBM Model

crash is less severe because of visible obstruction ahead and reduction in speed, or careful driving. However, the SHAP also explained that the large number of drivers, with low number of pedestrians may lead to serious crashes with fatalities which means a succession of crashes or multiple vehicles crashing at a junction, as the case in Pakistan, where on Motorways, especially M2 and M4, and interstate highways like Grand Trunk Road usually experience successive crashes

during the fog and rain, and multi-vehicle collisions at intersections/junctions on GT Road, (notoriously, Gujjjar Khan, Thapla, Morr Aimanabad and Muridke) intersections. The LightGBM SHAP analysis showed that crash severity predictions had been driven by the combined influence of human exposure, pedestrian involvement, collision mechanics, vehicle characteristics, occupant protection, driver attributes, and roadway conditions, consistent with the multidimensional severity patterns reported in previous studies [46–49]. The beeswarm plot indicated nonlinear and observation specific relationships, particularly for N_PERSONS, PEDS, and N_DRIVERS, while the waterfall case demonstrated how these factors acted together in an individual prediction. Although N_PERSONS was the most influential feature globally, it contributed towards a lower model output in the selected case, whereas pedestrian absence, vehicle body type, driver age, number of drivers, and collision manner collectively moved the output from the baseline value of 3.06 towards the final prediction of 4. This distinction between global importance and local contribution demonstrated the value of SHAP for explaining both the overall predictive structure and individual crash severity predictions [46, 48, 49].

The case study of an individual prediction is also explained here. The Waterfall Plot (shown in Figure 4.13) explains the HOW part of the question, using a case study (selected by SHAP algorithm), it is not a timeline, rather a score card of forces leading to the prediction. The LightGBM waterfall plot showed how the selected crash case moved from a baseline model output of 3.06 to the final output of 4. The strongest upward contributions came from the absence of pedestrians and the body type of the first vehicle, each adding approximately 0.20, followed by the age of the second driver at 0.14, the number of drivers at 0.13, and the manner of collision at 0.11. Vehicle deformation, the age of the first driver, harmful event, airbag status, restraint use, temporal conditions, speed limit, vehicle model year, surface condition, and travel speed provided additional upward contributions. The strongest downward contribution came from the presence of three persons, which reduced the model output by 0.22, while relative road position and road function reduced the output by 0.05 and 0.03 respectively.

The case indicated that the final prediction had resulted from the cumulative

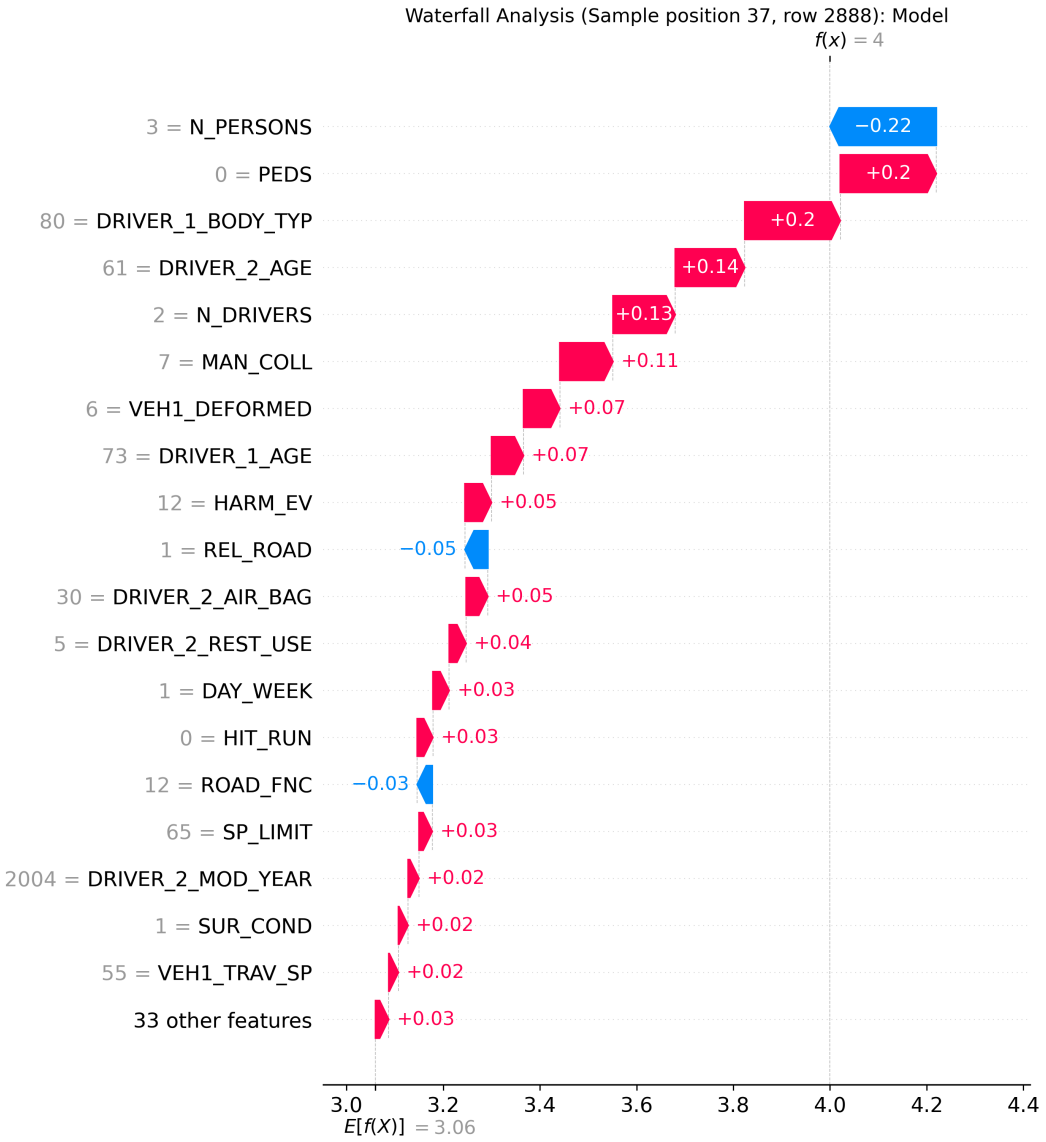

FIGURE 4.12: Waterfall Plot for LightGBM Model

influence of human exposure, vehicle characteristics, driver attributes, collision configuration, occupant protection, roadway conditions, and operating characteristics. Although the number of persons produced a substantial downward contribution, it had been outweighed by the combined upward effects of pedestrian absence, vehicle body type, driver age, number of drivers, collision manner, and other supporting factors. The local explanation also demonstrated that globally important features could act in opposing directions within an individual case, as N_PERSONS reduced the model output while several moderately ranked variables collectively moved the prediction towards the final output of 4. The Waterfall Plot was evaluated for the LightGBM model as discussed below:

4.2.7 Comparative SHAP Analysis of Random Forest, XGBoost and LightGBM Models

A comparative analysis of SHAP Analysis for all models was performed which is presented in Table 4.8. The comparative SHAP scores showed substantial agreement among Random Forest, XGBoost, and LightGBM regarding the principal dimensions of crash severity prediction, despite differences in the ranking and magnitude of individual features. N_PERSONS ranked first in all three models, with SHAP values of 0.4015, 0.3671, and 0.4049 respectively, while PEDS consistently ranked second. This agreement indicated that human exposure and pedestrian involvement formed the most stable predictive dimensions across the three modelling approaches. Collision and occupant related factors also showed strong consistency, as VEH1_DEFORMED, DRIVER_1_REST_USE, HARM_EV, and DRIVER_2 _IMP- ACT1 appeared among the leading predictors in all three models. Similarly, driver age, vehicle body type, impact location, second vehicle deformation, rollover, collision manner, and relative road position repeatedly appeared within the top 20 features, demonstrating broad agreement regarding the importance of human, vehicle, collision, and roadway characteristics.

Differences among the models reflected variations in how each algorithm represented the predictive structure of the data. Random Forest assigned comparatively greater importance to VEH1_DEFORMED, PEDESTRIAN_INVOLVED, and ALIGNMNT, while XGBoost placed greater emphasis on PEDS, N_DRIVERS, and restraint use. LightGBM assigned relatively greater importance to HARM_EV and DRIVER_2_AGE, while also including travel speed, speed limit, and hit and run status among its leading predictors. These variations indicated that the models did not rely on identical feature combinations even when predicting the same outcome. Nevertheless, the repeated dominance of exposure related variables and the recurring presence of collision mechanics, occupant protection, driver characteristics, vehicle attributes, and roadway conditions demonstrated a high degree of thematic agreement across the three models. The comparison therefore suggested that the principal crash severity relationships were sufficiently robust to be identified by different ensemble learning architectures, while model specific rankings

Table 4.7: Comparison of SHAP scores for Random Forest, XGBoost, and LightGBM Models

Random Forest		**XGBoost**		**LightGBM**	
Feature Name	**SHAP**	**Feature Name**	**SHAP**	**Feature Name**	**SHAP**
N_PERSONS	0.4015	N_PERSONS	0.3671	N_PERSONS	0.4049
PEDS	0.2309	PEDS	0.2873	PEDS	0.2199
VEH1_DEFORMED	0.1271	DRIVER_1_REST_USE	0.1090	HARM_EV	0.1218
DRIVER_1_REST_USE	0.1099	VEH1_DEFORMED	0.1057	DRIVER_1_REST_USE	0.1136
HARM_EV	0.0996	N_DRIVERS	0.1044	N_DRIVERS	0.0975
PEDESTRIAN_INVOLVED	0.0780	HARM_EV	0.0996	VEH1_DEFORMED	0.0945
DRIVER_2_IMPACT1	0.0734	DRIVER_2_IMPACT1	0.0600	DRIVER_2_AGE	0.0876
VEH_1_BODY_TYP	0.0629	DRIVER_1_BODY_TYP	0.0600	DRIVER_1_BODY_TYP	0.0694
ALIGNMNT	0.0598	DRIVER_1_IMPACT1	0.0593	DRIVER_2_IMPACT1	0.0647
DRIVER_1_IMPACT1	0.0580	DRIVER_2_AGE	0.0485	DRIVER_1_IMPACT1	0.0591
VEH2_DEFORMED	0.0566	DRIVER_1_AGE	0.0483	VEH2_DEFORMED	0.0565
VEH_2_BODY_TYP	0.0519	VEH2_DEFORMED	0.0444	DRIVER_1_AGE	0.0528
MAN_COLL	0.0508	VEH1_ROLLOVER	0.0418	VEH1_ROLLOVER	0.0471
VEH1_ROLLOVER	0.0474	DRIVER_1_AIR_BAG	0.0397	DRIVER_1_AIR_BAG	0.0416
DRIVER_1_AGE	0.0459	REL_ROAD	0.0390	VEH1_TRAV_SP	0.0384
REL_ROAD	0.0378	MAN_COLL	0.0362	MAN_COLL	0.0380
VEH2_2_MOD_YEAR	0.0378	VEH1_TRAV_SP	0.0307	SP_LIMIT	0.0353
VEH_2_AIR_BAG	0.0373	ALIGNMNT	0.0298	REL_ROAD	0.0351
N_DRIVERS	0.0371	DRIVER_2_BODY_TYP	0.0297	DRIVER_2_BODY_TYP	0.0339
DRIVER_2_AGE	0.0368	SP_LIMIT	0.0267	HIT_RUN	0.0295

reflected differences in their learning mechanisms and representation of nonlinear relationships and feature interactions.

4.3 Model Performance

To assess the model's performance, the performance metrics were calculated using the test data. It is evident from the model performance metrics tables that the Random Forest model performed the best with highest accuracy, precision, recall, and F1-score. The primary tool used for evaluation is the confusion matrix, an $N \times N$ table that maps the actual target values against the model's predictions. We utilize the matrix to compute aggregated performance indicators instead of detailing the absolute counts of every prediction outcome. The Random Forest model was evaluated on the held-out 20% test set (N = 35,984). The confusion matrix is shown in Table 4.8.

TABLE 4.8: Confusion Matrix from Random Forest Model (Test Set)

Actual Class	**Predicted Class**				
	0	**1**	**2**	**3**	**4**
0	2,236	71	139	92	2,410
1	54	1,159	52	35	115
2	167	108	1,710	92	107
3	152	79	96	1,926	164
4	2,253	75	153	172	22,367

Weighted recall is mathematically identical to accuracy in this multiclass single-label setting, so the two numbers coincide. The model's strongest performance is on the fatal class (Class 4), which dominates the dataset. The key limitation is the weak performance on Class 0 (Property Damage Only). From the confusion matrix, Class 0 precision is 0.460 and recall is 0.452; the model frequently misclassifies uninjured drivers as fatally injured. This is a structural consequence of the FARS census of fatal crashes: Class 0 cases are rare and their crash features often overlap with those of more severe outcomes. In the intended forensic and risk prioritization use cases, the focus is on distinguishing the higher injury levels,

where the model is considerably more reliable. The weighted F1-score of 0.816 was the primary selection criterion, and the Random Forest achieved the highest value among the three ensemble models. The model performance metrics are presented in Table 4.9.

TABLE 4.9: Model Performance Metrics for Random Forest Model

Accuracy	**Macro Precision**	**Macro F1**	**Weighted Precision**	**Weighted Recall**	**Weighted F1**
0.7789	0.7499	0.7494	0.7819	0.7789	0.7802

The XGBoost model was evaluated on the same held-out test set (N = 35,984). The confusion matrix is presented in Table 4.10.

TABLE 4.10: Confusion Matrix from XGBoost Model (Test Set)

Actual Class	**Predicted Class**				
	0	**1**	**2**	**3**	**4**
0	1,172	71	118	104	3,483
1	45	1,297	25	11	37
2	124	73	1,878	52	57
3	82	82	71	2,126	56
4	3,729	0	9	12	21,270

The model performance metrics for XGBoost were determined using the confusion matrix, and are presented in Table 4.11.

TABLE 4.11: Model Performance Metrics for XGBoost Model

Accuracy	**Macro Precision**	**Macro F1**	**Weighted Precision**	**Weighted Recall**	**Weighted F1**
0.7710	0.7499	0.7488	0.7748	0.7710	0.7728

The LightGBM model showed better metrics than XGBoost. The confusion matrix from the test set is presented in Table 4.12.

The model performance metrics for LightGBM were also determined using its confusion matrix, and are presented in Table 4.13.

TABLE 4.12: Confusion Matrix from LightGBM Model (Test Set)

Actual Class	**Predicted Class**				
	0	**1**	**2**	**3**	**4**
0	1,285	171	82	110	3,300
1	64	1,237	37	42	35
2	55	37	2,003	42	47
3	148	77	77	2,032	83
4	3,550	0	0	0	21,470

TABLE 4.13: Model Performance Metrics for LightGBM

Accuracy	**Macro Precision**	**Macro F1**	**Weighted Precision**	**Weighted Recall**	**Weighted F1**
0.7789	0.7498	0.7494	0.7819	0.7789	0.7802

4.4 Discussion

This study sought to answer three specific research questions concerning feature attribution, directional effects, and comparative model performance. The findings are discussed below in that order, with the principal limitations woven into the interpretation rather than treated as an afterthought.

4.4.1 Features most strongly associated with injury severity (Research Question 1)

The global SHAP analysis of the Random Forest model (Table 4.4) identified a multidimensional structure of feature importance. The top ranks were occupied by human exposure and pedestrian involvement variables (N_PERSONS, PEDS, PEDESTRIAN_INVOLVED, N_DRIVERS), collision mechanics (VEH1 _ DEFORMED, HARM _ EV, impact location, MAN_COLL), and occupant protection (DRIVER _ 1 _ REST _ USE, body type categories). This pattern was broadly consistent with earlier FARS based studies that have highlighted the role of participant type, collision configuration, and restraint use as strong predictors

of injury severity. The present results extend those findings by quantifying importance within a unified expected severity framework and by demonstrating that the signal is not monopolized by any single variable group.

However, the interpretation of feature importance must account for the mechanical relationship between N_PERSONS and the pedestrian related variables. A crash involving a pedestrian will almost always have a low vehicle occupancy, and vice versa, so the model's attribution of importance to each is a partitioning of a shared exposure signal rather than an independent causal estimate. Consequently, the combined importance of these variables should be understood as representing a broad exposure and vulnerable road user dimension. This limitation is inherent to any post hoc explanation method applied to correlated predictors.

Moreover, because the FARS dataset is a census of fatal crashes, the observed importance structure applies only to drivers involved in events with at least one death. The rankings do not describe the determinants of injury severity in the general crash population, where property damage only and minor injury crashes predominate. The low frequency of Class0 (Property Damage Only) cases in the sample and the model's correspondingly weak precision and recall for that class (Section 4.3) are direct consequences of this sampling design.

4.4.2 Direction and Magnitude of Feature Contributions (Research Question 2)

The beeswarm and waterfall plots showed that higher deformation extent, unbelted status, and specific impact configurations were associated with higher expected severity, whereas belted restraint use and certain body type categories pushed predictions downward. These directional effects are model learned associations and should not be interpreted as causal proof. For instance, the observation that more recent vehicle model years sometimes carried a positive SHAP contribution does not imply declining vehicle build quality; it may reflect fleet composition differences or confounding by crash type within the 2005-2009 time-frame.

A critical interpretive distinction separates pre crash features (road functional class, light condition, speed limit) from impact and post crash variables (deformation, airbag deployment). The presence of the latter means the model cannot function as a pre crash warning tool. Its utility, as defined in the research scope, lies in forensic structural analysis (mapping impact dynamics to trauma) and infrastructure risk prioritization (identifying roadway factors that amplify severity when a crash occurs). Pre crash features, though less prominent in the importance ranking, still contributed systematic information, particularly for intermediate severity classes, and their role underscores the potential for roadway targeted interventions.

It must also be noted that the model's feature space includes variables that are not known at the moment of the crash but are recorded afterward, such as vehicle deformation and airbag deployment. This ex post availability is compatible with the diagnostic use cases but limits the direct translation of SHAP effects into pre crash risk scores.

4.4.3 Comparative Model Performance (Research Question 3)

All three ensemble models were evaluated on the same hold out test set using accuracy, weighted precision, weighted recall, and weighted F1 score. Random Forest achieved the highest weighted F1 score (0.816), followed by LightGBM (0.780) and XGBoost (0.773). The confusion matrices revealed a shared strength, high recall for the fatal class, and a common weakness of poor precision and recall for Class0 (Property Damage Only), a structural limitation of the fatal crash sampling frame. SMOTE was applied within the training folds to mitigate imbalance during learning, but the test set preserved the natural distribution, so the reported metrics reflect real world class proportions.

The SHAP feature rankings were compared across models using a thematic ranking table (Table 4.Y), rather than by directly comparing absolute SHAP magnitudes. Although all models used the same expected severity scale (0-4) and the same

background sample, the absolute SHAP values are model specific quantities and cannot be treated as universally comparable effect sizes. The cross model comparison therefore focused on the stability of top ranking features and on the consistency of the thematic groups. The table shows that N_PERSONS and PEDS occupied the first two ranks in all three models, and that collision mechanics, occupant protection, and roadway characteristics formed similar clusters, albeit with minor rank shifts (e.g., HARM_EV ranked third in LightGBM but fifth in Random Forest). This qualitative consistency across different algorithmic families supports the robustness of the identified predictive structure, even though formal rank stability across cross validation folds was not assessed.

4.4.4 Integrated Limitations

Several limitations define the boundaries within which these findings should be interpreted.

i. Fatal crash sampling frame: FARS conditions on at least one fatality, so the results describe injury severity patterns among drivers involved in the most severe crashes. They cannot be generalized to all road crashes.

ii. Geographical transferability: The deliberate selection of region independent features was intended to make the framework conceptually portable to Pakistan. However, as stated in the methodology, feature deletion and engineering alone do not establish representativeness. Empirical recalibration with Pakistani crash data, including local vehicle fleet and post crash care variables, is a necessary precondition for operational deployment.

iii. Class imbalance: Despite the use of SMOTE during training, the test set retained the natural imbalance, and the model's performance on the rare Class0 was poor. This limitation is inherent to the sampling design and should be considered when the model is used for applications requiring reliable identification of uninjured drivers.

iv. Possible group level leakage: Because the unit of analysis is the individual driver, multiple drivers from the same crash event appear as separate observations. The train-test split did not group by crash identifier, so some dependencies may exist between training and test instances. The effect on performance estimates is expected to be small given the large sample, but it is acknowledged.

v. Test set evaluation: The metrics were computed on a single 20% hold out split with a fixed random seed. While this is standard practice, the reported numbers represent point estimates; cross validation of the test set was not performed, so the stability of the performance differences across multiple splits is not quantified.

vi. Variable availability: The model uses post crash variables that would not be known before a collision, which aligns with the forensic and diagnostic use cases but precludes use as a real time warning system.

4.5 Summary

This study has shown that a high performance ensemble model, trained on FARS and interpreted through expected severity SHAP, can extract transparent, multidimensional feature attribution patterns that support forensic safety analysis and infrastructure risk prioritization. The methodology offers a methodological blueprint for contexts where high quality crash data are scarce, but its direct application to Pakistan remains contingent on future local data collection and recalibration. Future work should prioritize per class SHAP analyses, formal stability assessment of feature rankings, and the incorporation of crash level clustering to strengthen inferential validity.

Chapter 5

Conclusion and Recommendations

5.1 Conclusion

This study set out to determine whether ensemble machine learning, combined with SHAP based explainable artificial intelligence, could extract transparent and actionable diagnostic knowledge from a large fatal crash dataset. Using the U.S. Fatality Analysis Reporting System and an ordinal expected severity framework, the analysis identified the features most closely associated with driver injury severity, characterized the direction and magnitude of their contributions, and compared the predictive performance of three tree based ensemble models. The principal findings, each bounded by the study's scope and limitations, are summarized below:

i. The predictive signal in the data is distributed across several dimensions rather than concentrated in a single variable group. Human exposure and pedestrian involvement captured by the number of persons in the vehicle, pedestrian presence, and pedestrian count occupied the highest importance ranks in all three models. Collision mechanics (vehicle deformation, harmful event, impact location) and occupant protection (driver restraint use, vehicle

body type) formed a second major cluster, followed by roadway characteristics and driver age. The prominence of exposure related features must be understood in light of the mechanical link between occupancy and pedestrian involvement; their individual importance values represent a shared signal from the broader exposure dimension.

ii. Directional SHAP contributions were consistent with physical expectations. Belted restraint use was associated with lower expected severity, while greater vehicle deformation and certain impact configurations were associated with higher severity. These relationships are model learned associations and are not presented as causal proof. The distinction between pre crash factors (road functional class, light conditions, speed limit) and post crash factors (deformation, airbag deployment) was critical: the model's diagnostic value lies in forensic safety analysis and infrastructure risk prioritization, not in real time collision warning.

iii. Pedestrian involvement was strongly linked to higher expected severity in the model, reinforcing the importance of protecting vulnerable road users. Restraint use emerged as one of the most influential modifiable factors, with consistent downward pressure on predicted severity. Airbag deployment also contributed to severity predictions, though with moderate average importance.

iv. Among the three models evaluated on the same hold out test set, Random Forest achieved the highest weighted F1 score, followed by LightGBM and XGBoost. The thematic feature rankings were qualitatively consistent across all three algorithms, lending credibility to the identified importance structure. The comparison was based on ranks rather than absolute SHAP magnitudes, respecting the model specific nature of SHAP values.

v. Several limitations define the boundaries of these conclusions. The FARS dataset is a census of fatal crashes, so the findings apply only to drivers involved in events that already resulted in at least one death. The deliberate selection of region independent features makes the methodological framework

conceptually portable to a context such as Pakistan, but empirical recalibration with local crash data is a prerequisite for operational deployment. The presence of multiple drivers from the same crash introduces possible dependencies that were not explicitly modelled. Class imbalance, inherent to the fatal crash sampling frame, resulted in weaker performance on the rare property damage only category. These limitations do not invalidate the contribution but specify the conditions under which the results should be interpreted.

In sum, this research demonstrates that a carefully designed ensemble model, interpreted through expected severity SHAP, can serve as a transparent diagnostic tool that bridges the gap between predictive accuracy and engineering interpretability. The approach yields a replicable methodological blueprint for extracting structural safety hypotheses from large scale crash data. Future work should focus on validating and recalibrating this blueprint with locally collected data in low and middle income settings, and on extending the explainability framework to per class SHAP analyses for the rarer, non fatal injury outcomes.

5.2 Future Recommendations

The findings of this research provide opportunities for further development of crash severity prediction and analysis. Future studies can build upon the present work by improving the quality and local relevance of crash data, introducing additional explanatory variables, and exploring advanced modelling approaches. Greater attention should also be given to the spatial and temporal dimensions of road crashes and to the practical use of predictive models by road safety authorities.

Following are the future recommendations:

i. Incorporation of AI models into Realtime Intelligent Transportation System (ITS), to enable early crash and severity identification for rapid response and control for prioritized emergency services and to alert the authorities about the crash site.

ii. The agencies should consider proper road geometrics, enforce control and discipline on the road.

iii. Data source augmentation from safe city cameras, sensors and meteorological data can significantly improve road safety and reduce lane closures, leading to reduction in travel time.

iv. Future studies should use locally collected and standardized crash data from Pakistan to develop models that better represent local traffic, roadway, vehicle, and road user conditions.

Bibliography

[1] World Health Organization, "Global status report on road safety 2023," World Health Organization, Geneva, Switzerland, Tech. Rep., 2023. [Online]. Available: https://www.who.int/publications/i/item/9789240086517

[2] ——, "Road traffic injuries: Fact sheet," https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries, World Health Organization, Geneva, Switzerland, May 2024.

