Enhancing Pedestrian Safety in Great Britain:

Analysis of Road Collision Data and Policy Effectiveness

Tian Tong

Introduction

Road traffic injuries are a major public health problem and a leading cause of death and injury around the world. According to WHO, globally, road traffic crashes kill approximately 1.3 million people every year - more than two every minute - with nine in ten deaths occurring in low-and middle-income countries. In Great Britain, pedestrians accounted for 24% of all road deaths (5,393 out of 27,450 total road users) in 2021, from the Department for Transport Annual Report and Accounts. In view of this, modeling pedestrian safety in relation to road traffic accidents is an essential research objective.

Despite the UK's significant progress in reducing road deaths over the last 30 years, the volume of traffic casualties remains considerable. The Transportation Department has implemented major changes to promote road safety, such as amendments to the Highway Code two years ago, requiring drivers entering or exiting a junction to give way to pedestrians, cyclists, and horse riders who are crossing or waiting to cross the road. However, many drivers perceive junctions as even more hazardous for pedestrians, aligning with the Government's Public Accounts Committee report of November 2023, which criticized the promotion of these safety changes as inadequate in fostering public compliance and awareness (Roberts, 2024).

The UK has made significant progress in reducing road deaths over the last 30 years, but the volume of traffic casualties remains considerable. Previous study by Choudhary et al. (2023) provides an exploratory data analysis on road data safety in Great Britain, identifying factors such as rapid urbanization, increasing vehicle numbers, speeding, and general negligence of road safety as contributing to the rise in pedestrian accidents. This underscores the need for further research to evaluate the efficacy of current policies and examine areas for improvement.

This study aims to develop predictive models that prioritize pedestrian safety, account for the vulnerability of pedestrians on roads, and aim to reduce disparities in injury risks among different social groups. Utilizing the Road Safety Data of 2023, published by the Department for Transport, we will construct machine learning models, such as Random Forest Tree models, to demonstrate the factors influencing pedestrian safety and collision severity. By identifying the most significant predictors, we aim to assess the effectiveness of current policy implementations and provide data-driven recommendations for further enforcement or policy adjustments, ultimately contributing to the development of more effective interventions and policies to reduce accidents and enhance road safety.

Road Safety | Pedestrian Safety Protection | UK Transportation Policy | Prediction Model

Methodology

Data

To create a model predicting pedestrian road safety, we utilize the collision dataset from the 2023 Road Accident and Safety Statistics, published by the Department for Transport for the UK Government. The

STATS19 dataset, named after the data collection form used by the police, is pivotal for analyzing road traffic collisions and pedestrian safety in Great Britain. It primarily includes statistics from collisions where at least one person was injured and reported to the police, who then relay the data to the Department for Transport. The dataset contains 49316 records, each representing a unique collision.

While historical statistics have used the terminology ”road traffic accidents” in line with the Road Traffic Act, the terms ”accidents” and ”collisions” are now used interchangeably in the current context of road safety since 2022. For a collision to be included, it must have involved at least one vehicle, and resulted in personal injury. Collisions that do not result in actual bodily harm, such as ”damage-only”, are not in scope and are not estimated.

The dataset captures a wide range of incidents, from accidents involving motor vehicles to those involving pedestrians or pedal cyclists falling off their bicycles on the road, even if no other vehicle or pedestrian was involved. However, the dataset has its limitations in that there is no legal obligation for drivers to report a road traffic collision if all parties exchange details, even if injuries occur, leading to potential under-reporting. Also, under the analysis target of pedestrian safety, the study of only-vehicle involved accidents may result in the underestimation.

Despite this limitation, the road safety dataset remains a valuable resource for predicting pedestrian safety due to its comprehensive collection. The inclusion of pedestrian casualties and the clear definitions of traffic terminologies contribute to the dataset's relevance for analyzing pedestrian safety. Moreover, the technological advancements in data collection methods, such as injury-based reporting systems and online platforms, greatly enhanced the quality and accuracy (Reported Road Casualty Statistics: Background Quality Report, 2023).

For the forthcoming analysis, Table 1 presents the variables from the original dataset with descriptions and notes from the official Background Quality Report to aid understanding. In Table 2, the addition of column 'Observation' presents the occurrence counts of each target feature before