[3] A. Sartori, A. Russo, A. Sardo, and D. Raniero, "Fatal road traffic accidents and injuries: a preliminary study," *International Journal of Legal Medicine*, vol. 139, no. 1, pp. 375–381, jan 2025.

[4] "Road accidents in saudi arabia: A comparative and analytical study," accessed: Jan. 20, 2026.

[5] A. Sulong, "The impact of socioeconomic and macroeconomic factors on road accident rates in malaysia," *Journal of Management and Muamalah*, vol. 15, no. 1, pp. 57–67, jun 2025.

[6] K. Boruzs, R. Boda, G. Bányai, V. Dombrádi, and K. Bíró, "Association between road traffic accidents and economic growth with a special focus on alcohol consumption – a comparison of six alpine countries."

[7] D. Levinson, J. M. Mathieu, D. Gillen, and A. Kanafani, "The full cost of high-speed rail: an engineering approach," *The Annals of Regional Science*, vol. 31, no. 2, pp. 189–215, may 1997.

[8] H. E. Rosen *et al.*, "Global road safety 2010–18: An analysis of global status reports," *Injury*, vol. 56, no. 6, p. 110266, jun 2025.

[9] K. Wang and Z. Li, "Global, regional, and national burdens of road injuries from 1990 to 2021: Findings from the 2021 global burden of disease study," *Injury*, vol. 56, no. 3, p. 112221, mar 2025.

[10] L. Huang, "Identifying risk factors for household burdens of road traffic fatalities: regression results from a cross-sectional survey in taiwan," *BMC Public Health*, vol. 16, no. 1, p. 1202, nov 2016.

[11] B. Bryant, R. Mayou, L. Wiggs, A. Ehlers, and G. Stores, "Psychological consequences of road traffic accidents for children and their mothers," *Psychological Medicine*, vol. 34, no. 2, pp. 335–346, mar 2004.

[12] C. M. Ferreira-Vanegas, J. I. Vélez, and G. A. García-Llinás, "Analytical methods and determinants of frequency and severity of road accidents: A 20-year systematic literature review," *Journal of Advanced Transportation*, 2022.

[13] C. Gutierrez-Osorio and C. Pedraza, "Modern data sources and techniques for analysis and forecast of road accidents: A review," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 7, no. 4, pp. 432–446, aug 2020.

[14] M. E. Nwafor and V. O. Onya, "Road transportation service in nigeria: Problems and prospects," *Advanced Journal of Economics and Marketing Research*, vol. 4, no. 3, pp. 104–117, 2019.

[15] J.-P. Rodrigue, "Transportation and globalization."

[16] A. Pandit, H. Jeong, J. C. Crittenden, and M. Xu, "An infrastructure ecology approach for urban infrastructure sustainability and resiliency," in *2011 IEEE/PES Power Systems Conference and Exposition*, mar 2011, pp. 1–2.

[17] G. Yago, "The sociology of transportation," *Annual Review of Sociology*, vol. 9, pp. 171–190, aug 1983.

[18] E. de Boer, *Transport Sociology: Social Aspects of Transport Planning*. Elsevier, 2013.

[19] Local Government & Community Development Department, "Government of the punjab business rules 2017," Government of the Punjab, Pakistan, Tech. Rep., 2017.

[20] Z. Umar *et al.*, "Risky driving behaviours and their association with fracture injury severity in road traffic accident patients," *Cureus*, dec 2025.

[21] M. S. Luqman, U. Tanveer, B. A. Qureshi, A. Zeba, N. Tanveer, and S. B. Al-Mhanna, "A retrospective analysis of risk factors associated with road traffic accident fatalities," *Pakistan Journal of Medical and Health Sciences*, vol. 19, no. 08, pp. 14–19, sep 2025.

[22] M. J. Ullah, D. H. Khan, D. I. Wahid, and D. A. Ullah, "Substance use and drivers' perception of road accidents: Evidence from district dir lower, khyber pakhtunkhwa, pakistan," *Journal of Social Sciences Research and Policy*, vol. 3, no. 03, pp. 504–514, sep 2025.

[23] A. Hayat, H. Irfan, S. Haider, and M. Zaman, "Normalization of risk within road traffic culture in islamabad, pakistan," nov 2025, sSRN ID: 5830339.

[24] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal*, vol. 20, no. 1, 2020, accessed: Jul. 01, 2026.

[25] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, oct 2001.

[26] K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: from early developments to recent advancements," *Systems Science & Control Engineering*, vol. 2, no. 1, pp. 602–609, dec 2014.

[27] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, accessed: Jul. 01, 2026.

[28] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *IEEE Access*, 2022.

[29] National Highway Traffic Safety Administration (NHTSA), "National highway traffic safety administration: Introduction," U.S. Department of Transportation, Tech. Rep. DOT HS 810 552, 2006, accessed: Feb. 11, 2025.

[30] L. J. Blincoe, T. R. Miller, E. Zaloshnja, and B. Lawrence, "The economic and societal impact of motor vehicle crashes, 2010 (revised)," National Highway Traffic Safety Administration, Tech. Rep., 2015, accessed: Jan. 21, 2026. Available: https://rosap.ntl.bts.gov.

[31] I. R. H. Rockett *et al.*, "Leading causes of unintentional and intentional injury mortality: United states, 2000–2009," *American Journal of Public Health*, vol. 102, no. 11, pp. e84–e92, nov 2012.

[32] S. Hakim, D. Shefer, A. S. Hakkert, and I. Hocherman, "A critical review of macro models for road accidents," *Accident Analysis & Prevention*, vol. 23, no. 5, pp. 379–400, oct 1991.

[33] V. Sharma, R. Upadhyay, and S. Kumar, "A systematic review on road traffic accident: Causes and control measures," *Journal of Research in Applied Science and Engineering Technology*, apr 2023.

[34] P. Poumanyvong, S. Kaneko, and S. Dhakal, "Impacts of urbanization on national transport and road energy use: Evidence from low, middle and high income countries," *Energy Policy*, vol. 46, pp. 268–277, jul 2012.

[35] R. d. Sharma, S. Jain, and K. Singh, "Growth rate of motor vehicles in india - impact of demographic and economic development," *Journal of Economic and Social Studies*, vol. 1, no. 1, pp. 137–153, jan 2011.

[36] C.-W. Wang and C. L. W. Chan, "Estimated trends and patterns of road traffic fatalities in china, 20022012," *Traffic Injury Prevention*, vol. 17, no. 2, pp. 164–169, feb 2016.

[37] X. Zhang, H. Yao, G. Hu, M. Cui, Y. Gu, and H. Xiang, "Basic characteristics of road traffic deaths in china," *Iranian Journal of Public Health*, vol. 42, no. 1, pp. 7–15, jan 2013.

[38] A. F. Williams, S. L. Oesch, A. T. McCartt, E. R. Teoh, and L. B. Sims, "On-road all-terrain vehicle (atv) fatalities in the united states," *Journal of Safety Research*, vol. 50, pp. 117–123, sep 2014.

[39] F. A. Wilson and J. P. Stimpson, "Trends in fatalities from distracted driving in the united states, 1999 to 2008," *American Journal of Public Health*, vol. 100, no. 11, pp. 2213–2219, nov 2010.

[40] S. Saha, P. Schramm, A. Nolan, and J. Hess, "Adverse weather conditions and fatal motor vehicle crashes in the united states, 1994-2012," *Environmental Health*, vol. 15, no. 1, p. 104, nov 2016.

[41] I. Savage, "Comparing the fatality risks in united states transportation across modes and over time," *Research in Transportation Economics*, vol. 43, no. 1, pp. 9–22, jul 2013.

[42] R. L. Sanders and R. J. Schneider, "An exploration of pedestrian fatalities by race in the united states," *Transportation Research Part D: Transport and Environment*, vol. 107, p. 103298, jun 2022.

[43] E. Zaloshnja and T. R. Miller, "Cost of crashes related to road conditions, united states, 2006," *Annals of Advances in Automotive Medicine / Annual Scientific Conference*, vol. 53, pp. 141–153, oct 2009.

[44] I. Siddique, "Advanced analytics for predicting traffic collision severity assessment," feb 2024, sSRN ID: 4894183, Accessed: Jan. 21, 2026.

[45] M. Manzoor *et al.*, "Rfcnn: Traffic accident severity prediction based on decision level fusion of machine and deep learning model," *IEEE Access*, vol. 9, pp. 128 359–128 371, 2021.

[46] V. Machaca Arceda and E. Laura Riveros, "Fast car crash detection in video," in *2018 XLIV Latin American Computer Conference (CLEI)*, oct 2018, pp. 632–637.

[47] R. Desai, A. Jadhav, S. Sawant, and N. Thakur, "Accident detection using ml and ai techniques," 2024.

[48] "Computer vision-based accident detection in traffic surveillance," accessed: Jan. 21, 2026.

[49] K. Kovačić, E. Ivanjko, and H. Gold, "Computer vision systems in road vehicles: A review," oct 2013.

[50] M. Essam, N. M. Ghanem, and M. A. Ismail, "Detection of road traffic crashes based on collision estimation," in *Artificial Intelligence and Machine Learning*, jul 2022, pp. 169–179.

[51] U. Lokala, S. Nowduri, and P. K. Sharma, "Road accidents bigdata mining and visualization using support vector machines," *World Academy of Science, Engineering and Technology - International Journal of Computer and Systems Engineering*, vol. 10, no. 8, jul 2017.

[52] U. S. Kumar, "An efficient analysis of road accidents severity using novel support vector machines over artificial neural networks with improved accuracy rate," *Revista de Gestão, Inovação e Tecnologias*, vol. 11, no. 2, pp. 1109–1122, jun 2021.

[53] J. Li, F. Guo, Y. Zhou, W. Yang, and D. Ni, "Predicting the severity of traffic accidents on mountain freeways with dynamic traffic and weather data," *Transportation Safety and Environment*, vol. 5, no. 4, p. tdad001, sep 2023.

[54] S. L. Karri, L. C. De Silva, D. T. C. Lai, and S. Y. Yong, "Classification and prediction of driving behaviour at a traffic intersection using svm and knn," *SN Computer Science*, vol. 2, no. 3, p. 209, may 2021.

[55] I. K. Umar, "Machine learning approach for predicting truck drivers involvement in an injury accident," *Sigma Journal of Engineering and Natural Sciences*, pp. 808–815, 2025.

[56] R. Zottis Junges, M. Braga De Paula, and M. Sanchotene De Aguiar, "Brazilian traffic signs detection and recognition in videos using clahe, hog feature extraction and svm cascade classifier with temporal coherence," in *Lecture Notes in Computer Science*, L. Martínez-Villaseñor, I. Batyrshin, and A. Maríhn-Hernández, Eds. Cham: Springer, 2019, vol. 11835, pp. 589–600.

[57] G. P. Jose, J. Joseph, O. S. Paimattam, M. M. Joseph, and N. T. Joseph, "System for predicting traffic violations using machine learning," accessed: Jan. 21, 2026.

[58] S. F. Abdul Razak, S. N. M. Sayed Ismail, B. Hii Ben Bin, S. Yogarayan, M. F. A. Abdullah, and N. H. Kamis, "Monitoring physiological state of drivers using in-vehicle sensing of non-invasive signal," *Civil Engineering Journal*, vol. 10, no. 4, pp. 1221–1231, apr 2024.

[59] M. Ding, T. Suzuki, and T. Ogasawara, "Estimation of driver's posture using pressure distribution sensors in driving simulator and on-road experiment," in *2017 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, oct 2017, pp. 215–220.

[60] H. Zeng *et al.*, "A lightgbm-based eeg analysis method for driver mental states classification," *Computational Intelligence and Neuroscience*, vol. 2019, pp. 1–11, sep 2019.

[61] F. Qanouni, H. E. Massari, N. Gherabi, and M. E. Badaoui, "Road accident detection using svm and learning: A comparative study," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 5, may 2024.

[62] A. Mengistu *et al.*, "Predicting car accident severity in northwest ethiopia: A machine learning approach leveraging driver, environmental, and road conditions," 2025.

[63] R. Saleh, "Towards smart maintenance machine-learning based prediction of retroreflectivity and color of road traffic signs," Ph.D. dissertation, Doctoral Dissertation, 2025.

[64] "Integrated analysis of skid resistance, texture, and geometric factors for enhanced road safety on the southern expressway," in *ResearchGate*, dec 2025.

[65] G. Jacobs, A. Aeron-Thomas, and A. Astrop, "Estimating global road fatalities," 2000.

[66] M. Lutz, *Programming Python*. O'Reilly Media, Inc., 2001.

[67] "What is python? executive summary," Python.org, accessed: Jan. 22, 2026.

[68] Open Source Initiative, "The open source definition," opensource.org, accessed: Jul. 01, 2026. Available: https://opensource.org/osd.

[69] "pandas - python data analysis library," accessed: Jan. 22, 2026. Available: https://pandas.pydata.org/about/.

[70] W. McKinney, "pandas: a foundational python library for data analysis and statistics," 2011.

[71] "Io tools (text, csv, hdf5, ...) — pandas 3.0.0 documentation," accessed: Jan. 22, 2026.

[72] W. McKinney, *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter*. O'Reilly Media, 2022, accessed: Jan. 22, 2026.

[73] ——, "Apache arrow and the '10 things i hate about pandas'," Wes McKinney Blog, accessed: Jan. 22, 2026.

[74] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," jun 2018.

[75] National Highway Traffic Safety Administration, "2022 fars/crss coding and validation manual," National Highway Traffic Safety Administration, Tech. Rep. DOT HS 813 556, apr 2024, accessed: Jul. 01, 2026. Available: https://trid.trb.org/View/2362004.

[76] ——, "Budget estimates fiscal year 2017 nhtsa," National Highway Traffic Safety Administration, Tech. Rep., 2016.

[77] National Highway Traffic Safety Administration (NHTSA), "Fatality analysis reporting system (fars) analytical users manual, 19752023," National Highway Traffic Safety Administration (NHTSA), Tech. Rep. Manual DOT HS 813706, 2024, accessed: Feb. 11, 2025.

Appendix-A (Feature Selection and Correlation with Pakistan)

A Feature Selection and Correlation with Pakistan

The correlation does not end with the time span, for a better reflection of Pakistani conditions, only the features reflecting Pakistani situation are selected.

For example, the feature of Drunk/Alcoholic driver is not selected. For any possible value of the feature which does not appear on the Pakistani road and traffic, the entire record/row is dropped.

A comprehensive summary of each feature is presented here. The details of data manipulation and results are provided as part of Jupyter Notebook along with the data for reproducibility.

A.1 MINUTE

Like hours, minutes are also numeric and represent the actual data (non-encoded). Minutes hold valid values from 0 to 59 and 99 for unknown/missing values.

Region Dependency: This column holds temporal data and is not region dependent.

A.2 PERSONS

This column holds the number of "persons" involved in the accident. It is also a representation of actual date.

Region Dependency: This column holds a "count", and the count of people is not region dependent.

A.3 PEDS

Like PERSONS, PEDS are also the people who were involved in the crash, but PEDS column holds only those who were outside the vehicle in transport (moving). These exclude the people who were either driving or were passengers of the vehicle.

Region Dependency: This column holds the number of pedestrians involved in the crash. This data is not region dependent.

Case for Pakistan: In Pakistan, a number of road traffic accidents involve the bystanders or pedestrians. This data closely relates to the situation in Pakistan.

A.4 ROAD_FNC

This feature describes the functional class of the road on which the crash occurred. Since the selected data source compiles crash data from across the nation, irrespective of rural or urban areas, we see the clear distinction between the Rural and Urban sub-features.

Table A.1: Categories under ROAD_FNC

CODE	Rural	CODE	Urban
1	Principal Arterial – Interstate	11	Principal Arterial – Interstate
2	Principal Arterial – Other	12	Principal Arterial – Other Freeways or Expressways
3	Minor Arterial	13	Other Principal Arterial
4	Major Collector	14	Minor Arterial
5	Minor Collector	15	Collector
6	Local Road or Street	16	Local Road or Street
9	Unknown	19	Unknown
99		Unknown if Rural or Urban	

Region Dependency: The road classification is a transportation engineering component not a regional component.

Case for Pakistan: From the table above, it is evident that the road classification in the United States and Pakistan is principally the same. For example, the Grand Trunk Road is classified as Principal Arterial Inter Provincial (Interstate) within the urban limits and Principal Arterial Inter Provincial within the rural limits. The road maintenance is taken over by the district governments and the Local Governments in urban areas of Punjab. While the rural areas and overall maintenance is managed by the NHA maintenance circles and beats.

Similarly, the Motorways are both interstate and principal arterials. We have Expressways in Islamabad and Karachi. The Makran Coastal Highway is a rural interstate highway.

Recently, Pakistan is building China Pakistan Economic Corridor which includes Principal Arterials, Expressways and collectors in both rural and urban settings.

This variable records the functional classification of the road where the crash occurred, following the U.S. Federal Highway Administration scheme (rural/urban principal arterial, minor arterial, collector, local). Functional road class was retained because the concept of a road hierarchy distinguishing high-speed, limited-access facilities from lower-speed local streets is a universal engineering principle relevant to crash severity. The model was trained using the numeric FARS codes as they appear in the original dataset.

No formal mapping between U.S. functional classes and Pakistani administrative or inter-provincial road categories is undertaken in this study. The classification systems differ in their conceptual basis: the U.S. scheme is mobility-oriented and geometry-based, while Pakistani road categories often reflect ownership and administrative responsibility (e.g., National Highway, Motorway, provincial road, municipal street). Any discussion of potential correspondences between the two systems is therefore tentative and is not used to filter, relabel, or otherwise modify the data. The feature is used solely as a proxy for the physical and operational characteristics of the roadway, under the assumption that these characteristics influence crash severity in a broadly comparable manner. The transfer of the model to Pakistan would require recoding local road data into a compatible engineering classification, a step that lies beyond the current scope and is discussed in the limitations.

A.5 MAN_COLL (Manner of Collision)

This data element describes the manner which the vehicles or vehicles collided as the harmful event (above) occurred.

Table A.2: Categories under MAN_COLL

CODE	COLISSION
0	Not Collision with Motor Vehicle In-Transport
1	Front-to-Rear
2	Front-to-Front
3	Angle – Front-to-Side, Same Direction
4	Angle – Front-to-Side, Opposite Direction
5	Angle – Front-to-Side, Right Angle (Includes Broadside)
6	Angle – Front-to-Side/Angle-Direction Not Specified
7	Sideswipe – Same Direction
8	Sideswipe – Opposite Direction
9	Rear-to-Side
10	Rear-to-Rear
11	Other (End-Swipes and Others)
98	Not Reported
99	Reported as Unknown

Regional Dependency: The dataset only describes the orientation of crashing vehicles and is region independent.

Case for Pakistan: Looking at the enumerated orientations, it is evident that the manner in which the vehicle collisions occur and are reported, is no different than in Pakistan. Thus, this data element closely corresponds to the ground reality of Pakistan.

A.6 REL_JUNC (Related to Junction)

This data element identifies the crash's location with respect to road junctions (interchanges)

Table A.3: Categories under REL_JUNC

CODE	NON-INTERCHANGE AREA	CODE	INTERCHANGE AREA
1	Non-Junction	10	Intersection
2	Intersection	11	Intersection-Related
3	Intersection-Related	12	Driveway Access
4	Driveway, Alley Access, etc.	13	Entrance/Exit Ramp-Related
5	Entrance/Exit Ramp-Related	14	In Crossover
6	Railway Grade Crossing	15	Other Location in Interchange
7	In Crossover	19	Unknown, Interchange Area
8	Driveway Access Related	99	Unknown
9	Unknown, Non-Interchange		

Regional Dependency: The dataset only describes the location with respect to nearby junction/crossing. This is a transportation feature not a region dependent setting.

Case for Pakistan: In Pakistan, intersections are common to the village level, while junctions are becoming commonplace in the urban areas. Moreover the crashes during the crossover or railway grade crossing are common news. Furthermore, the driveway and alley (small street) related crashes are frequent.

A.7 REL_ROAD (Relevance with Roadway)

This data element determines the location of the crash within or outside the trafficway based on the "First Harmful Event.", as it relates to its position

Table A.4: Categories under REL_ROAD

CODE	Roadway Relation
1	On Roadway
2	On Shoulder
3	On Median
4	On Roadside

5	Outside Trafficway/Outside Right-Of-Way
5	Outside Trafficway
6	Off Roadway – Location Unknown
7	In Parking Lane/Zone (Since 2007)
8	Gore
10	Separator
11	Continuous Left-Turn Lane
12	Pedestrian Refuge Island or Traffic Island
98	Not Reported
99	Reported as Unknown

Regional Dependency: The relevance to the road is purely a trafficway phenomenon nothing in this data column is regional dependent and is applicable to all regions in the world, especially Pakistan.

Case for Pakistan: Looking at the current situation in Pakistan and the possible values for the column, it is easy to spot that these type collisions are the case for Pakistan too. For example, in Pakistan, the shoulder is usually occupied or the cars try to use the shoulder and hence strike another vehicle or the shoulder.

A.8 SP_LIMIT (Speed Limit)

As the name suggests, the data column holds the designated speed limit on the highway section related to the crash, at the time of crash.

Table A.5:: Categories under SP_LIMIT

CODE	SPEED LIMIT
0	No Statuary Speed Limit
1 to 95	Speed Limit in mph
98	Not Reported
99	Unkonwn

A.9 ALIGNMNT

It is the roadway alignment prior to this vehicle's critical precrash event based on the case material.

Table A.6: Categories under ALIGNMENT

CODE	ALIGNMENT
0	Non-Trafficway Area
1	Straight

2	Curved
8, 9	Not Reported, Unknown

A.10 PROFILE

It is the roadway grade prior to this vehicle's critical precrash harmful event.

Table A.7: Categories under PROFILE

CODE	PROFILE
0	Non-Trafficway Area
1	Level
2	Grade
3	Hillcrest
4	Sag
5	Uphill
6	Downhill
8	Not Reported
9	Unknown

A.11 PAVE_TYP (Pavement Type)

This data column describes the roadway surface type prior to this vehicle's critical precrash harmful event

Table A.8: Categories under PAV_TYP

CODE	PAVEMENT TYPE
1	Concrete
2	Blacktop, Bituminous, or Asphalt
3	Brick or Block
4	Slag, Gravel or Stone
5	Dirt
7	Other
8	Other/Not Reported
9	Unknown

A.12 SUR_COND (Surface Condition)

The SUR_COND explains the roadway surface condition prior to this vehicle's critical precrash harmful event.

Table A.9: Categories under SUR_COND

CODE	SURFACE CONDITION
1	Dry
2	Wet
3	Snow or Slush
4	Ice/Frost
5	Sand, Dirt, Mud, Gravel
6	Water (Standing or Moving)
7	Oil
8	Other
9	Unknown
10	Slush
11	Mud
98	Not Reported
99	Unknown

Appendix-B (Data Preprocessing)

A. Data Preprocessing

A.1 Data Preprocessing

Data preprocessing holds great significance in both data analysis and machine learning processes. It involves transforming raw data into a cleaned, organized, and prepared format suitable for analysis. Real-world datasets often contain flaws like missing values, irregular formatting, redundant records, irrelevant data, and noise, which can undermine the reliability of statistical models and machine learning outcomes. By meticulously cleaning and standardizing the data during preprocessing, we eliminate these inconsistencies, ensuring that subsequent analyses are based on accurate and coherent information. This rigorous preparation step enhances the credibility of study results and ultimately contributes to more robust and reliable insights.

A.1.1 Handling Missing Values

The crash reporting system does not allow a missing value, it explicitly requires to fill the appropriate code of ‘unknown’, for example if the seat-belt was not known to be used, it marks it as unknown under seatbelt deployed, rather than leaving it blank. Thus, the term “missing value” means any unknown value. Exploratory data analysis was performed on each of the variables to gain an insight of the data quality and distributions of variables. Distribution charts and heatmaps were produced to understand the data and to spot the missing (unknown) values and to determine the imputation technique suitable for each variable. The dataset contains missing values and labels them as unknowns with proper enumeration. In order to impute the missing values, statistical procedures were adopted. Based on the statistical analysis, the imputation strategy such as mode, mean or ratio were applied. A poorly chosen imputation method can distort the signal and introduce artificial bias, which ultimately skews the feature importance and Gini impurity scores during model training. So instead of using one method of all missing values, each feature was statistically analyzed for suitable method of imputation.

The imputation method of Mode was applied when the feature was discrete (low cardinality) and the distribution was extremely skewed towards a single value. If a value is overwhelmingly dominant, predicting anything else is statistically useless. In practice, while imputing missing values for KABCO injury severity levels, road surface conditions, or light conditions at the time of an incident, if 80% of crashes in your dataset happened in daylight, filling a blank with "Daylight" is the safest approach avoiding any data distortion. The “Ratio” method was used

when the variable was continuous, highly dispersed (no clear center), but strongly correlated with other available features in the dataset with a high Gini (approaching 1.0) or high entropy, with negligible mode frequency. After that the missing values were imputed using relevant and suitable statistical methods determined during the EDA, to assess the data quality after the imputation, descriptive statistical analysis was performed for each of the variables. To analyze the cross-classification of categorical data, Karl Pearson's Chi-Square tests were performed to investigate the associations or differences between categorical features, while for each of the numerical features, Kruskal-Wallis H test was performed.

A.1.1.1 Minute of Crash Analysis

The following table shows the distribution for MINUTE.

Table A.1: Minute of Crash Distribution

Code	Frequency	Percentage
0.0	10939	6.07
1.0	1896	1.05
2.0	1917	1.06
3.0	1883	1.04
4.0	1898	1.05
5.0	5300	2.94
6.0	1839	1.02
7.0	2059	1.14
8.0	2291	1.27
9.0	1857	1.03
10.0	6532	3.62
11.0	1728	0.96
12.0	2130	1.18
13.0	1990	1.1
14.0	2030	1.13
15.0	7413	4.11
16.0	1767	0.98
17.0	1973	1.09
18.0	2207	1.22
19.0	1985	1.1

20.0	7128	3.95
21.0	1745	0.97
22.0	1976	1.1
23.0	1979	1.1
24.0	1957	1.09
25.0	5587	3.1
26.0	1848	1.02
27.0	1946	1.08
28.0	2210	1.23
29.0	1818	1.01
30.0	10422	5.78
31.0	1592	0.88
32.0	1906	1.06
33.0	1907	1.06
34.0	1893	1.05
35.0	5394	2.99
36.0	1885	1.05
37.0	1944	1.08
38.0	2178	1.21
39.0	1862	1.03
40.0	6717	3.72
41.0	1763	0.98
42.0	2128	1.18
43.0	2050	1.14
44.0	1960	1.09
45.0	7496	4.16
46.0	1798	1.0
47.0	1878	1.04
48.0	2131	1.18
49.0	1838	1.02
50.0	7118	3.95
51.0	1778	0.99

52.0	1996	1.11
53.0	1977	1.1
54.0	1964	1.09
55.0	5398	2.99
56.0	1846	1.02
57.0	1997	1.11
58.0	2215	1.23
59.0	1845	1.02
99.0	1632	0.9

The distribution is visualized as below:

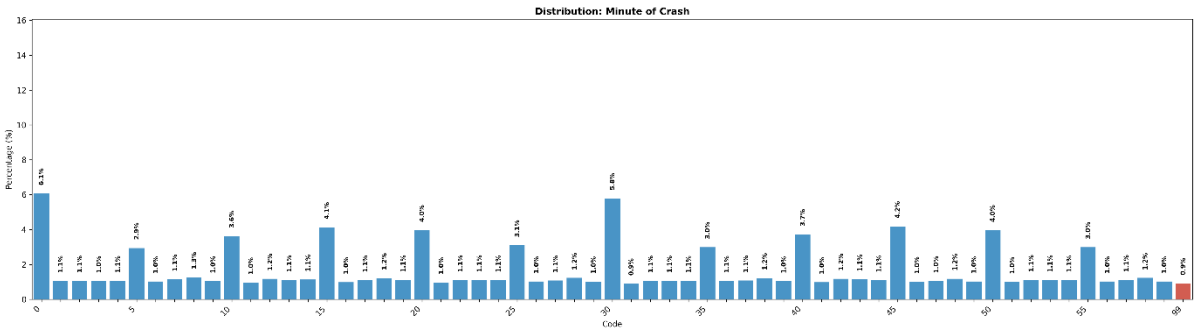

Figure A.1: Distribution Chart for Minute of Crash

To impute the unknown values for MINUTE, statistical assessment was performed results of which are presented below:

Table A.2: Imputation Assessment: Minute of Crash

Metric	Value
Count	178704
Unique	60
Entropy	5.6052
Gini	0.9741
Mode	0.0
Mode_Freq	0.0612
Strategy	Ratio

Imputation Strategy	Ratio

The distribution after imputation is presented below:

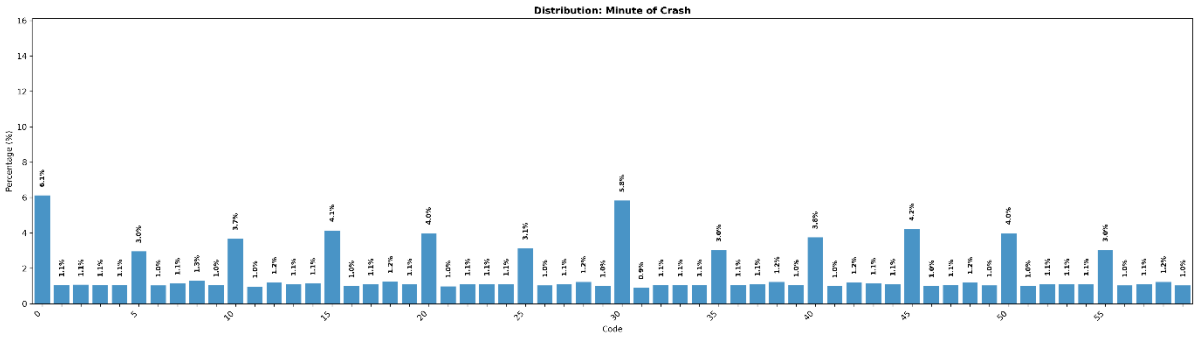

Figure A.2: Post Imputation Distribution Chart

The heatmap of the MINUTE is presented below:

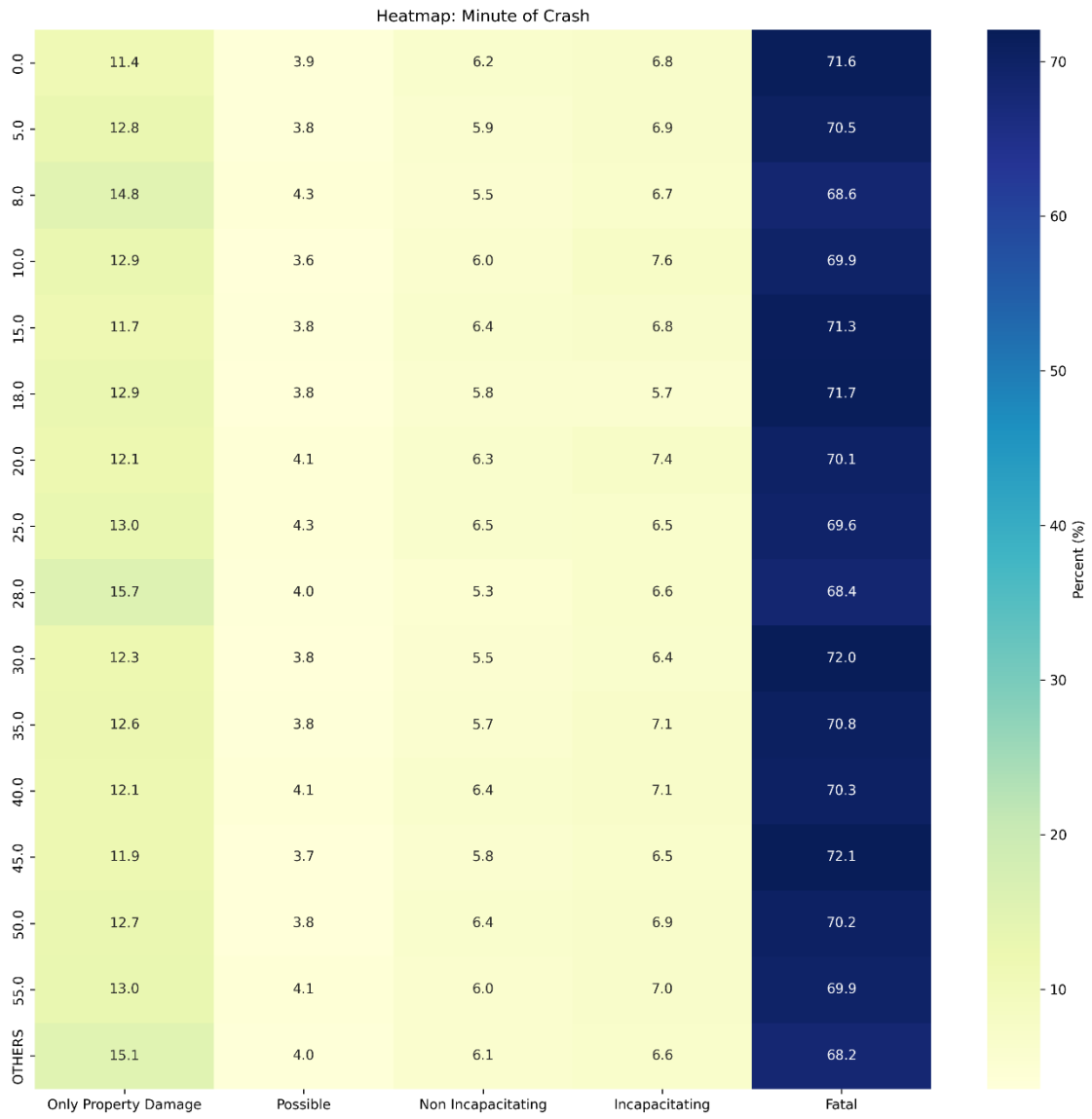

Figure A.3: Heatmap: Minute of Crash vs Target - HIGHEST_DRIVER_SEV

A.1.1.2 Number of Persons Involved in the Crash Analysis

The following table shows the distribution for N_PERSONS.

Table A.3: Number of Persons Involved in the Crash Distribution

Class	Frequency	Percentage
1 - 3	114683	63.59
3 - 5	46906	26.01
5 - 7	13149	7.29
7 - 9	3670	2.04
9 - 11	1113	0.62

11 - 13	396	0.22
13 - 15	166	0.09
15 - 17	79	0.04
17 - 19	42	0.02
19 - 21	29	0.02
21 - 23	28	0.02
23 - 25	8	0.0
25 - 27	10	0.01
27 - 29	8	0.0
29 - 31	3	0.0
31 - 33	8	0.0
33 - 35	3	0.0
35 - 37	2	0.0
37 - 39	6	0.0
41 - 43	2	0.0
43 - 45	2	0.0
45 - 47	5	0.0
47 - 49	1	0.0
49 - 51	4	0.0
51 - 53	3	0.0
53 - 55	2	0.0
55 - 57	2	0.0
63 - 65	1	0.0
65 - 67	2	0.0
69 - 71	1	0.0
145 - 147	1	0.0
157 - 159	1	0.0

The distribution is visualized as below:

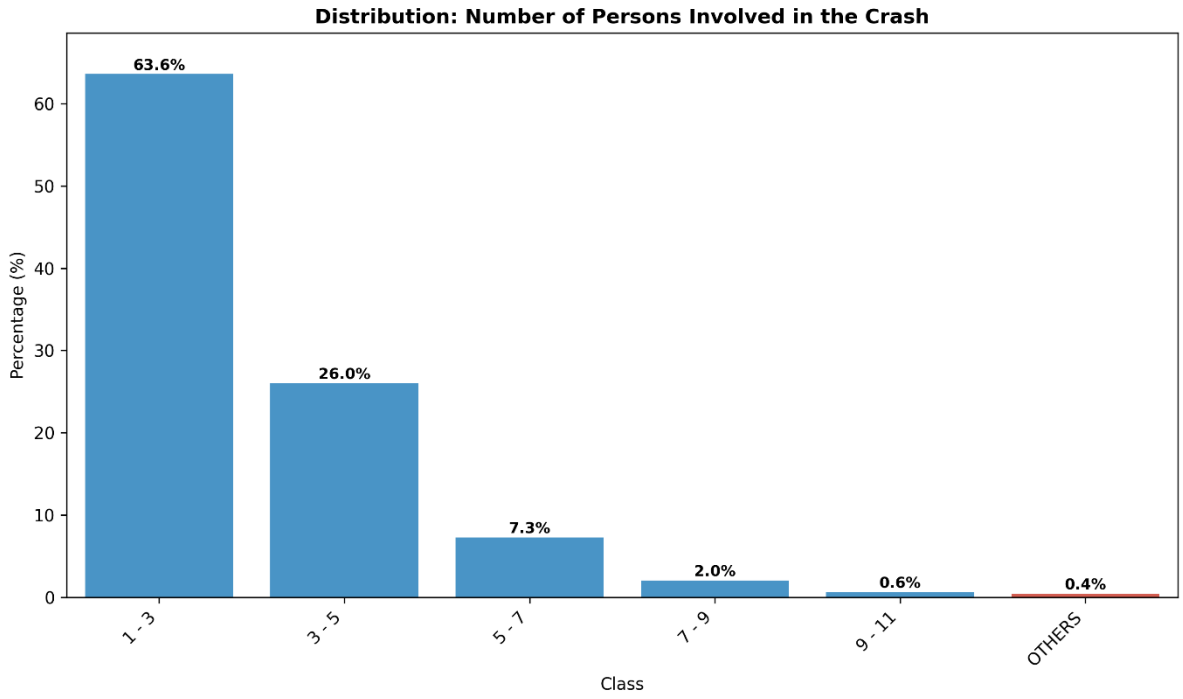

Figure A.4: Distribution Chart for Number of Persons Involved in the Crash

To impute the unknown values for N_PERSONS, statistical assessment was performed results of which are presented below:

Table A.4: Imputation Assessment: Number of Persons Involved in the Crash

Metric	Value
Count	180336
Min	1.0
Max	158.0
Mean	2.52
Median	2.0
Skewness	10.2331
Strategy	Median
Imputation Strategy	Median

The distribution after imputation is presented below:

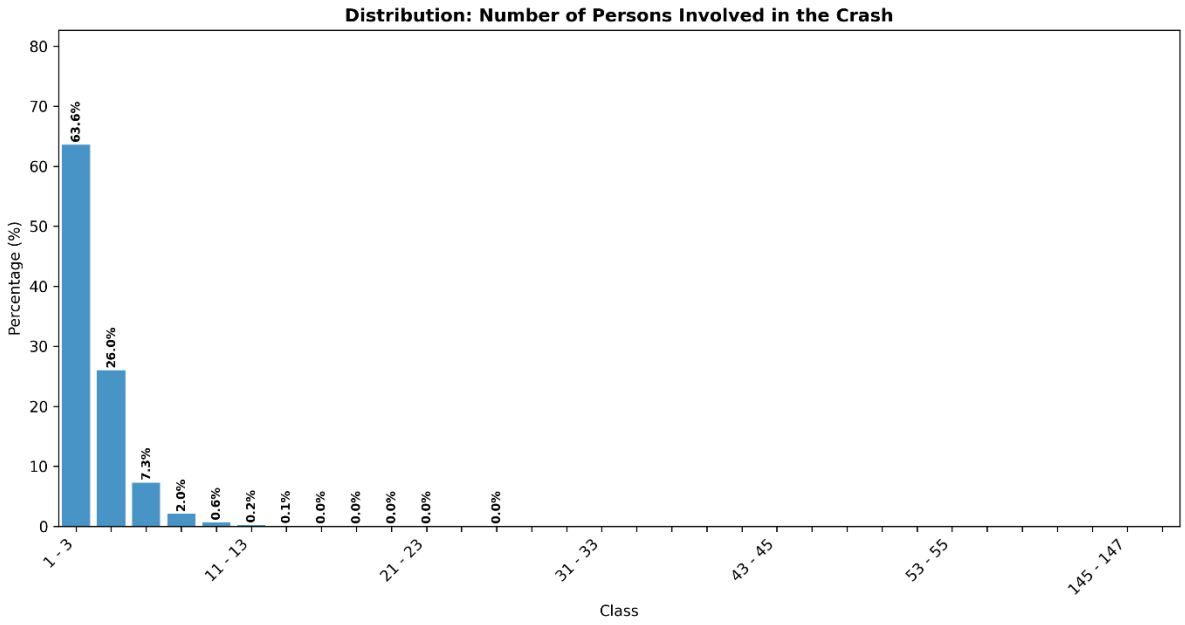

Figure A.5: Post Imputation Distribution Chart

The heatmap of the N_PERSONS is presented below:

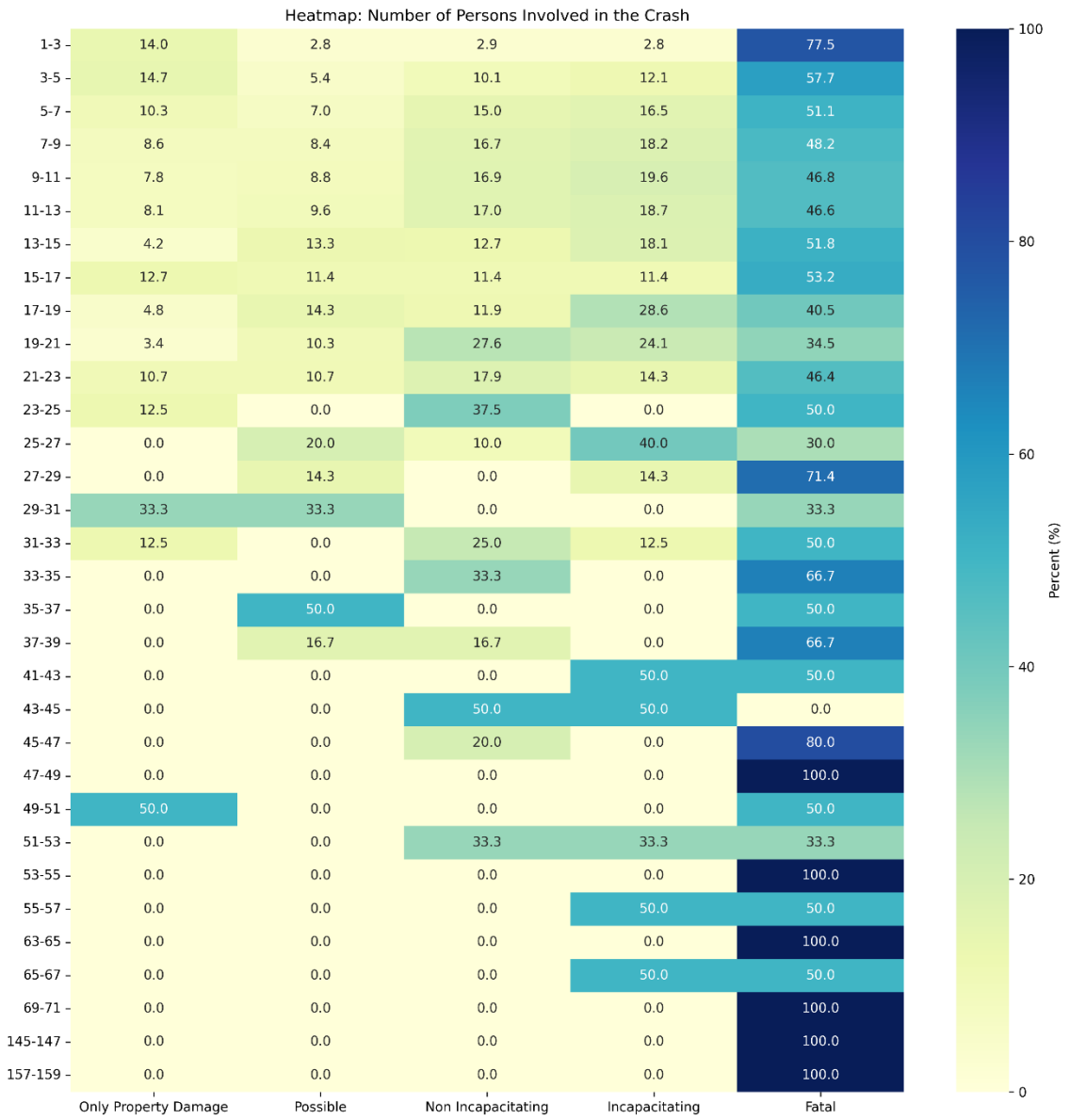

Figure A.6: Heatmap: Number of Persons Involved in the Crash vs Target - HIGHEST_DRIVER_SEV

A.1.1.3 Number of Pedestrians Involved in the Crash Analysis

The following table shows the distribution for PEDS.

Table A.5: Number of Pedestrians Involved in the Crash Distribution

Class	Frequency	Percentage
0 - 1	153323	85.02
1 - 2	25280	14.02
2 - 3	1374	0.76
3 - 4	229	0.13

4 - 5	79	0.04
5 - 6	20	0.01
6 - 7	13	0.01
7 - 8	5	0.0
8 - 9	4	0.0
9 - 10	3	0.0
10 - 11	1	0.0
11 - 12	2	0.0
12 - 13	1	0.0
17 - 18	1	0.0

The distribution is visualized as below:

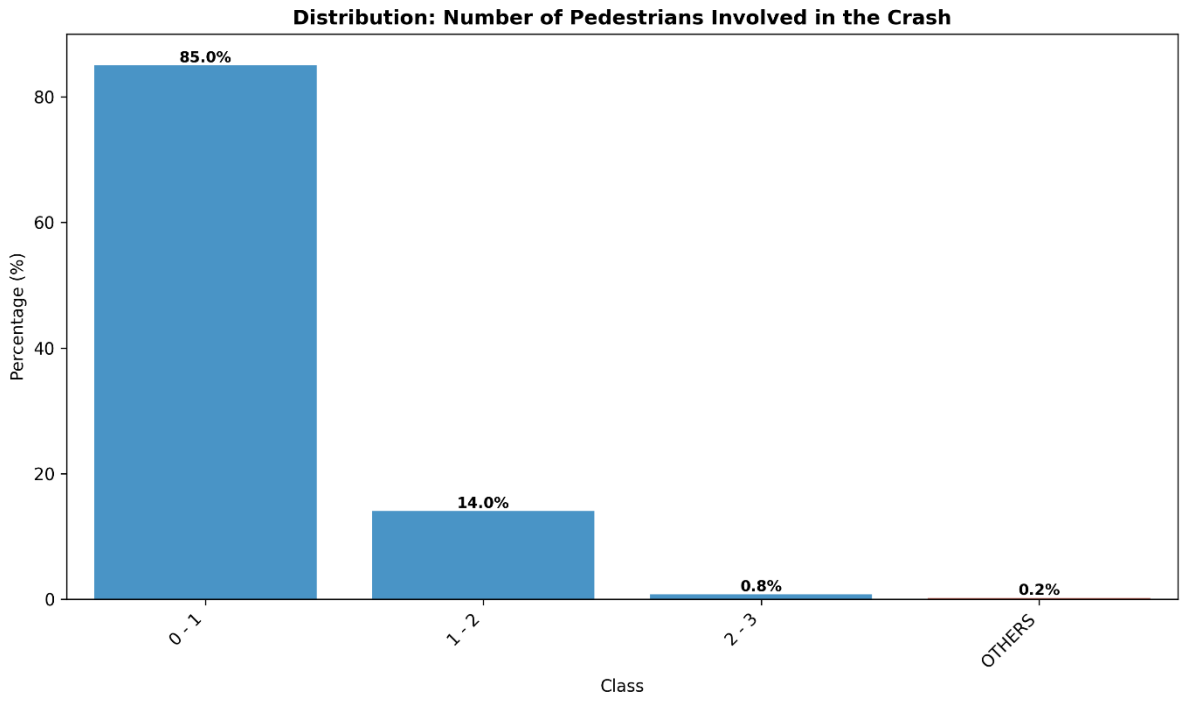

Figure A.7: Distribution Chart for Number of Pedestrians Involved in the Crash

To impute the unknown values for PEDS, statistical assessment was performed results of which are presented below:

Table A.6: Imputation Assessment: Number of Pedestrians Involved in the Crash

Metric	Value
Count	180336
Min	0.0
Max	22.0
Mean	0.16
Median	0.0
Skewness	4.6238
Strategy	Median
Imputation Strategy	Median

The distribution after imputation is presented below:

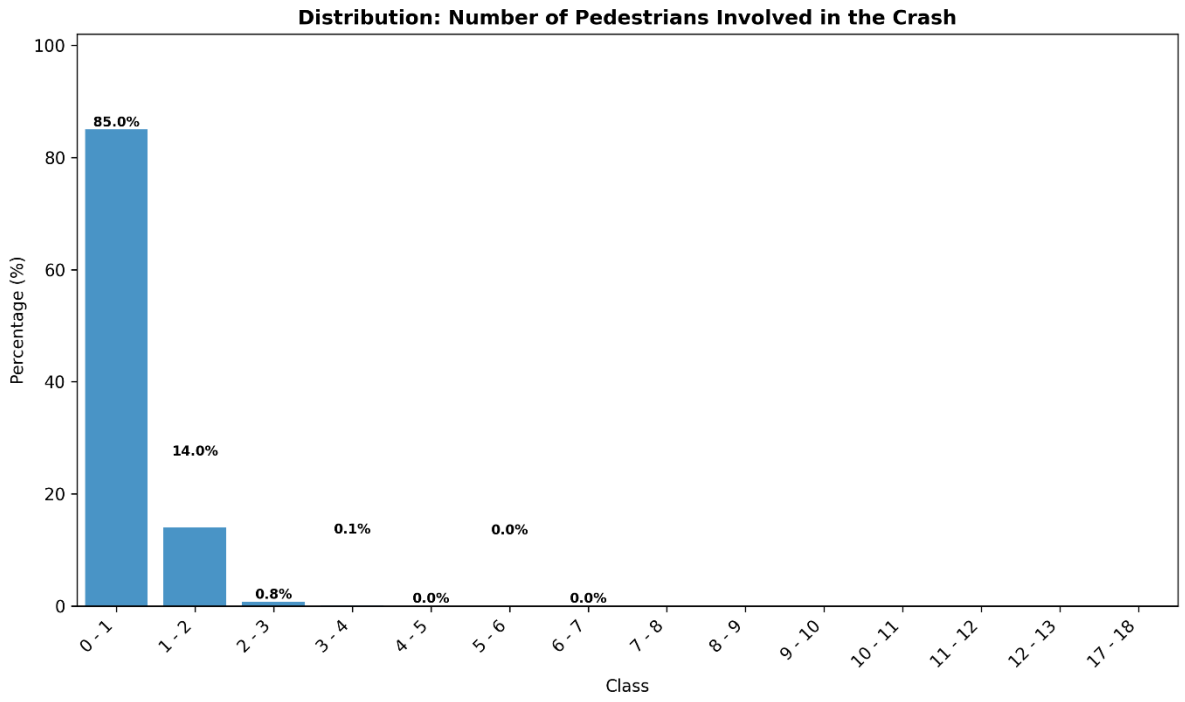

Figure A.8: Post Imputation Distribution Chart

The heatmap of the PEDS is presented below:

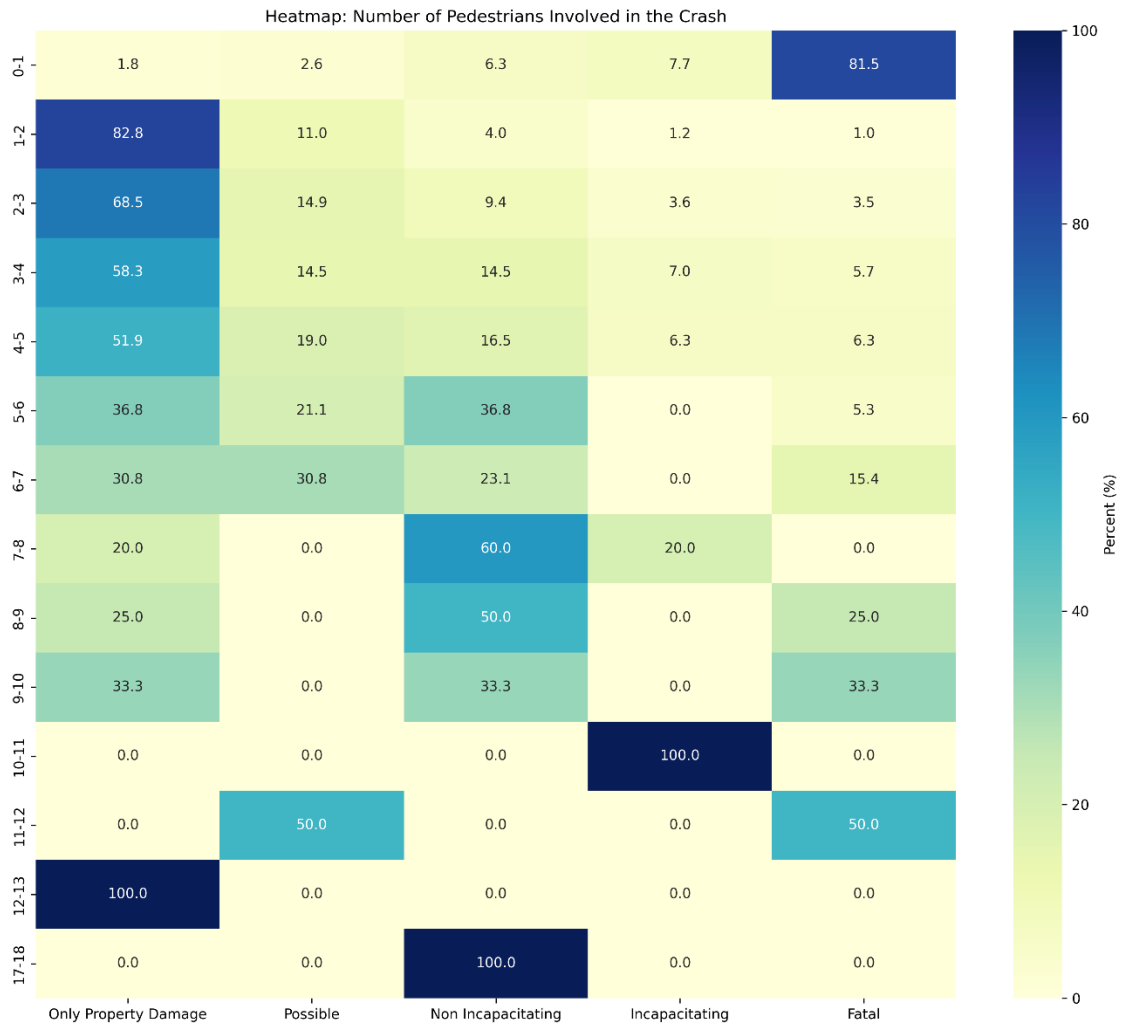

Figure A.9: Heatmap: Number of Pedestrians Involved in the Crash vs Target - HIGHEST_DRIVER_SEV

A.1.1.4 Road Functional Class Analysis

The following table shows the distribution for ROAD_FNC.

Table A.7: Road Functional Class Distribution

category	Code	Frequency	Percentage
Principal Arterial - Interstate - Rural	1	11242	6.23
Principal Arterial - Rural	2	20098	11.14
Minor Arterial-Rural	3	17142	9.51
Major Collector-Rural	4	24144	13.39
Minor Collector-Rural	5	6935	3.85
Local Road/Street-Rural	6	19227	10.66
Unknown-Rural	9	763	0.42

Principal Arterial	11	11182	6.2
Expressway	12	7089	3.93
Other Principal Arterial	13	22667	12.57
Minor Arterial	14	16111	8.93
Collector	15	6380	3.54
Street	16	16273	9.02
Unknown-Urban	19	207	0.11
Unknown	99	876	0.49

The distribution is visualized as below:

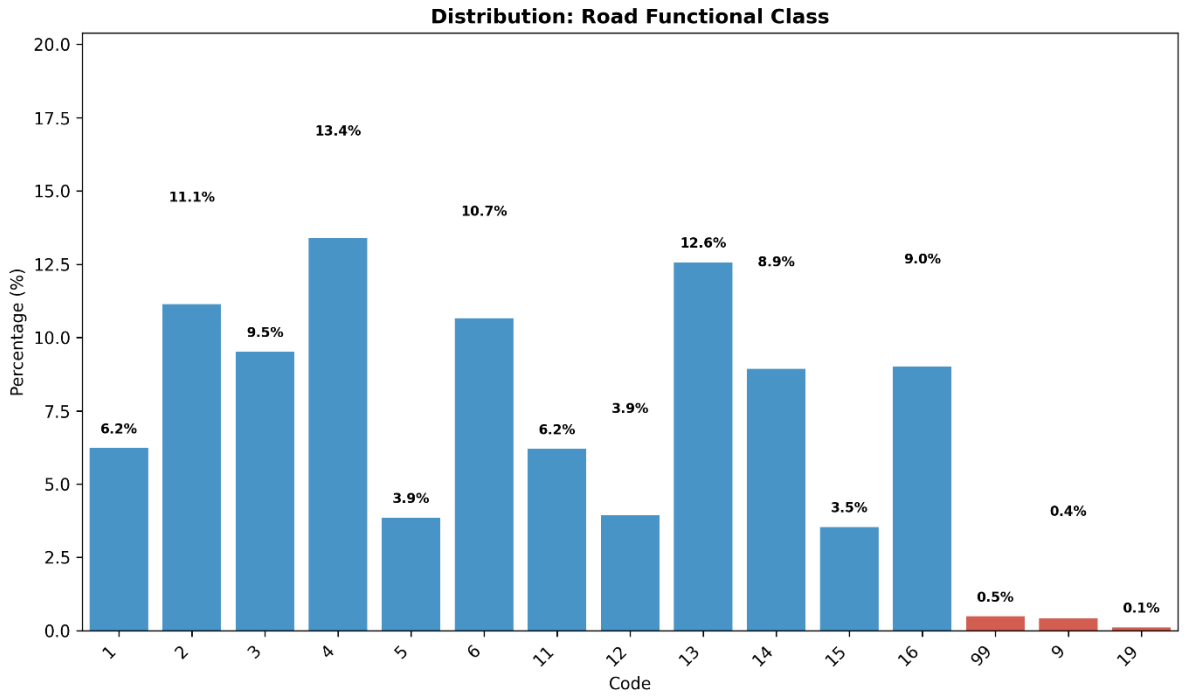

Figure A.10: Distribution Chart for Road Functional Class

To impute the unknown values for ROAD_FNC, statistical assessment was performed results of which are presented below:

Table A.8: Imputation Assessment: Road Functional Class

Metric	Value
Count	178490

Unique	12
Entropy	3.4627
Gini	0.9034
Mode	4.0
Mode_Freq	0.1353
Strategy	Ratio
Imputation Strategy	Ratio

The distribution after imputation is presented below:

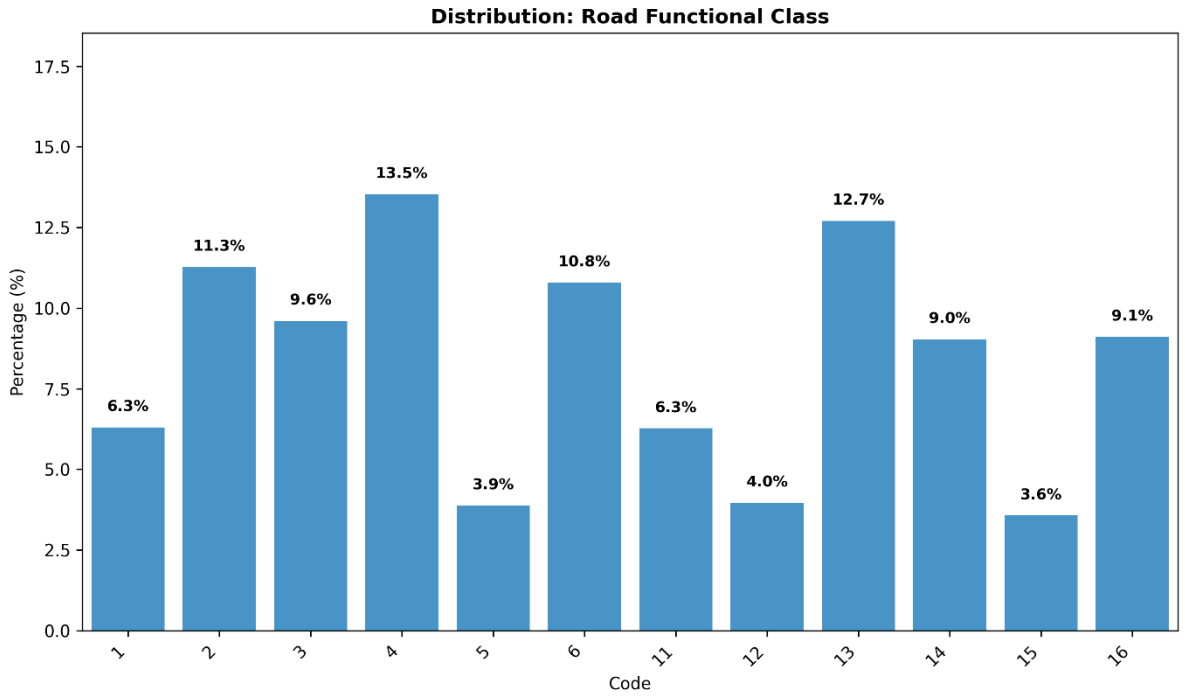

Figure A.11: Post Imputation Distribution Chart

The heatmap of the ROAD_FNC is presented below:

Figure A.12: Heatmap: Road Functional Class vs Target - HIGHEST_DRIVER_SEV

A.1.1.5 Junction Relation with Crash Location Analysis

The following table shows the distribution for REL_JUNC.

Table A.9: Junction Relation with Crash Location Distribution

category	Code	Frequency	Percentage
Non-Junction	1	130119	72.15
Intersection	2	31980	17.73
Intersection-Related	3	6606	3.66
Driveway, Alley Access, etc.	4	2075	1.15
Entrance/Exit Ramp-Related	5	872	0.48
Railway Grade Crossing	6	933	0.52
In Crossover	7	188	0.1
Driveway Access Related	8	2375	1.32
Unknown, Non-Interchange	9	156	0.09
Intersection-Interchange	10	788	0.44

Intersection-Related-Interchange	11	373	0.21
Driveway Access-Interchange	12	89	0.05
Entrance/Exit Ramp-Interchange	13	1865	1.03
In Crossover-Interchange	14	32	0.02
Other Location in Interchange	15	1700	0.94
Unknown, Interchange Area	19	28	0.02
Unknown	99	157	0.09

The distribution is visualized as below:

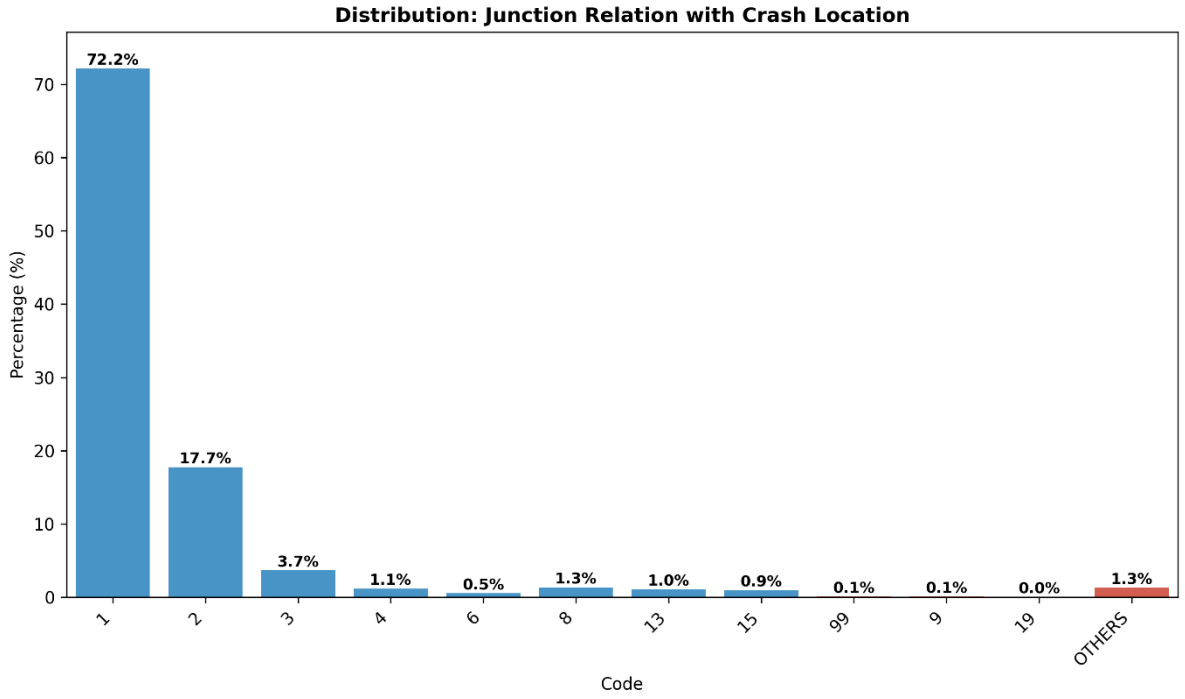

Figure A.13: Distribution Chart for Junction Relation with Crash Location

To impute the unknown values for REL_JUNC, statistical assessment was performed results of which are presented below:

Table A.10: Imputation Assessment: Junction Relation with Crash Location

Metric	Value
Count	179995
Unique	14

Entropy	1.3921
Gini	0.4439
Mode	1.0
Mode_Freq	0.7229
Strategy	Mode
Imputation Strategy	Mode

The distribution after imputation is presented below:

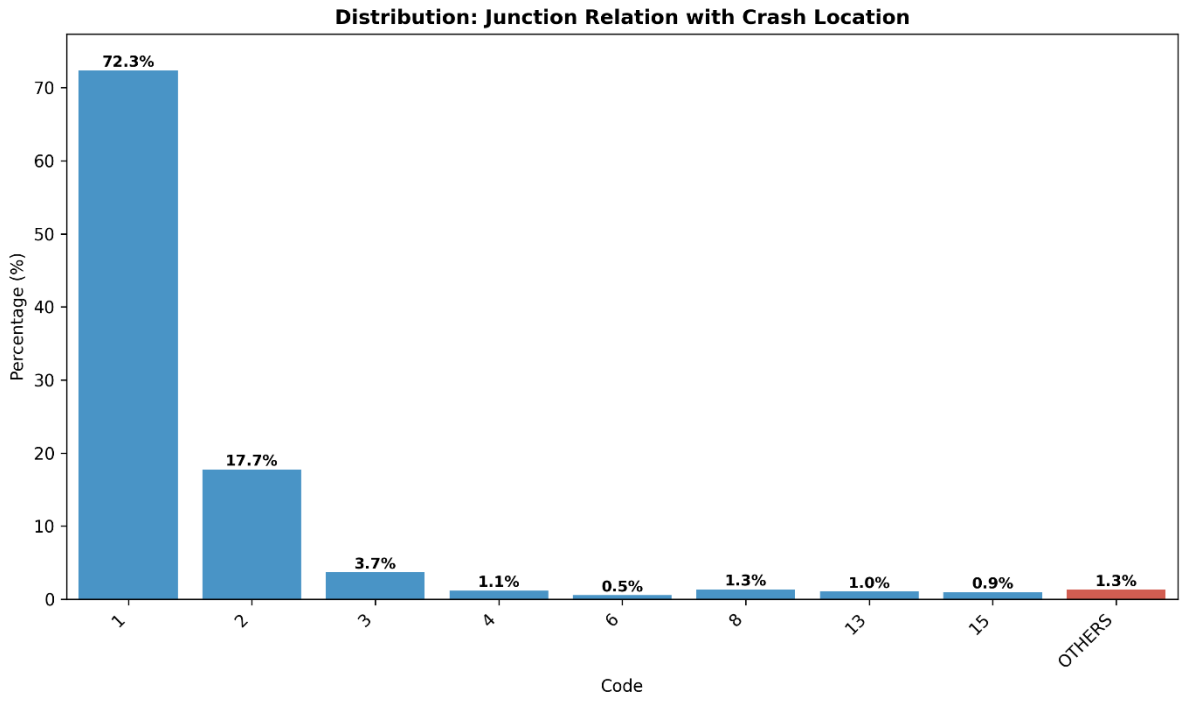

Figure A.14: Post Imputation Distribution Chart

The heatmap of the REL_JUNC is presented below:

Figure A.15: Heatmap: Junction Relation with Crash Location vs Target - HIGHEST_DRIVER_SEV

A.1.1.6 Vehicle Location Related to Roadway Analysis

The following table shows the distribution for REL_ROAD.

Table A.11: Vehicle Location Related to Roadway Distribution

category	Code	Frequency	Percentage
Non-Junction	1	100549	55.76
Intersection	2	12114	6.72
Intersection-Related	3	5566	3.09
Driveway, Alley Access, etc.	4	43621	24.19
Entrance/Exit Ramp-Related	5	6808	3.78
Railway Grade Crossing	6	9933	5.51
In Crossover	7	384	0.21

Driveway Access Related	8	613	0.34
Unknown, Non-Interchange	9	0	0.0
Intersection-Interchange	10	298	0.17
Intersection-Related-Interchange	11	68	0.04
Driveway Access-Interchange	12	0	0.0
Entrance/Exit Ramp-Interchange	13	0	0.0
In Crossover-Interchange	14	0	0.0
Other Location in Interchange	15	0	0.0
Unknown, Interchange Area	19	0	0.0
Unknown	99	382	0.21

The distribution is visualized as below:

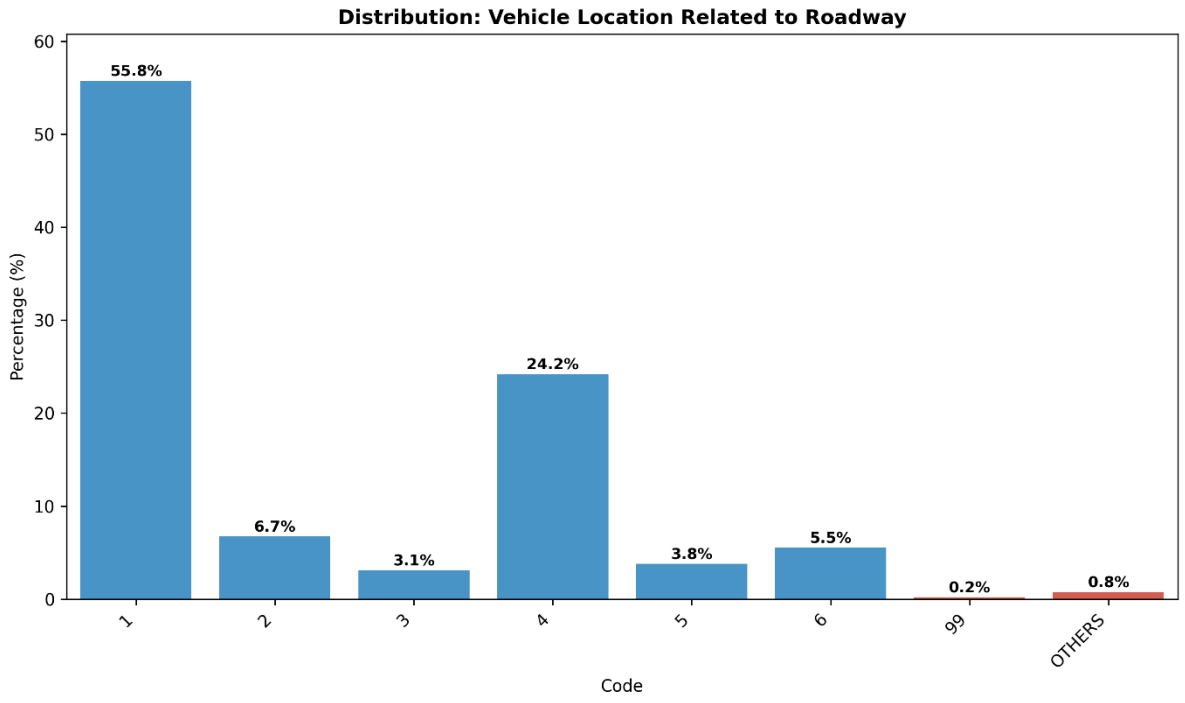

Figure A.16: Distribution Chart for Vehicle Location Related to Roadway

To impute the unknown values for REL_ROAD, statistical assessment was performed results of which are presented below:

Table A.12: Imputation Assessment: Vehicle Location Related to Roadway

Metric	Value
Count	179954
Unique	10
Entropy	1.8578
Gini	0.6191
Mode	1.0
Mode_Freq	0.5587
Strategy	Mode
Imputation Strategy	Mode

The distribution after imputation is presented below:

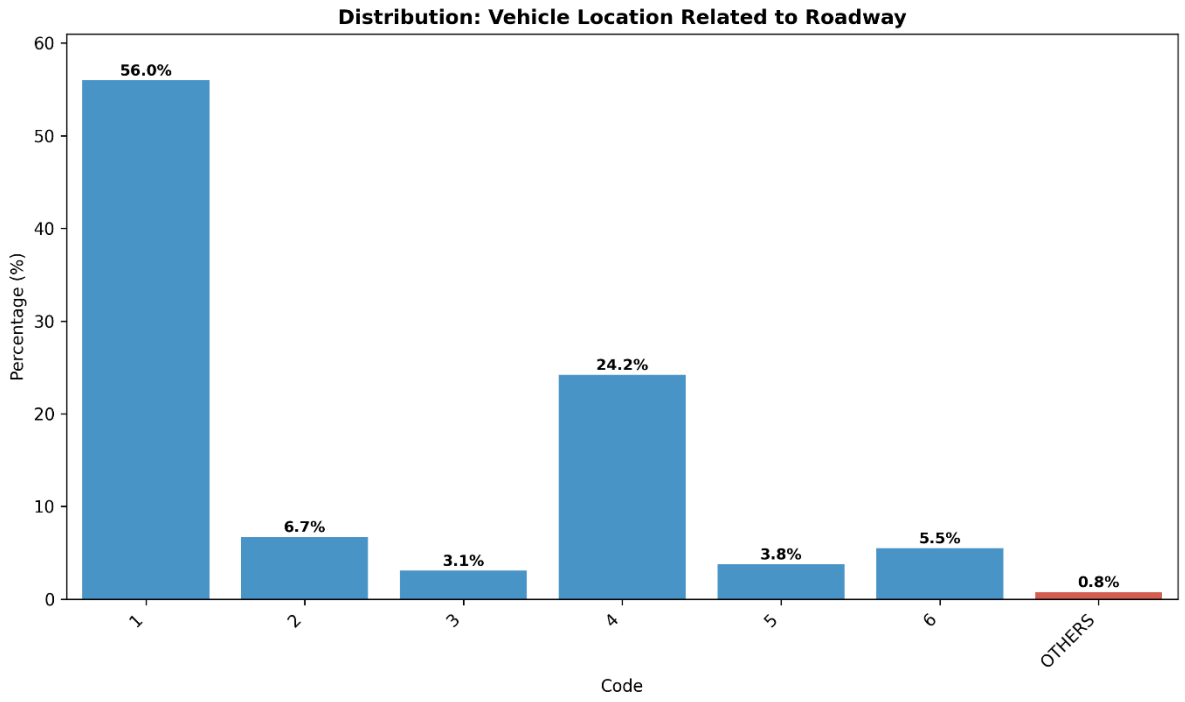

Figure A.17: Post Imputation Distribution Chart

The heatmap of the REL_ROAD is presented below:

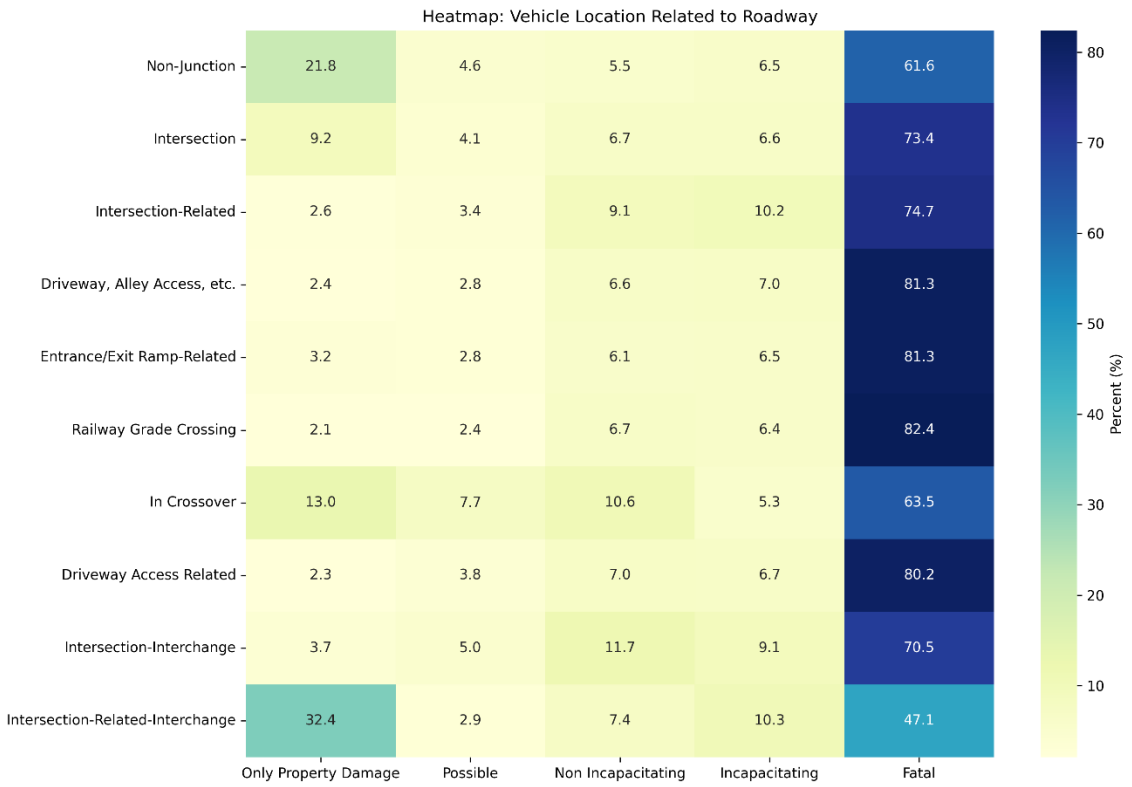

Figure A.18: Heatmap: Vehicle Location Related to Roadway vs Target - HIGHEST_DRIVER_SEV

A.1.1.7 Speed Limit Analysis

The following table shows the distribution for SP_LIMIT.

Table A.13: Speed Limit Distribution

Class	Frequency	Percentage
0 - 10	615	0.34
10 - 20	394	0.22
20 - 30	8574	4.75
30 - 40	29540	16.38
40 - 50	38544	21.37
50 - 60	61259	33.97
60 - 70	24548	13.61
70 - 80	12140	6.73
80 - 90	59	0.03
90 - 100	4663	2.59

The distribution is visualized as below:

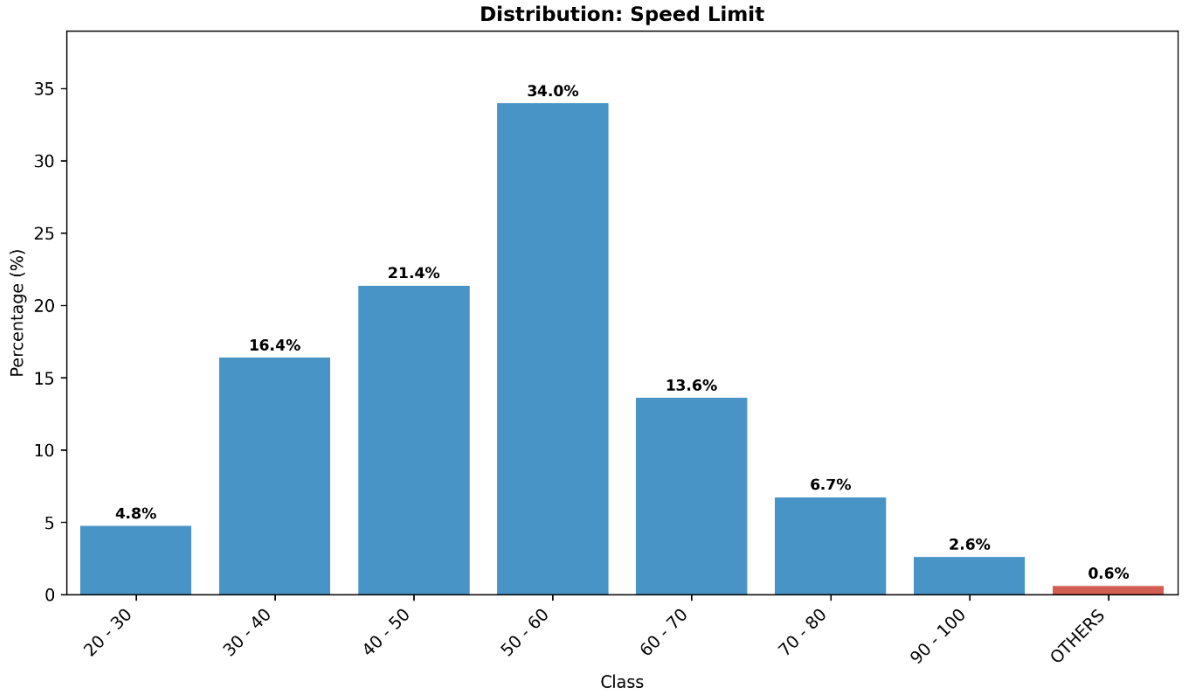

Figure A.19: Distribution Chart for Speed Limit

To impute the unknown values for SP_LIMIT, statistical assessment was performed results of which are presented below:

Table A.14: Imputation Assessment: Speed Limit

Metric	Value
Count	175685
Min	0.0
Max	95.0
Mean	49.14
Median	55.0
Skewness	-0.3399
Strategy	Mean
Imputation Strategy	Mean

The distribution after imputation is presented below:

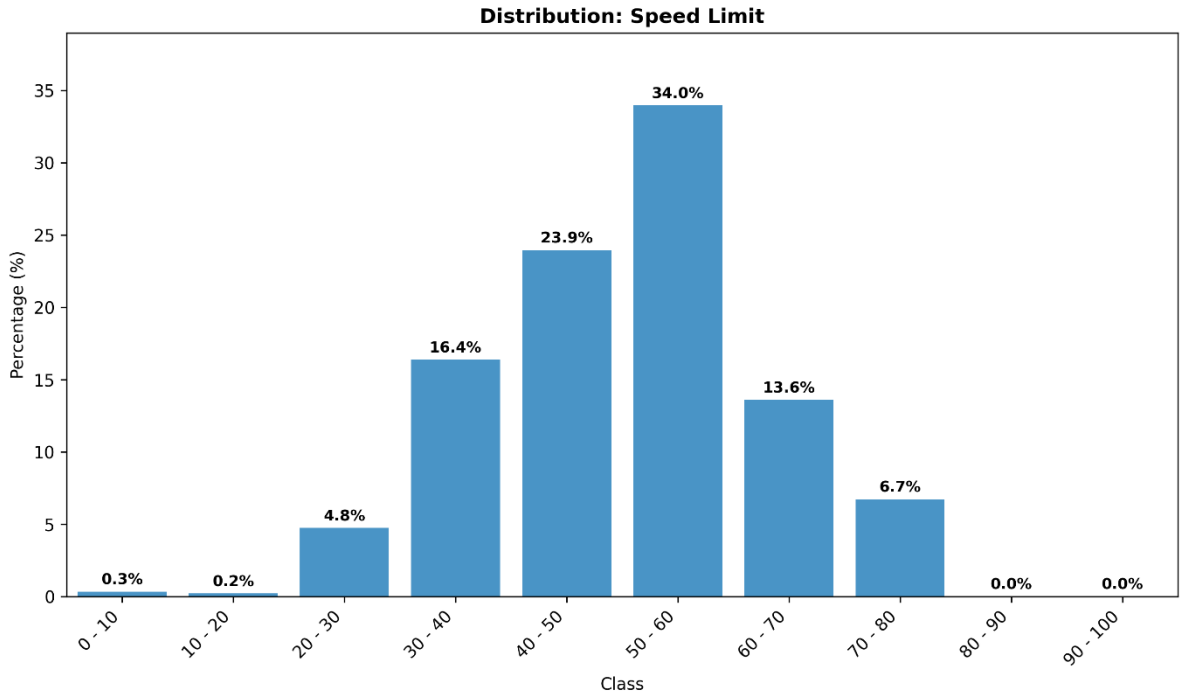

Figure A.20: Post Imputation Distribution Chart

The heatmap of the SP_LIMIT is presented below:

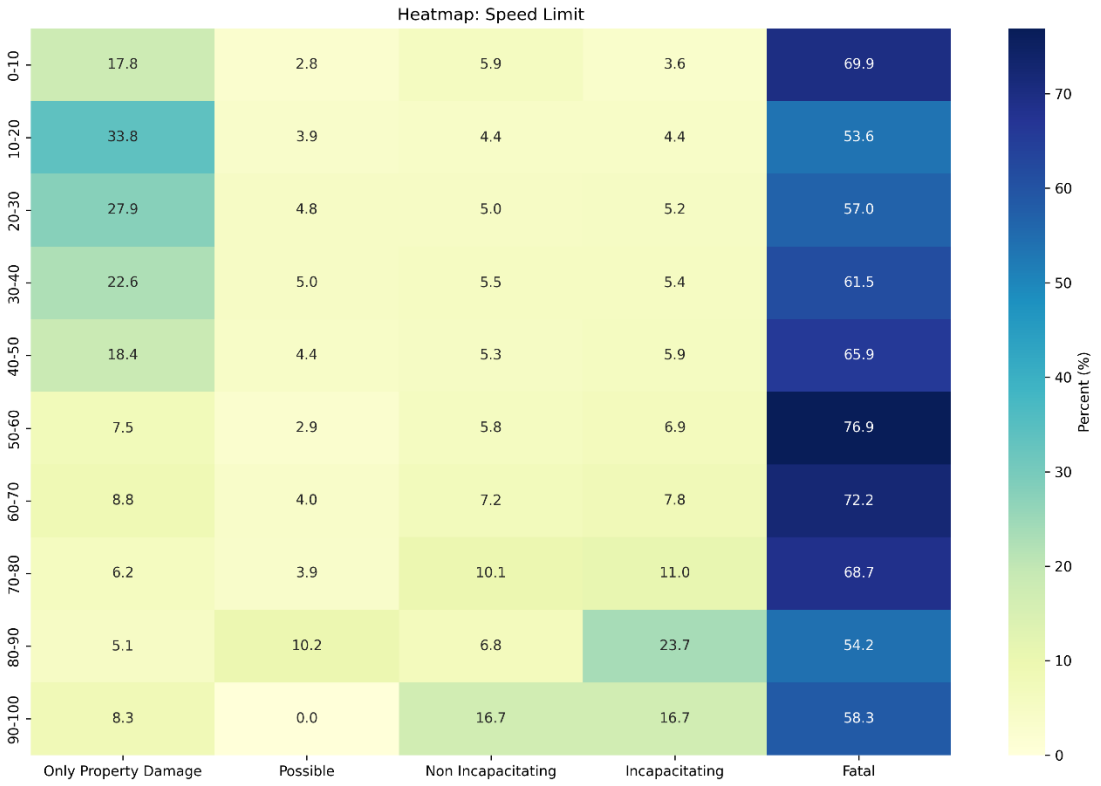

Figure A.21: Heatmap: Speed Limit vs Target - HIGHEST_DRIVER_SEV

A.1.1.8 Road Alignment Analysis

The following table shows the distribution for ALIGNMNT.

Table A.15: Road Alignment Distribution

category	Code	Frequency	Percentage
Non-Trafficway Area	0	0	0.0
Straight	1	130325	72.27
Curved	2	48966	27.15
Unknown	9	1045	0.58

The distribution is visualized as below:

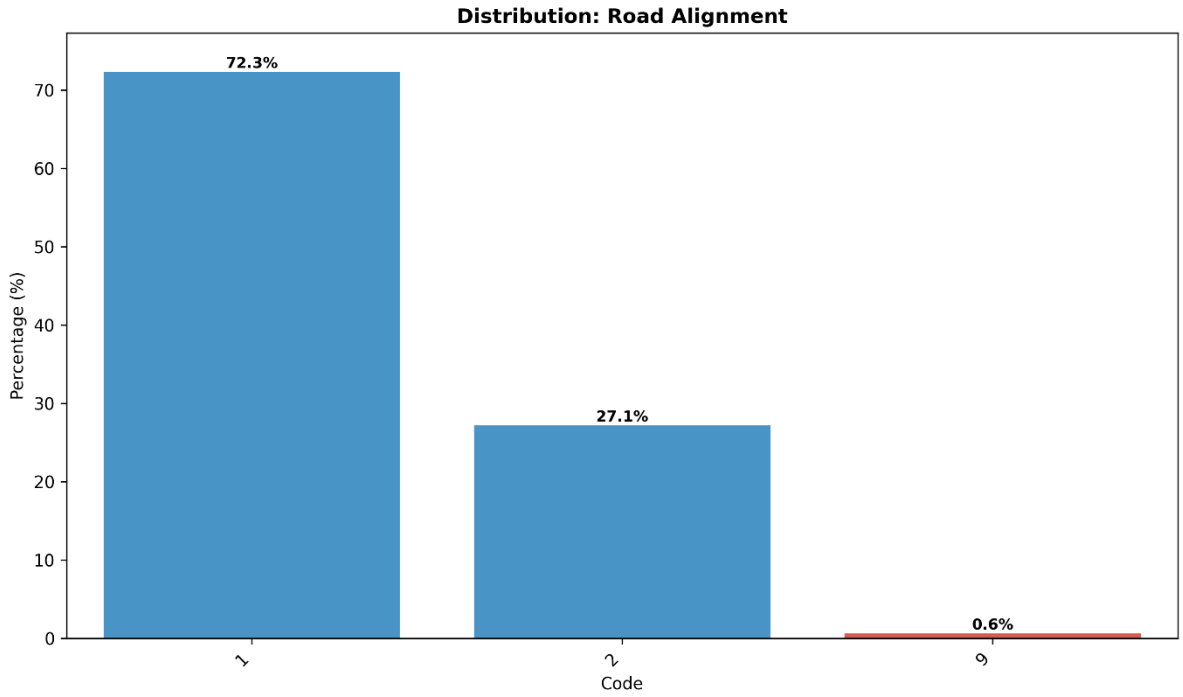

Figure A.22: Distribution Chart for Road Alignment

To impute the unknown values for ALIGNMNT, statistical assessment was performed results of which are presented below:

Table A.16: Imputation Assessment: Road Alignment

Metric	Value
Count	179291

Unique	2
Entropy	0.8459
Gini	0.397
Mode	1.0
Mode_Freq	0.7269
Strategy	Mode
Imputation Strategy	Mode

The distribution after imputation is presented below:

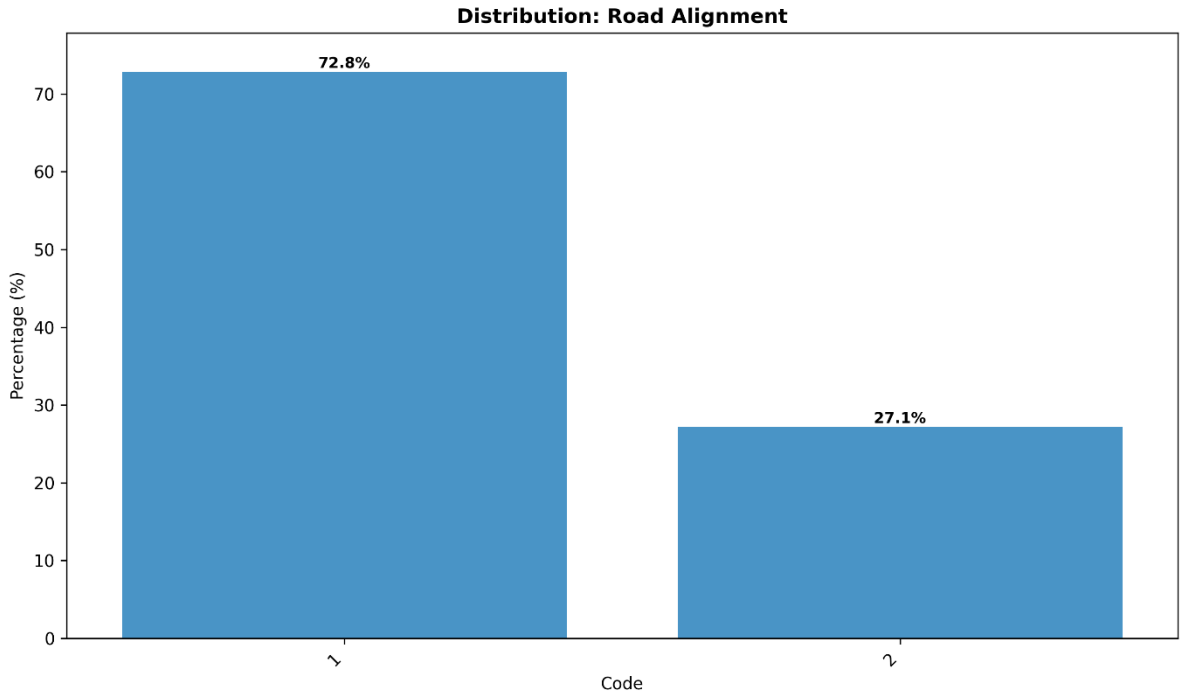

Figure A.23: Post Imputation Distribution Chart

The heatmap of the ALIGNMNT is presented below:

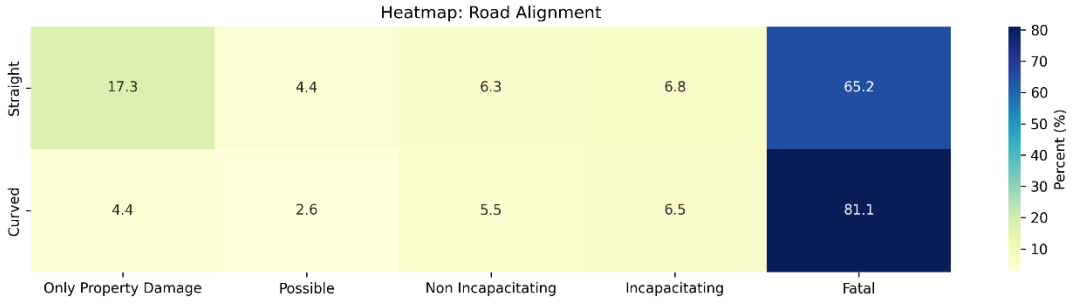

Figure A.24: Heatmap: Road Alignment vs Target - HIGHEST_DRIVER_SEV

A.1.1.9 Road Profile Analysis

The following table shows the distribution for PROFILE.

Table A.17: Road Profile Distribution

category	Code	Frequency	Percentage
Non-Trafficway Area	0	0	0.0
Level	1	128038	71.0
Grade	2	44134	24.47
Hillcrest	3	4467	2.48
Sag	4	574	0.32
Uphill	5	0	0.0
Downhill	6	0	0.0
Not Reported	8	0	0.0
Unknown	9	3123	1.73

The distribution is visualized as below:

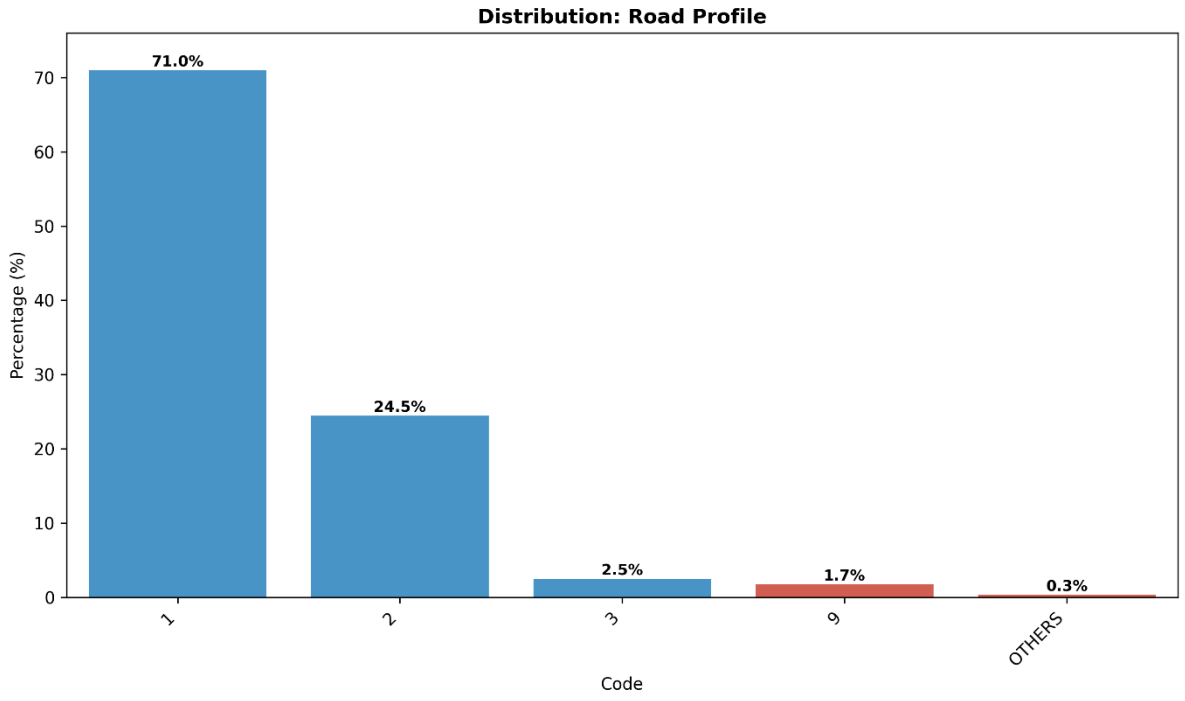

Figure A.25: Distribution Chart for Road Profile

To impute the unknown values for PROFILE, statistical assessment was performed results of which are presented below:

Table A.18: Imputation Assessment: Road Profile

Metric	Value
Count	177213
Unique	4
Entropy	0.9989
Gini	0.4153
Mode	1.0
Mode_Freq	0.7225
Strategy	Mode
Imputation Strategy	Mode

The distribution after imputation is presented below:

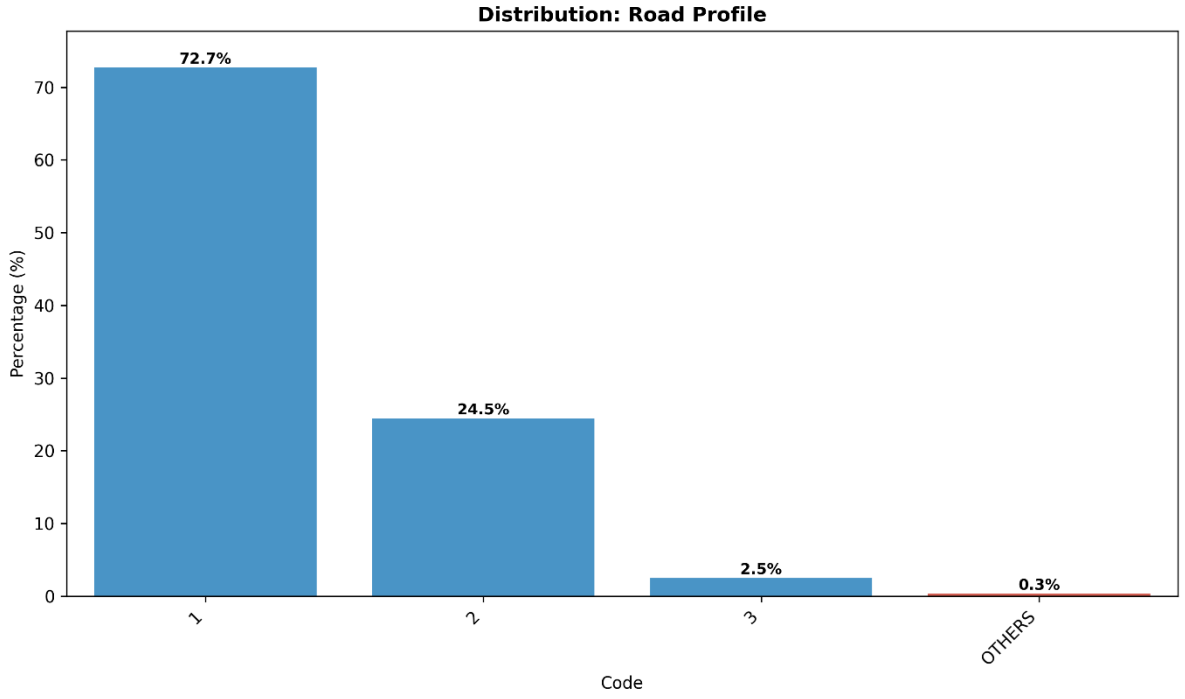

Figure A.26: Post Imputation Distribution Chart

The heatmap of the PROFILE is presented below:

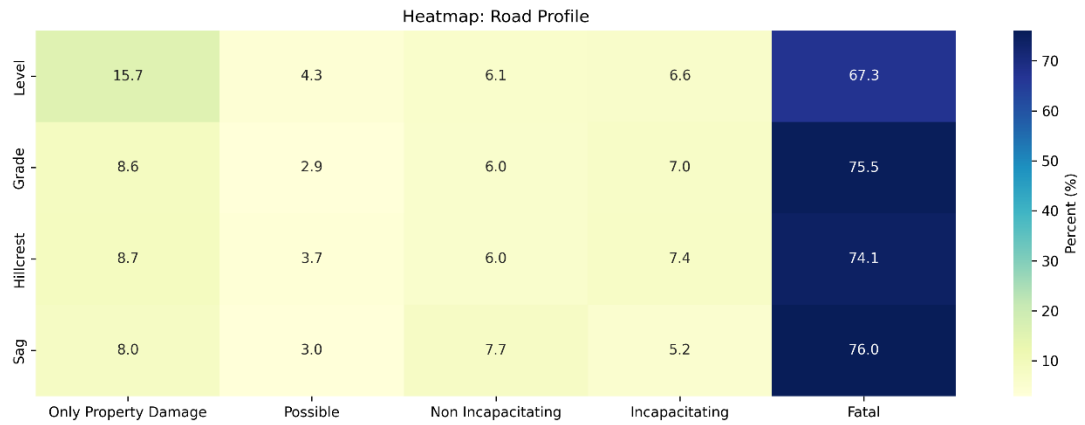

Figure A.27: Heatmap: Road Profile vs Target - HIGHEST_DRIVER_SEV

A.1.1.10 Pavement Type Analysis

The following table shows the distribution for PAVE_TYP.

Table A.19: Pavement Type Distribution

category	Code	Frequency	Percentage
Concrete	1	15731	8.72
Blacktop, Bituminous, or Asphalt	2	157344	87.25
Brick or Block	3	76	0.04
Slag, Gravel or Stone	4	2490	1.38
Dirt	5	1369	0.76
Other	7	0	0.0
Other/Not Reported	8	240	0.13
Unknown	9	3086	1.71

The distribution is visualized as below:

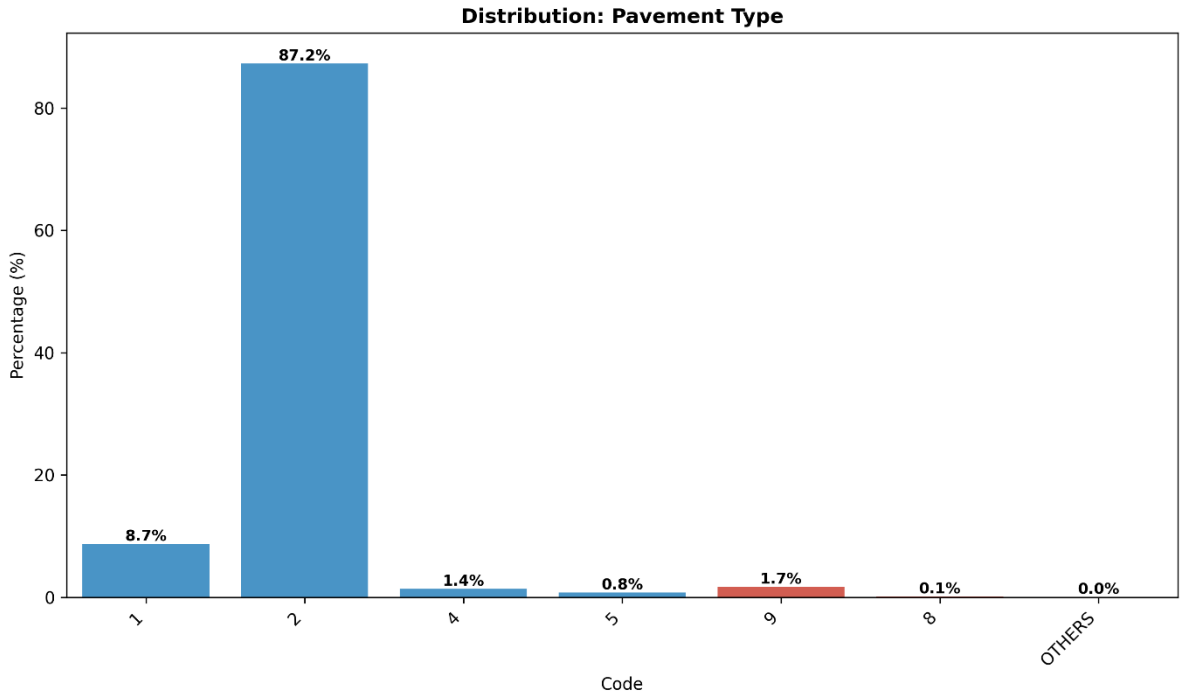

Figure A.28: Distribution Chart for Pavement Type

To impute the unknown values for PAVE_TYP, statistical assessment was performed results of which are presented below:

Table A.20: Imputation Assessment: Pavement Type

Metric	Value
Count	177010
Unique	5
Entropy	0.607
Gini	0.2017
Mode	2.0
Mode_Freq	0.8889
Strategy	Mode
Imputation Strategy	Mode

The distribution after imputation is presented below:

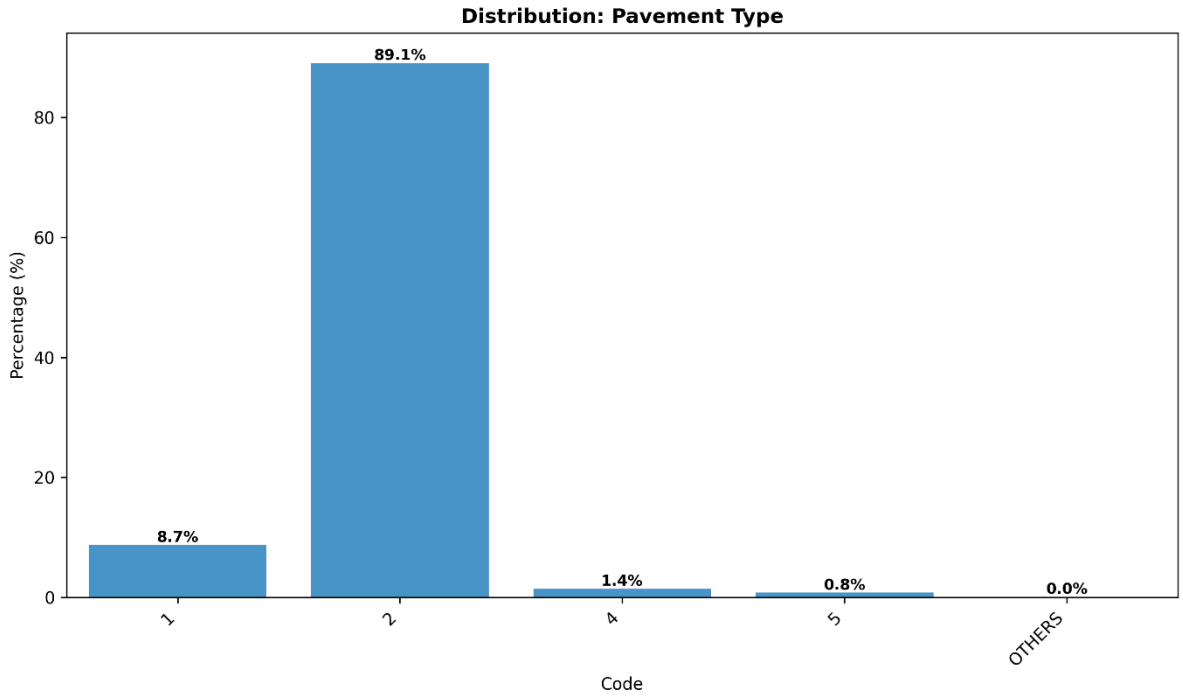

Figure A.29: Post Imputation Distribution Chart

The heatmap of the PAVE_TYP is presented below:

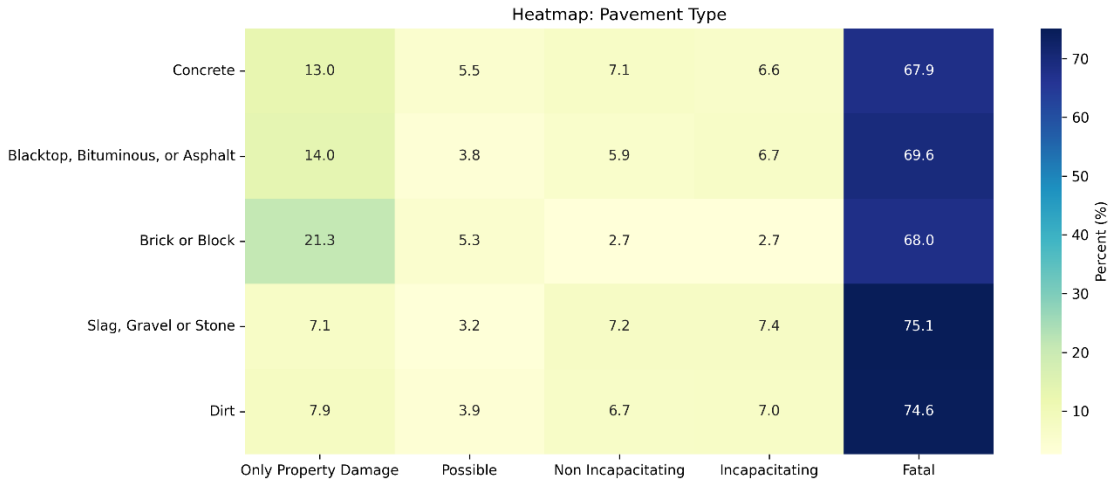

Figure A.30: Heatmap: Pavement Type vs Target - HIGHEST_DRIVER_SEV

A.1.1.11 Roadway Surface Condition Analysis

The following table shows the distribution for SUR_COND.

Table A.21: Roadway Surface Condition Distribution

category	Code	Frequency	Percentage
Dry	1	151288	83.89

Wet	2	21755	12.06
Snow or Slush	3	2598	1.44
Ice/Frost	4	2744	1.52
Sand, Dirt, Mud, Gravel	5	393	0.22
Water (Standing or Moving)	6	131	0.07
Oil	7	13	0.01
Other	8	213	0.12
Unknown	9	1201	0.67
Slush	10	0	0.0
Mud	11	0	0.0
Not Reported	98	0	0.0
Unknown	99	0	0.0

The distribution is visualized as below:

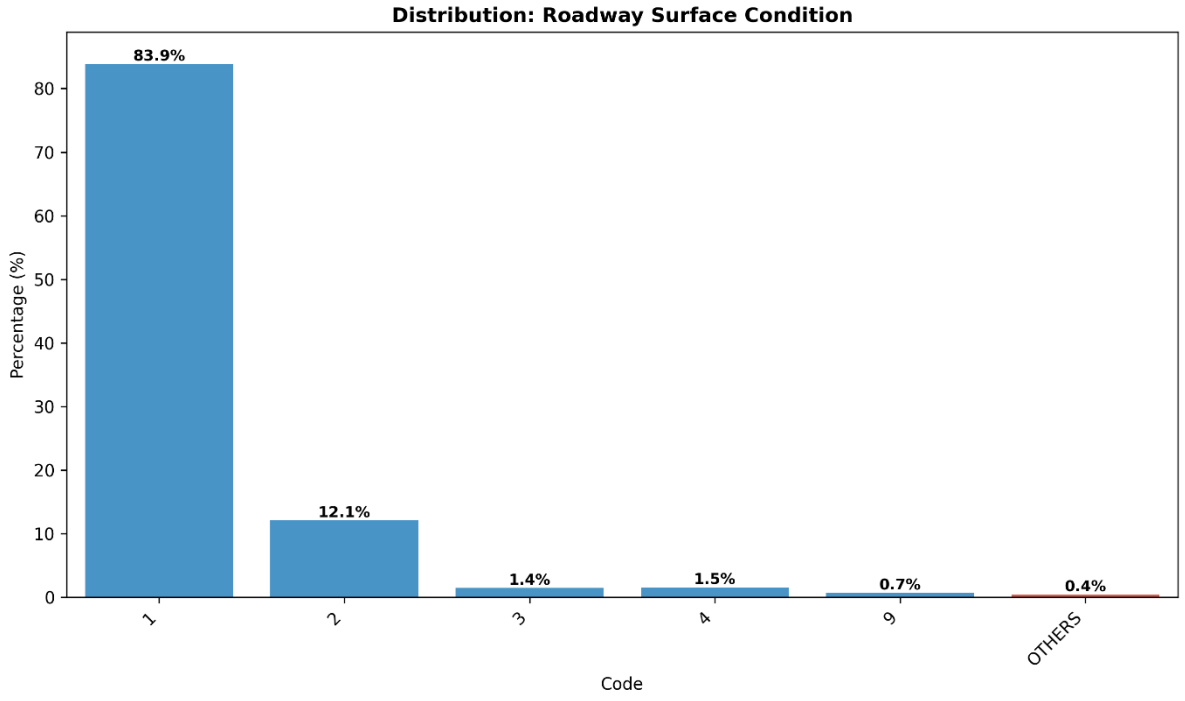

Figure A.31: Distribution Chart for Roadway Surface Condition

To impute the unknown values for SUR_COND, statistical assessment was performed results of which are presented below:

Table A.22: Imputation Assessment: Roadway Surface Condition

Metric	Value
Count	180336
Unique	9
Entropy	0.8481
Gini	0.2812
Mode	1.0
Mode_Freq	0.8389
Strategy	Mode
Imputation Strategy	Mode

The distribution after imputation is presented below:

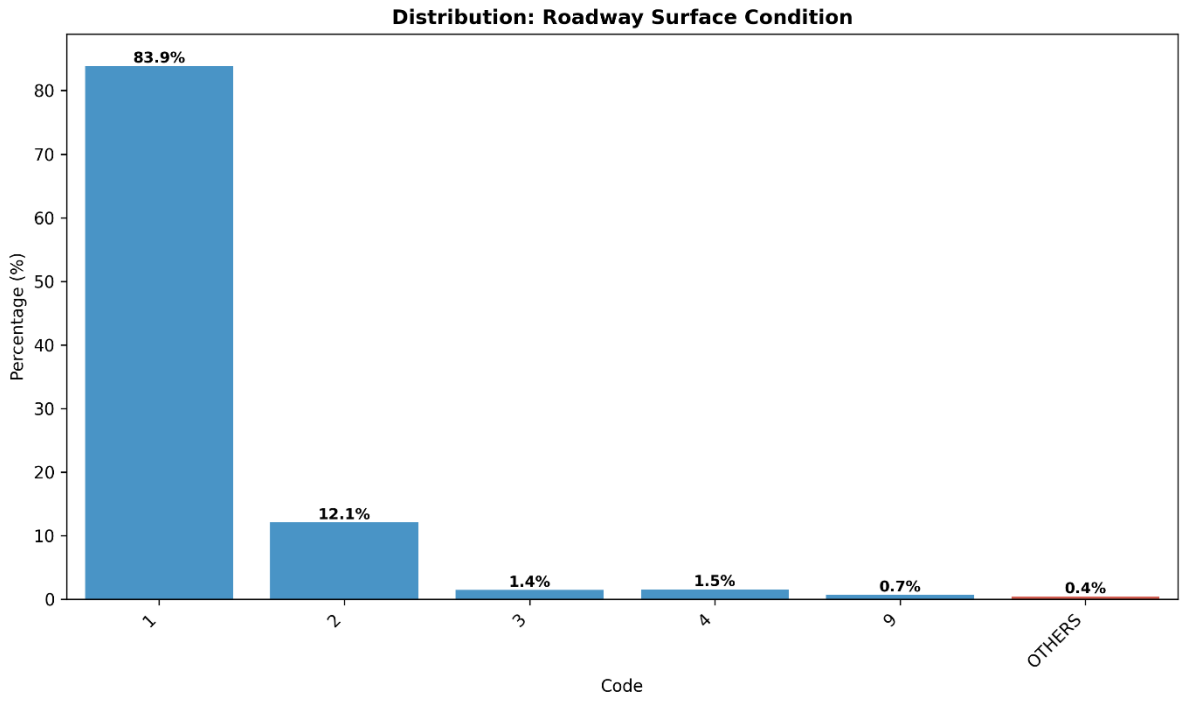

Figure A.32: Post Imputation Distribution Chart

The heatmap of the SUR_COND is presented below:

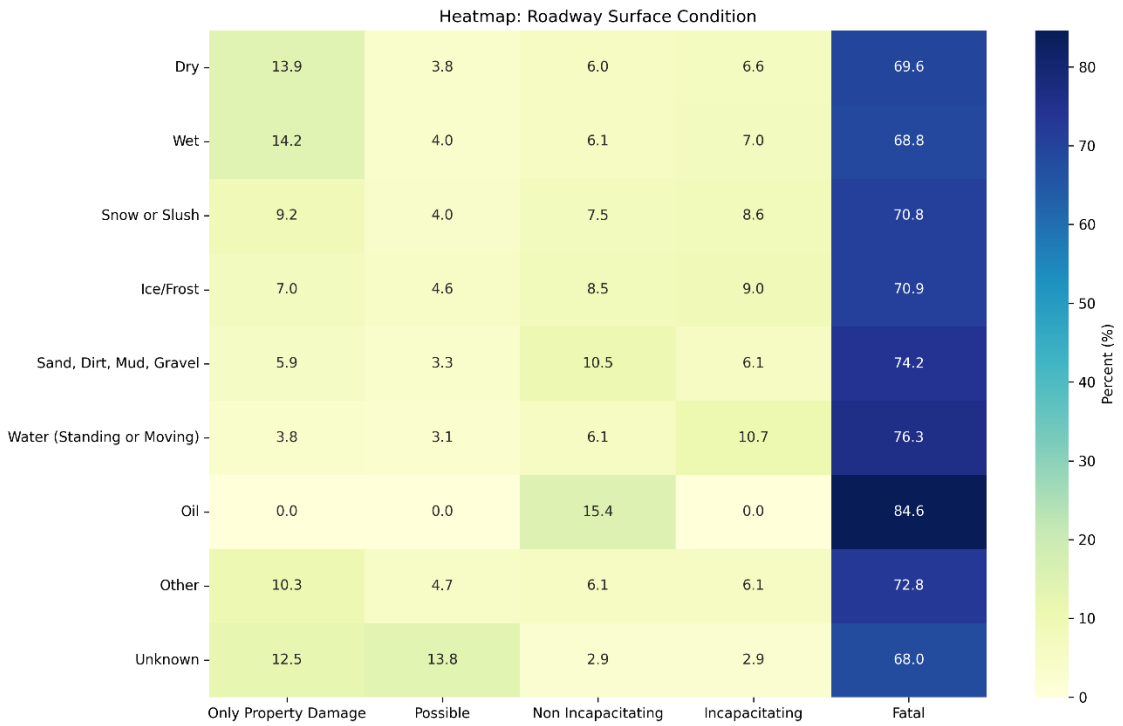

Figure A.33: Heatmap: Roadway Surface Condition vs Target - HIGHEST_DRIVER_SEV

A.1.1.12 Manner of Collision Analysis

The following table shows the distribution for MAN_COLL.

Table A.23: Manner of Collision Distribution

category	Code	Frequency	Percentage
Not Collision with Motor Vehicle In-Transport	0	109876	60.93
Front-to-Rear	1	11087	6.15
Front-to-Front	2	17989	9.98
Angle - Front-to-Side, Same Direction	3	2297	1.27
Angle - Front-to-Side, Opposite Direction	4	9113	5.05
Angle - Front-to-Side, Right Angle (Includes Broadside)	5	23382	12.97
Angle - Front-to-Side/Angle-Direction Not Specified	6	1006	0.56
Sideswipe - Same Direction	7	2406	1.33
Sideswipe - Opposite Direction	8	2128	1.18
Rear-to-Side	9	338	0.19
Rear-to-Rear	10	13	0.01
Other (End-Swipes and Others)	11	379	0.21
Not Reported	98	0	0.0

Reported as Unknown	99	322	0.18

The distribution is visualized as below:

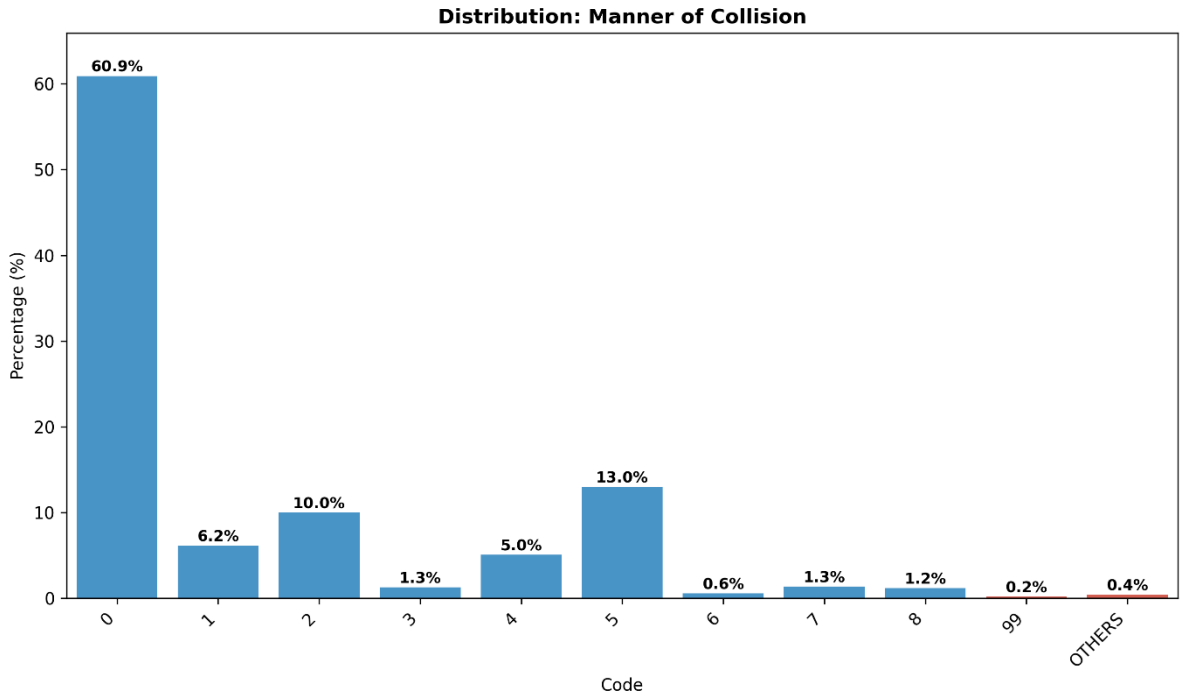

Figure A.34: Distribution Chart for Manner of Collision

To impute the unknown values for MAN_COLL, statistical assessment was performed results of which are presented below:

Table A.24: Imputation Assessment: Manner of Collision

Metric	Value
Count	180014
Unique	12
Entropy	1.9325
Gini	0.5937
Mode	0.0
Mode_Freq	0.6104
Strategy	Mode
Imputation Strategy	Mode

The distribution after imputation is presented below:

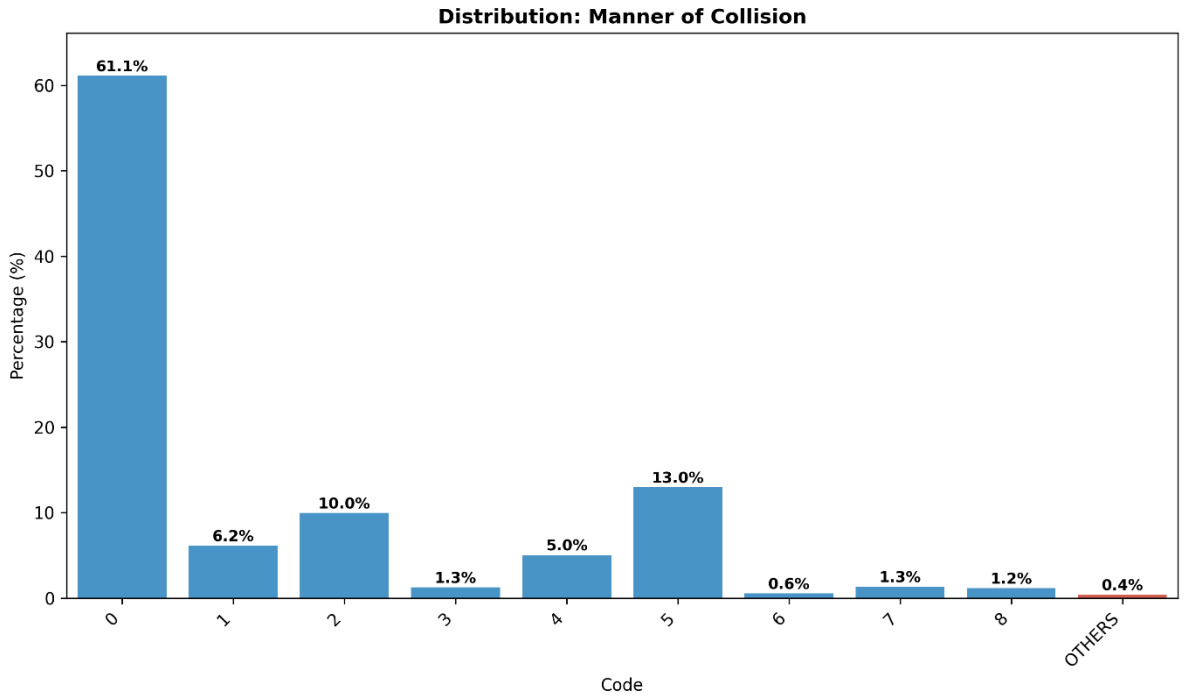

Figure A.35: Post Imputation Distribution Chart

The heatmap of the MAN_COLL is presented below:

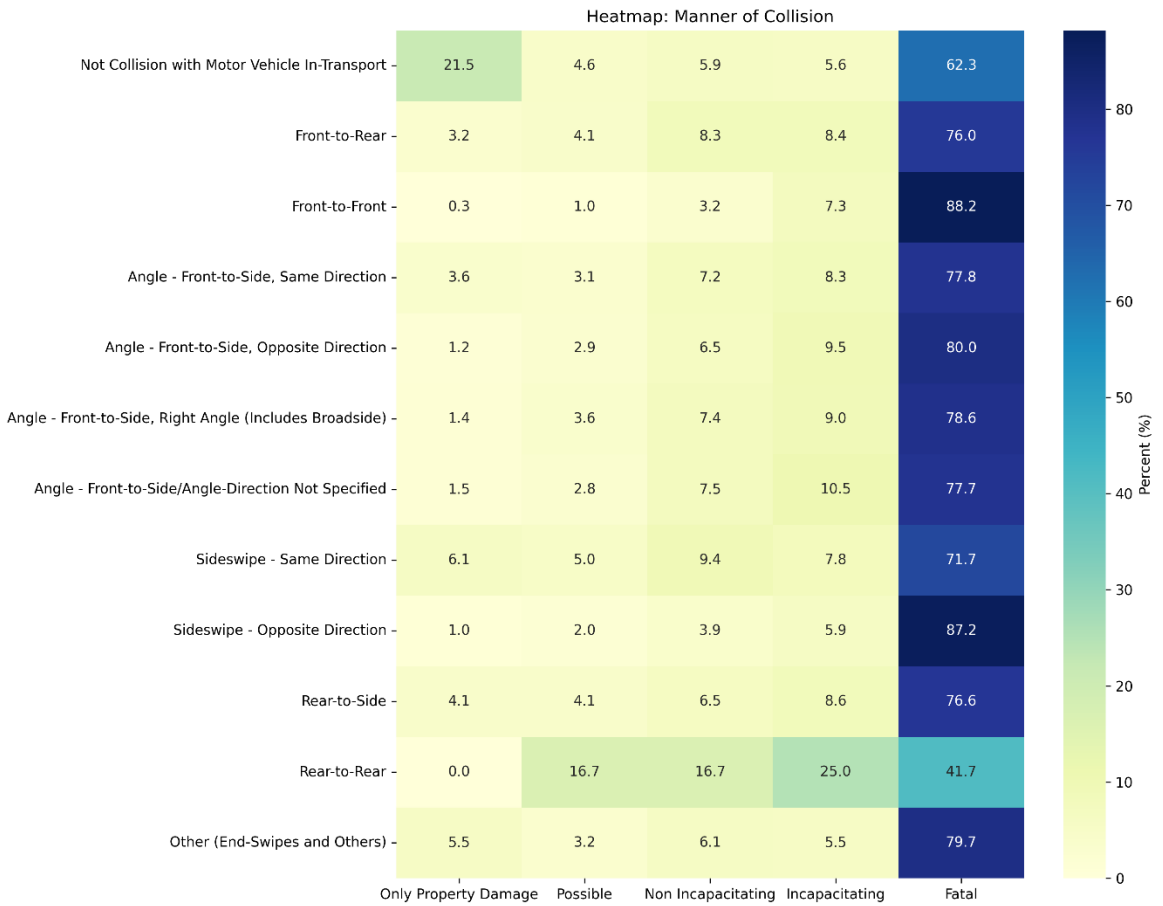

Figure A.36: Heatmap: Manner of Collision vs Target - HIGHEST_DRIVER_SEV

A.1.2 Descriptive Statistical Analysis of Dataset

For a deeper understanding of the underlying distribution and central tendencies of any large-scale observational dataset, the descriptive statistical analysis serves as the foundational step. It systematically evaluates individual variables, such as the exact MINUTE of an occurrence. Researchers identify temporal patterns, anomalies and data quality before advancing to complex modeling. Thus, producing comprehensive frequency tables is a critical initial procedure. The statistical analysis tables report the frequency, which quantifies the raw absolute count of instances recorded for each specific minute, alongside the total percentage, which represents this count as a fraction of the entire dataset, including any missing or unrecorded entries. Importantly, the researcher must isolate the valid percentage, which is unlike the total percentage. The valid percentage adjusts the computational baseline by excluding the missing values or corrupted data points, hence providing a mathematically accurate reflection of the variable's true distribution within the observable data. Descriptive statistical analysis is performed after imputing the missing or unknown values to ensure the proper imputation and

assess the data quality. The descriptive statistical analysis of each feature was performed and is presented feature-wise.

A.1.2.1 Descriptive Statistics of Minute of Crash

Table A.25: Minute of Crash Distribution

Code	Label	Frequency	Pct_Total	Pct_Valid
0.0	0.0	11040	6.12%	6.12%
30.0	30.0	10516	5.83%	5.83%
45.0	45.0	7571	4.2%	4.2%
15.0	15.0	7465	4.14%	4.14%
20.0	20.0	7195	3.99%	3.99%
50.0	50.0	7191	3.99%	3.99%
40.0	40.0	6765	3.75%	3.75%
10.0	10.0	6600	3.66%	3.66%
25.0	25.0	5650	3.13%	3.13%
55.0	55.0	5449	3.02%	3.02%
35.0	35.0	5448	3.02%	3.02%
5.0	5.0	5342	2.96%	2.96%
8.0	8.0	2318	1.29%	1.29%
18.0	18.0	2233	1.24%	1.24%
28.0	28.0	2227	1.23%	1.23%
58.0	58.0	2225	1.23%	1.23%
38.0	38.0	2196	1.22%	1.22%
12.0	12.0	2154	1.19%	1.19%
42.0	42.0	2152	1.19%	1.19%
48.0	48.0	2145	1.19%	1.19%
7.0	7.0	2079	1.15%	1.15%
43.0	43.0	2070	1.15%	1.15%
14.0	14.0	2048	1.14%	1.14%
57.0	57.0	2013	1.12%	1.12%
19.0	19.0	2011	1.12%	1.12%
52.0	52.0	2006	1.11%	1.11%
13.0	13.0	2004	1.11%	1.11%

23.0	23.0	2001	1.11%	1.11%
22.0	22.0	1994	1.11%	1.11%
53.0	53.0	1994	1.11%	1.11%
17.0	17.0	1994	1.11%	1.11%
44.0	44.0	1979	1.1%	1.1%
54.0	54.0	1979	1.1%	1.1%
24.0	24.0	1973	1.09%	1.09%
37.0	37.0	1967	1.09%	1.09%
27.0	27.0	1964	1.09%	1.09%
2.0	2.0	1937	1.07%	1.07%
32.0	32.0	1919	1.06%	1.06%
4.0	4.0	1919	1.06%	1.06%
33.0	33.0	1917	1.06%	1.06%
34.0	34.0	1912	1.06%	1.06%
1.0	1.0	1908	1.06%	1.06%
36.0	36.0	1904	1.06%	1.06%
47.0	47.0	1895	1.05%	1.05%
3.0	3.0	1891	1.05%	1.05%
9.0	9.0	1883	1.04%	1.04%
39.0	39.0	1877	1.04%	1.04%
56.0	56.0	1866	1.03%	1.03%
26.0	26.0	1865	1.03%	1.03%
59.0	59.0	1859	1.03%	1.03%
6.0	6.0	1856	1.03%	1.03%
49.0	49.0	1850	1.03%	1.03%
29.0	29.0	1835	1.02%	1.02%
46.0	46.0	1819	1.01%	1.01%
51.0	51.0	1798	1.0%	1.0%
16.0	16.0	1785	0.99%	0.99%
41.0	41.0	1777	0.99%	0.99%
21.0	21.0	1762	0.98%	0.98%
11.0	11.0	1743	0.97%	0.97%

31.0	31.0	1601	0.89%	0.89%

A.1.2.2 Descriptive Statistics of Number of Persons Involved in the Crash

Table A.26: Number of Persons Involved in the Crash Distribution

Metric	Value
Valid %	100.0%
Skewness	10.2331

A.1.2.3Descriptive Statistics of Number of Pedestrians Involved in the Crash

Table A.27: Number of Pedestrians Involved in the Crash Distribution

Metric	Value
Valid %	100.0%
Skewness	4.6238

A.1.2.4Descriptive Statistics of Road Functional Class

Table A.28: Road Functional Class Distribution

Code	Label	Frequency	Pct_Total	Pct_Valid
4.0	Major Collector-Rural	24399	13.53%	13.53%
13.0	Other Principal Arterial	22897	12.7%	12.7%
2.0	Principal Arterial - Rural	20316	11.27%	11.27%
6.0	Local Road/Street-Rural	19454	10.79%	10.79%
3.0	Minor Arterial-Rural	17317	9.6%	9.6%
16.0	Street	16429	9.11%	9.11%
14.0	Minor Arterial	16278	9.03%	9.03%
1.0	Principal Arterial - Interstate - Rural	11357	6.3%	6.3%
11.0	Principal Arterial	11306	6.27%	6.27%
12.0	Expressway	7142	3.96%	3.96%
5.0	Minor Collector-Rural	7001	3.88%	3.88%
15.0	Collector	6440	3.57%	3.57%

A.1.2.5Descriptive Statistics of Junction Relation with Crash Location

Table A.29: Junction Relation with Crash Location Distribution

Code	Label	Frequency	Pct_Total	Pct_Valid
1.0	Non-Junction	130460	72.34%	72.34%
2.0	Intersection	31980	17.73%	17.73%
3.0	Intersection-Related	6606	3.66%	3.66%
8.0	Driveway Access Related	2375	1.32%	1.32%
4.0	Driveway, Alley Access, etc.	2075	1.15%	1.15%
13.0	Entrance/Exit Ramp-Interchange	1865	1.03%	1.03%
15.0	Other Location in Interchange	1700	0.94%	0.94%
6.0	Railway Grade Crossing	933	0.52%	0.52%
5.0	Entrance/Exit Ramp-Related	872	0.48%	0.48%
10.0	Intersection-Interchange	788	0.44%	0.44%
11.0	Intersection-Related-Interchange	373	0.21%	0.21%
7.0	In Crossover	188	0.1%	0.1%
12.0	Driveway Access-Interchange	89	0.05%	0.05%
14.0	In Crossover-Interchange	32	0.02%	0.02%

A.1.2.6 Descriptive Statistics of Vehicle Location Related to Roadway

Table A.30: Vehicle Location Related to Roadway Distribution

Code	Label	Frequency	Pct_Total	Pct_Valid
1.0	Non-Junction	100931	55.97%	55.97%
4.0	Driveway, Alley Access, etc.	43621	24.19%	24.19%
2.0	Intersection	12114	6.72%	6.72%
6.0	Railway Grade Crossing	9933	5.51%	5.51%
5.0	Entrance/Exit Ramp-Related	6808	3.78%	3.78%
3.0	Intersection-Related	5566	3.09%	3.09%
8.0	Driveway Access Related	613	0.34%	0.34%
7.0	In Crossover	384	0.21%	0.21%
10.0	Intersection-Interchange	298	0.17%	0.17%
11.0	Intersection-Related-Interchange	68	0.04%	0.04%

A.1.2.7Descriptive Statistics of Speed Limit

Table A.31: Speed Limit Distribution

Metric	Value
Valid %	100.0%
Skewness	-0.3399

A.1.2.8Descriptive Statistics of Road Alignment

Table A.32: Road Alignment Distribution

Code	Label	Frequency	Pct_Total	Pct_Valid
1.0	Straight	131370	72.85%	72.85%
2.0	Curved	48966	27.15%	27.15%

A.1.2.9Descriptive Statistics of Road Profile

Table A.33: Road Profile Distribution

Code	Label	Frequency	Pct_Total	Pct_Valid
1.0	Level	131161	72.73%	72.73%
2.0	Grade	44134	24.47%	24.47%
3.0	Hillcrest	4467	2.48%	2.48%
4.0	Sag	574	0.32%	0.32%

A.1.2.10 Descriptive Statistics of Pavement Type

Table A.34: Pavement Type Distribution

Code	Label	Frequency	Pct_Total	Pct_Valid
2.0	Blacktop, Bituminous, or Asphalt	160670	89.09%	89.09%
1.0	Concrete	15731	8.72%	8.72%
4.0	Slag, Gravel or Stone	2490	1.38%	1.38%
5.0	Dirt	1369	0.76%	0.76%
3.0	Brick or Block	76	0.04%	0.04%

A.1.2.11 Descriptive Statistics of Roadway Surface Condition

Table A.35: Roadway Surface Condition Distribution

Code	Label	Frequency	Pct_Total	Pct_Valid
1.0	Dry	151288	83.89%	83.89%
2.0	Wet	21755	12.06%	12.06%
4.0	Ice/Frost	2744	1.52%	1.52%
3.0	Snow or Slush	2598	1.44%	1.44%
9.0	Unknown	1201	0.67%	0.67%
5.0	Sand, Dirt, Mud, Gravel	393	0.22%	0.22%
8.0	Other	213	0.12%	0.12%
6.0	Water (Standing or Moving)	131	0.07%	0.07%
7.0	Oil	13	0.01%	0.01%

A.1.2.12 Descriptive Statistics of Manner of Collision

Table A.36: Manner of Collision Distribution

Code	Label	Frequency	Pct_Total	Pct_Valid
0.0	Not Collision with Motor Vehicle In-Transport	110198	61.11%	61.11%
5.0	Angle - Front-to-Side, Right Angle (Includes Broadside)	23382	12.97%	12.97%
2.0	Front-to-Front	17989	9.98%	9.98%
1.0	Front-to-Rear	11087	6.15%	6.15%
4.0	Angle - Front-to-Side, Opposite Direction	9113	5.05%	5.05%
7.0	Sideswipe - Same Direction	2406	1.33%	1.33%
3.0	Angle - Front-to-Side, Same Direction	2297	1.27%	1.27%
8.0	Sideswipe - Opposite Direction	2128	1.18%	1.18%
6.0	Angle - Front-to-Side/Angle-Direction Not Specified	1006	0.56%	0.56%
11.0	Other (End-Swipes and Others)	379	0.21%	0.21%
9.0	Rear-to-Side	338	0.19%	0.19%
10.0	Rear-to-Rear	13	0.01%	0.01%

A.1.3 Feature Significance Testing

The descriptive statistical analysis of each feature was performed and is presented feature-wise.

A.1.3.1 Feature Significance of Month of Crash

The summary of chi-square significance test is as below:

Table A.37: Month of Crash significance test summary

Metric	Value	Asymp. Sig (2-sided)
Pearson Chi-Square	615.7762537343149	0.0000*
Likelihood Ratio	613.1205834885291	0.0000*
Linear-by-Linear Association	61.0570496327507	0.0000*
N of Valid Cases	179919.0	-
Degrees of Freedom	44.0	-

The graph of significance is presented below:

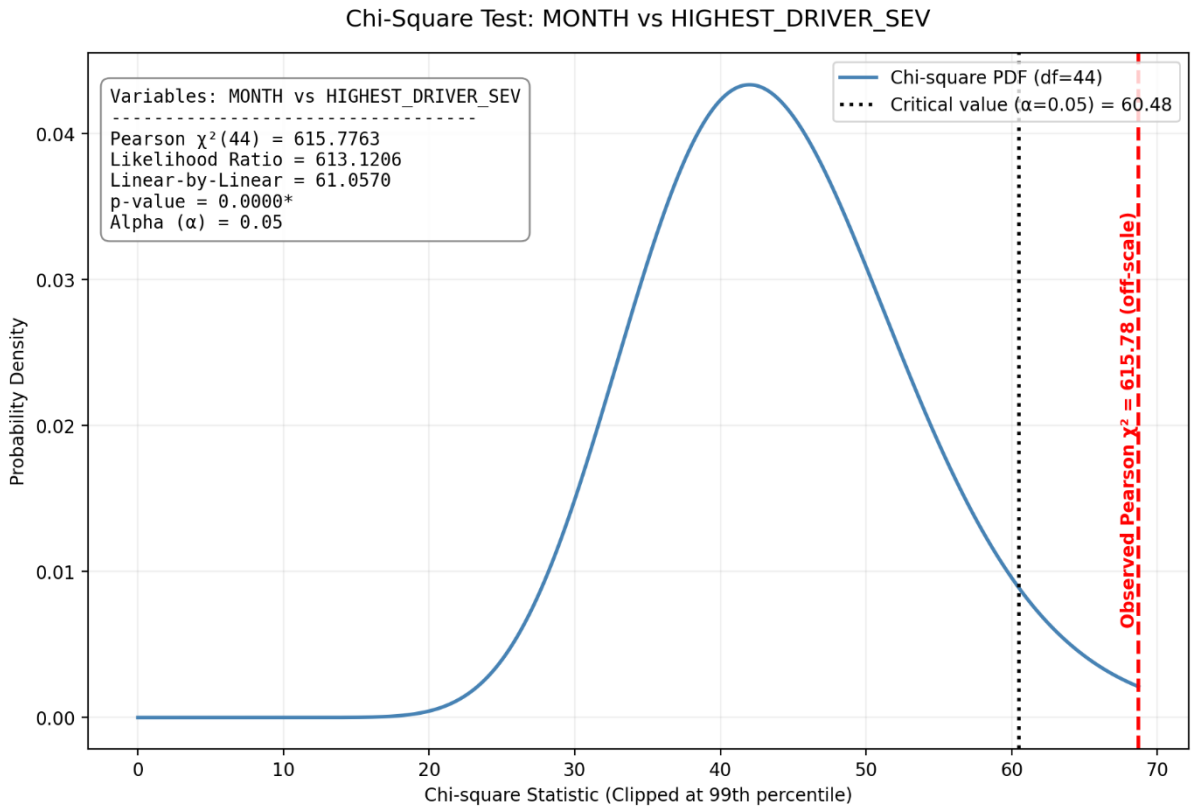

Figure A.37: Chi-Square Test for MONTH

A.1.3.2Feature Significance of Minute of Crash

The summary of chi-square significance test is as below:

Table A.38: Minute of Crash significance test summary

Metric	Value	Asymp. Sig (2-sided)
Pearson Chi-Square	567.6536933426298	0.0000*
Likelihood Ratio	568.1311928501993	0.0000*
Linear-by-Linear Association	6.19984353994001	0.0128*
N of Valid Cases	179919.0	-
Degrees of Freedom	236.0	-

The graph of significance is presented below:

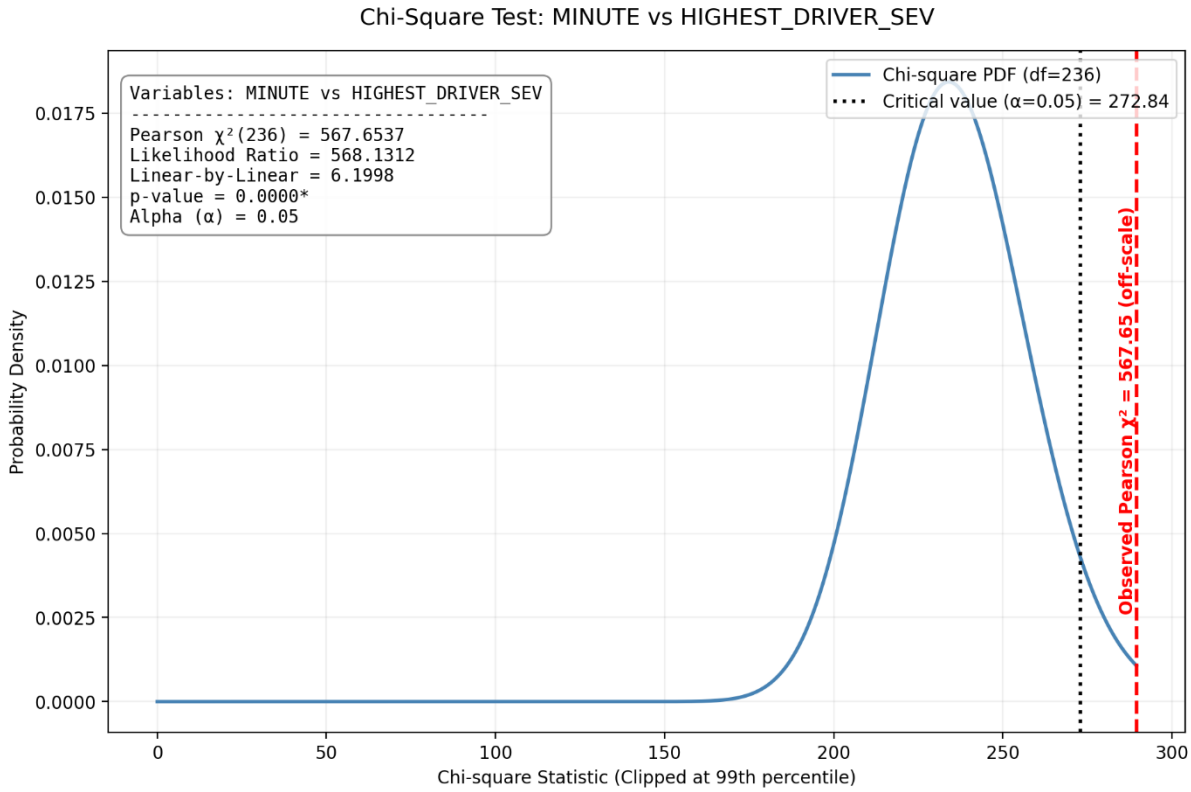

Figure A.38: Chi-Square Test for MINUTE

A.1.3.3Feature Significance of Number of Persons Involved in the Crash

The summary of kruskal-wallis test is as below:

Table A.39: Number of Persons Involved in the Crash significance test summary

	Test	Feature	Target	H_stat	p_value	N_valid
0	Kruskal-Wallis	N_PERSONS	HIGHEST_DRIVER_SEV	28193.42508875786	0.0	179919

A.1.3.4Feature Significance of Number of Pedestrians Involved in the Crash

The summary of kruskal-wallis test is as below:

Table A.40: Number of Pedestrians Involved in the Crash significance test summary

	Test	Feature	Target	H_stat	p_value	N_valid
0	Kruskal-Wallis	PEDS	HIGHEST_DRIVER_SEV	132371.85445160413	0.0	179919

A.1.3.5Feature Significance of Road Functional Class

The summary of chi-square significance test is as below:

Table A.41: Road Functional Class significance test summary

Metric	Value	Asymp. Sig (2-sided)
Pearson Chi-Square	11781.36949168608	0.0000*
Likelihood Ratio	11399.328015457657	0.0000*
Linear-by-Linear Association	6621.926730565271	0.0000*
N of Valid Cases	179919.0	-
Degrees of Freedom	44.0	-

The graph of significance is presented below:

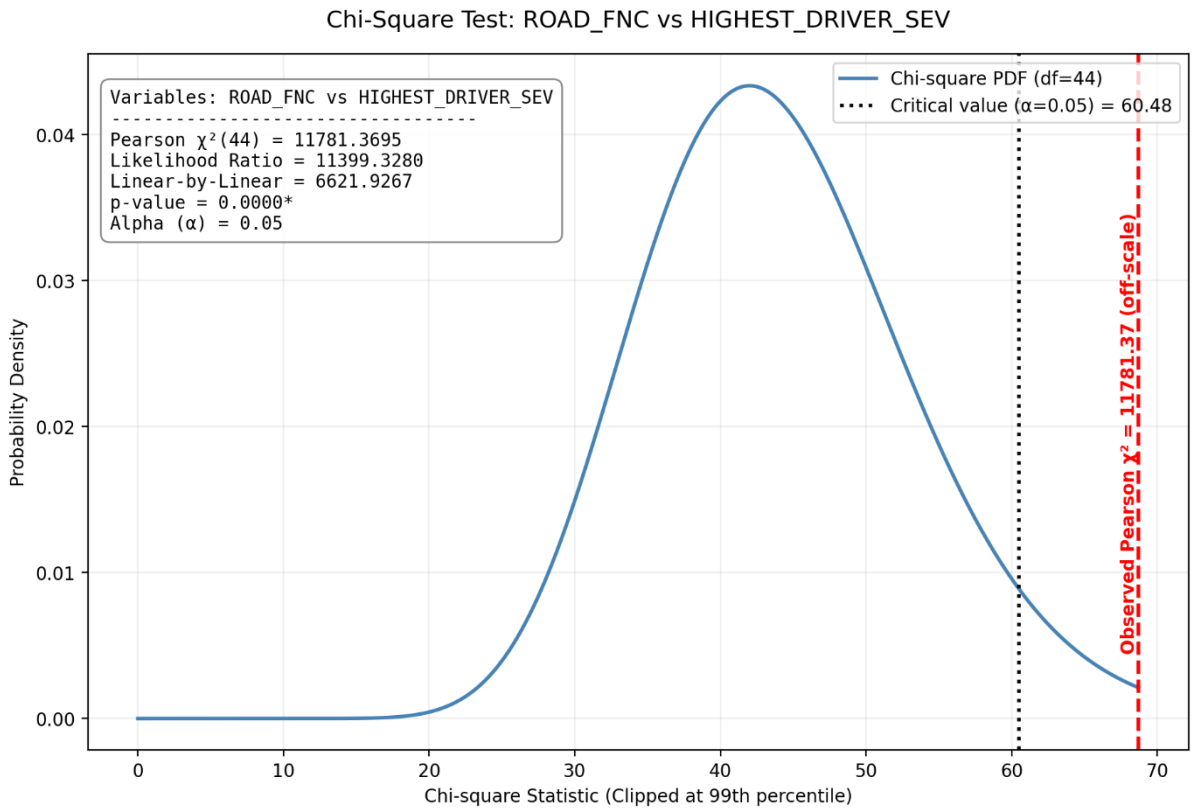

Figure A.39: Chi-Square Test for ROAD_FNC

A.1.3.6Feature Significance of Junction Relation with Crash Location

The summary of chi-square significance test is as below:

Table A.42: Junction Relation with Crash Location significance test summary

Metric	Value	Asymp. Sig (2-sided)
Pearson Chi-Square	2341.5193385161137	0.0000*
Likelihood Ratio	2075.263321747403	0.0000*
Linear-by-Linear Association	1.4812040968713605	0.2236
N of Valid Cases	179919.0	-
Degrees of Freedom	52.0	-

The graph of significance is presented below:

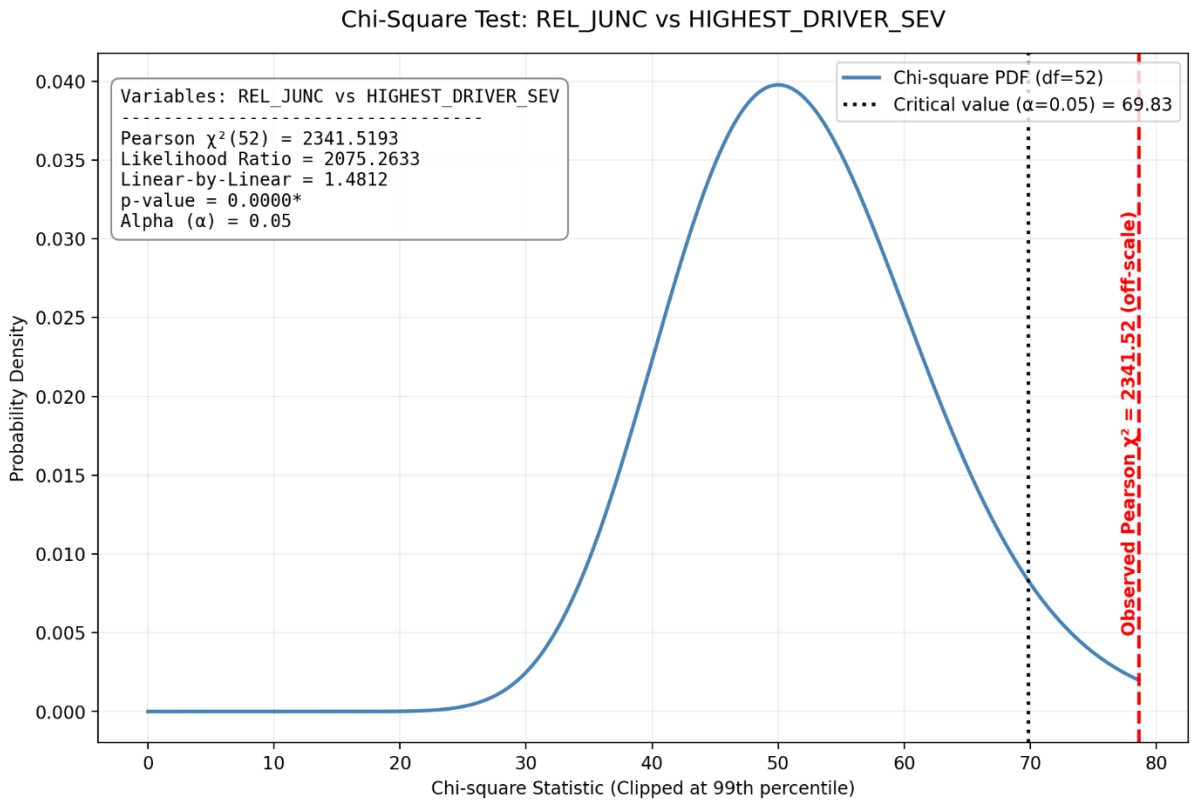

Figure A.40: Chi-Square Test for REL_JUNC

A.1.3.7Feature Significance of Vehicle Location Related to Roadway

The summary of chi-square significance test is as below:

Table A.43: Vehicle Location Related to Roadway significance test summary

Metric	Value	Asymp. Sig (2-sided)
Pearson Chi-Square	14014.088354851678	0.0000*
Likelihood Ratio	16550.92029394711	0.0000*
Linear-by-Linear Association	10164.703077185704	0.0000*
N of Valid Cases	179919.0	-
Degrees of Freedom	36.0	-

The graph of significance is presented below:

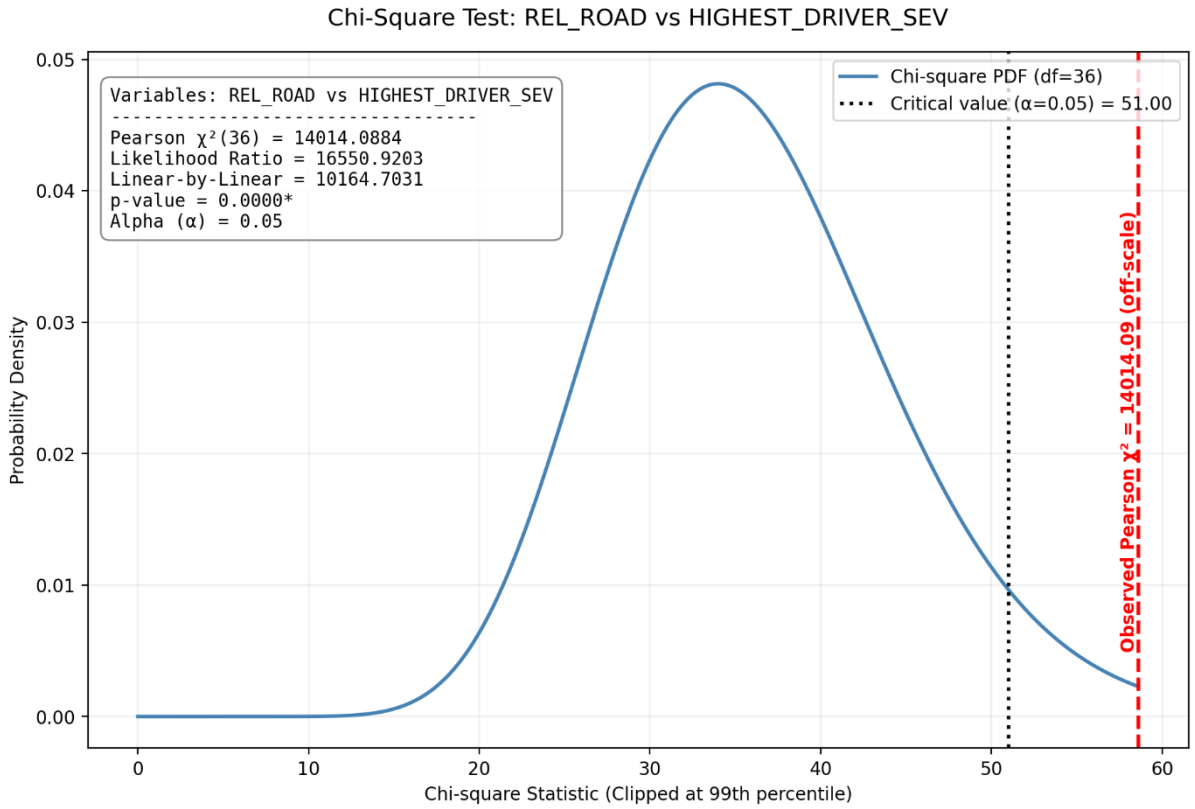

Figure A.41: Chi-Square Test for REL_ROAD

A.1.3.8Feature Significance of Speed Limit

The summary of kruskal-wallis test is as below:

Table A.44: Speed Limit significance test summary

	Test	Feature	Target	H_stat	p_value	N_valid
0	Kruskal-Wallis	SP_LIMIT	HIGHEST_DRIVER_SEV	7017.160836774733	0.0	179919

A.1.3.9Feature Significance of Road Alignment

The summary of chi-square significance test is as below:

Table A.45: Road Alignment significance test summary

Metric	Value	Asymp. Sig (2-sided)
Pearson Chi-Square	5946.818965544749	0.0000*
Likelihood Ratio	7000.547489454859	0.0000*
Linear-by-Linear Association	5821.362572351285	0.0000*

N of Valid Cases	179919.0	-
Degrees of Freedom	4.0	-

The graph of significance is presented below:

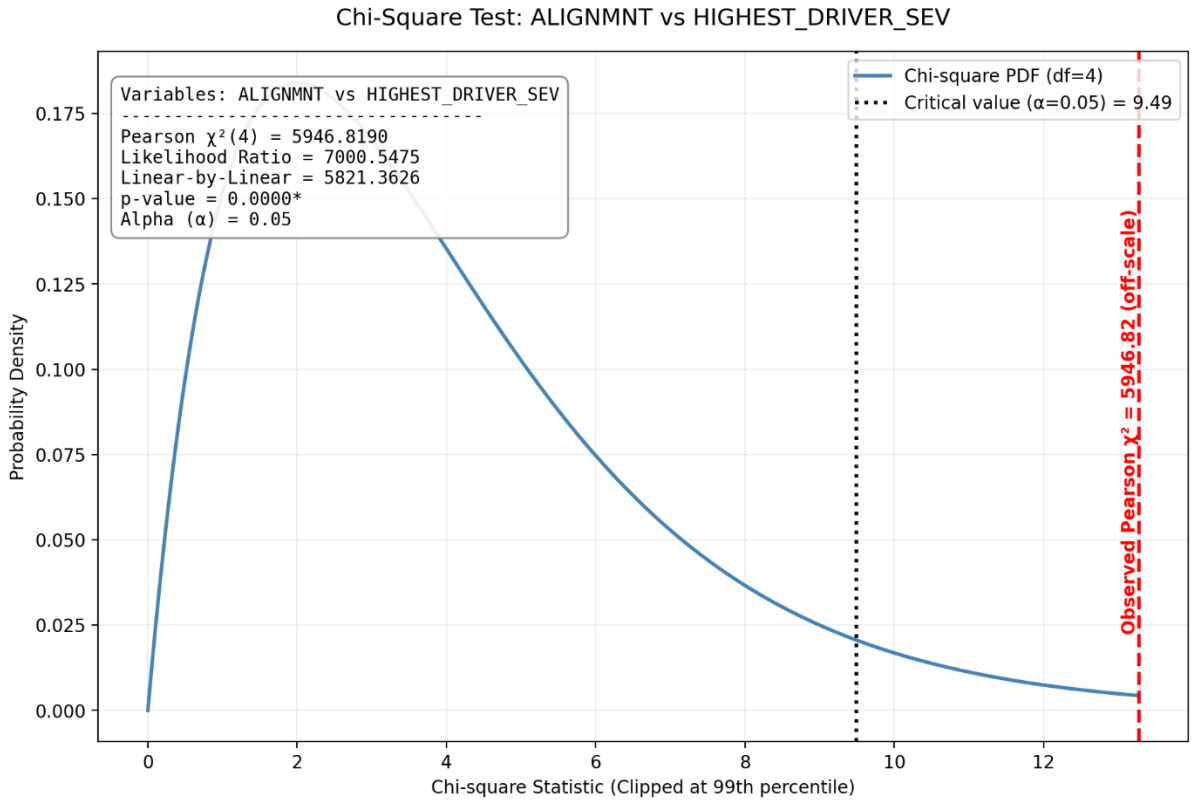

Figure A.42: Chi-Square Test for ALIGNMNT

A.1.3.10 Feature Significance of Road Profile

The summary of chi-square significance test is as below:

Table A.46: Road Profile significance test summary

Metric	Value	Asymp. Sig (2-sided)
Pearson Chi-Square	1791.7608658649715	0.0000*
Likelihood Ratio	1925.6717694277438	0.0000*
Linear-by-Linear Association	1463.7494450729598	0.0000*
N of Valid Cases	179919.0	-
Degrees of Freedom	12.0	-

The graph of significance is presented below:

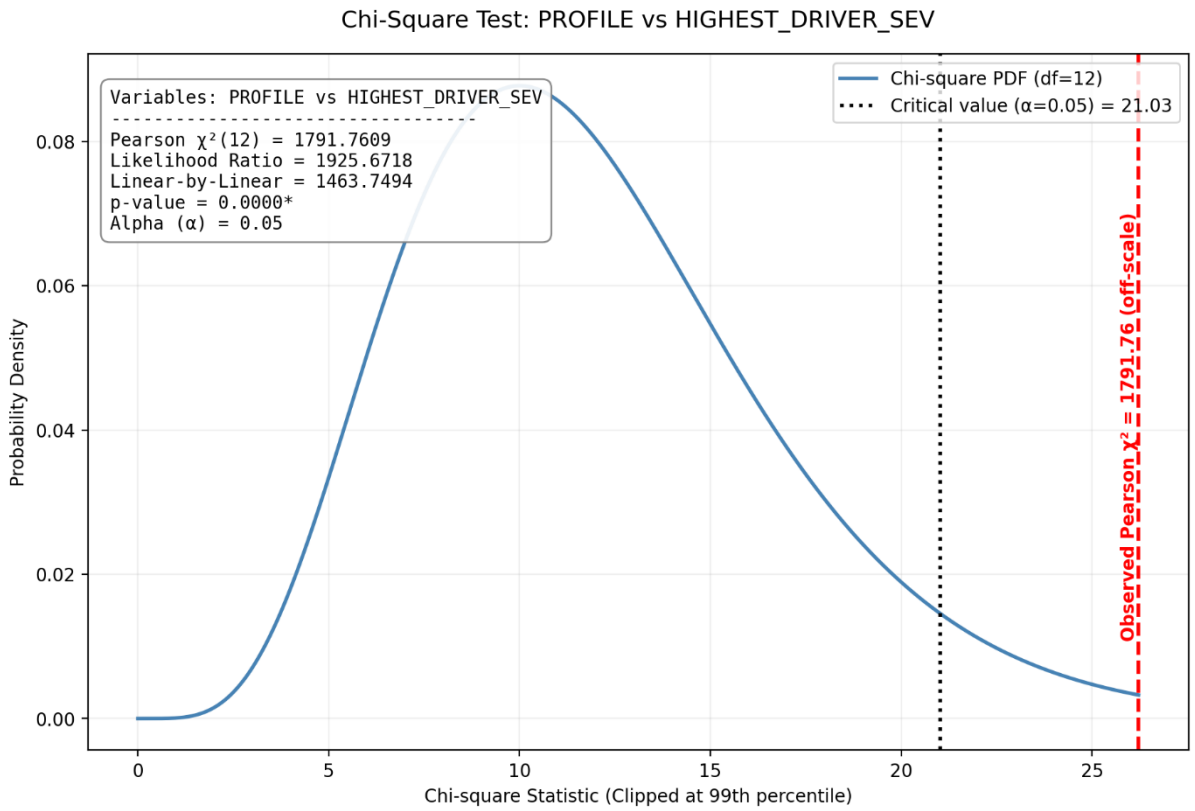

Figure A.43: Chi-Square Test for PROFILE

A.1.3.11 Feature Significance of Pavement Type

The summary of chi-square significance test is as below:

Table A.47: Pavement Type significance test summary

Metric	Value	Asymp. Sig (2-sided)
Pearson Chi-Square	300.9270524993437	0.0000*
Likelihood Ratio	312.8228126434507	0.0000*
Linear-by-Linear Association	92.10446254512014	0.0000*
N of Valid Cases	179919.0	-
Degrees of Freedom	16.0	-

The graph of significance is presented below:

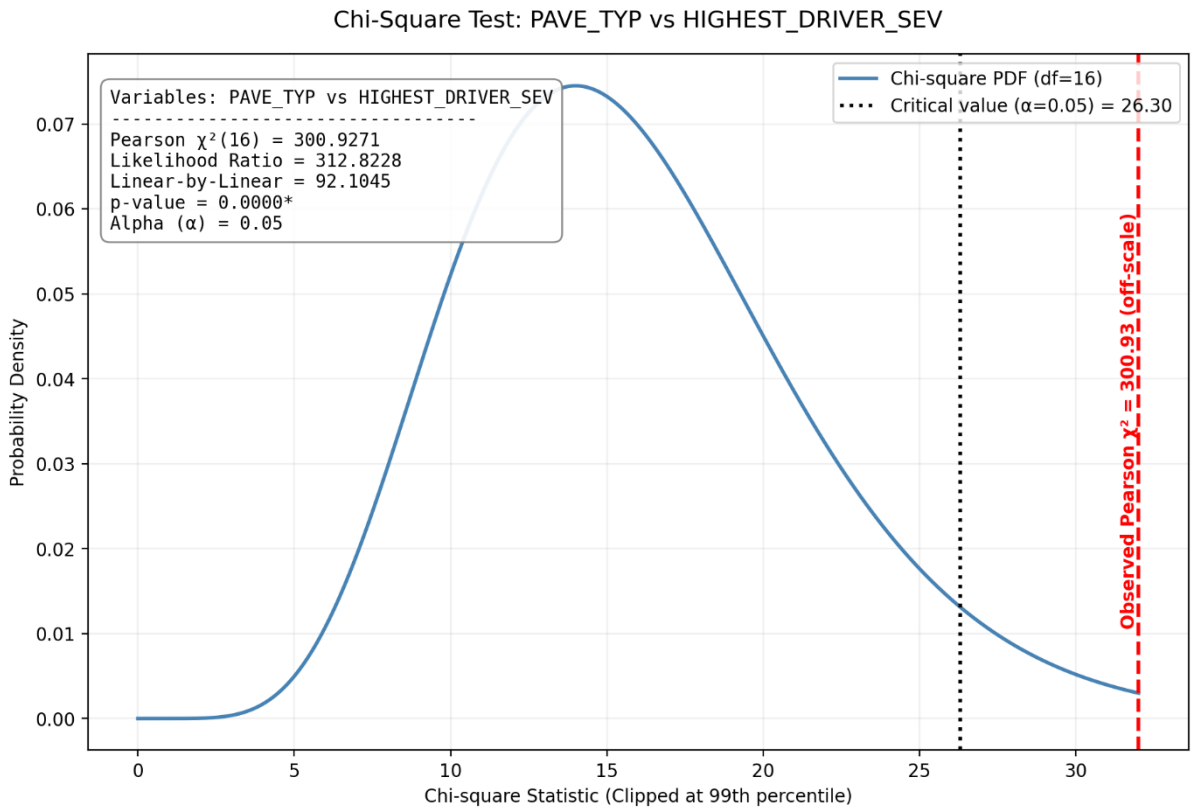

Figure A.44: Chi-Square Test for PAVE_TYP

A.1.3.12 Feature Significance of Roadway Surface Condition

The summary of chi-square significance test is as below:

Table A.48: Roadway Surface Condition significance test summary

Metric	Value	Asymp. Sig (2-sided)
Pearson Chi-Square	605.8578932468682	0.0000*
Likelihood Ratio	532.0698892817551	0.0000*
Linear-by-Linear Association	5.142330896702372	0.0233*
N of Valid Cases	179919.0	-
Degrees of Freedom	32.0	-

The graph of significance is presented below:

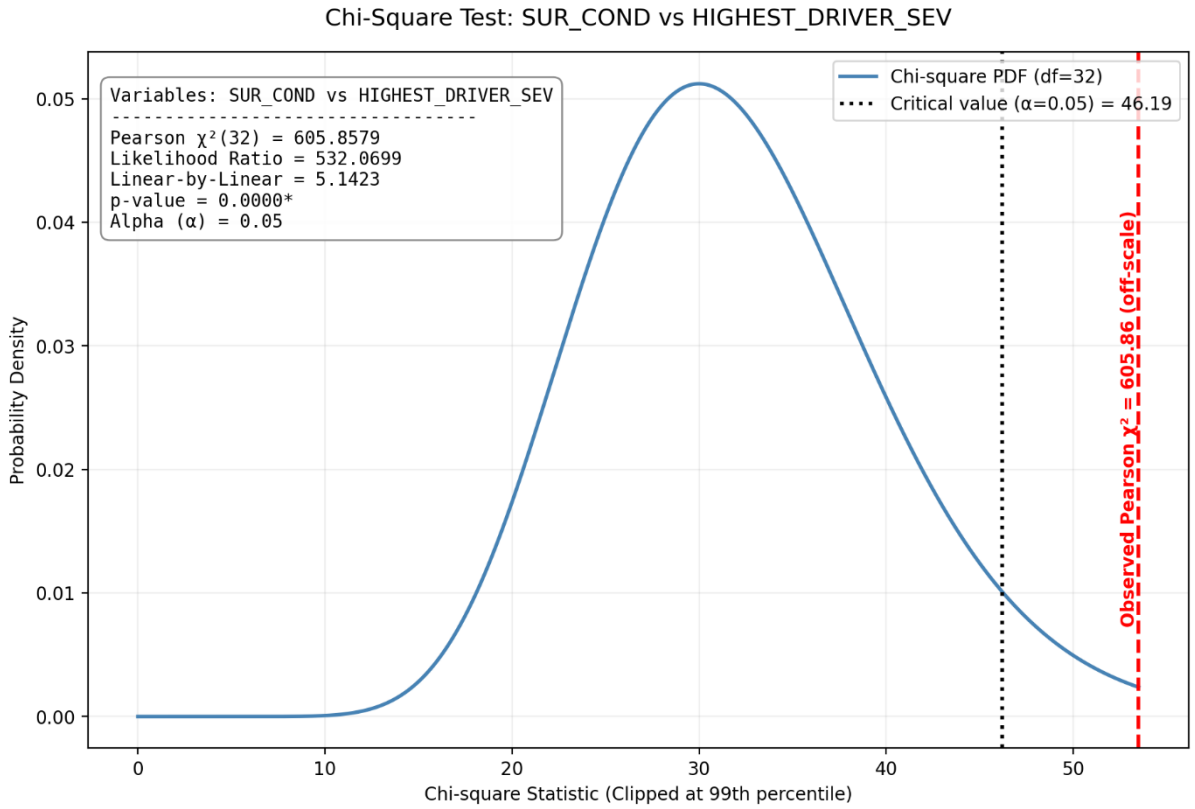

Figure A.45: Chi-Square Test for SUR_COND

A.1.3.13 Feature Significance of Manner of Collision

The summary of chi-square significance test is as below:

Table A.49: Manner of Collision significance test summary

Metric	Value	Asymp. Sig (2-sided)
Pearson Chi-Square	16372.068961305149	0.0000*
Likelihood Ratio	21021.630918509178	0.0000*
Linear-by-Linear Association	8258.590188669747	0.0000*
N of Valid Cases	179919.0	-
Degrees of Freedom	44.0	-

The graph of significance is presented below:

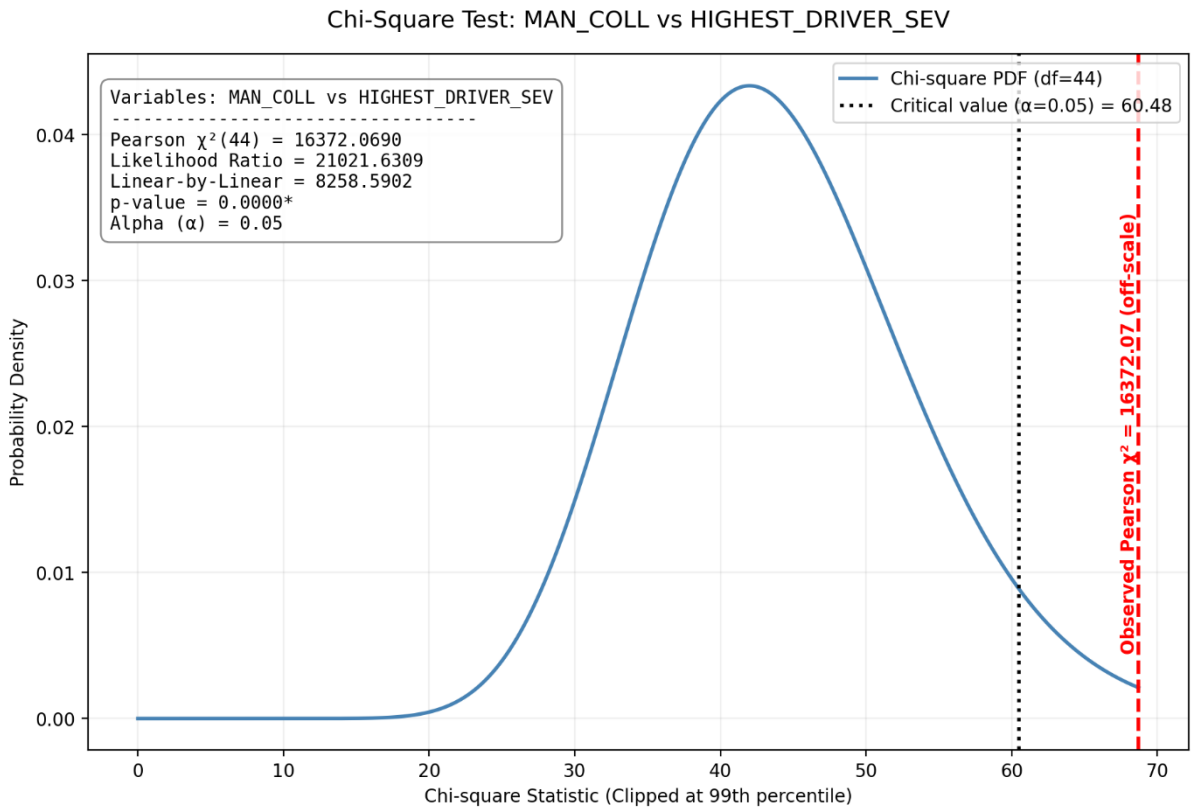

Figure A.46: Chi-Square Test for MAN_COLL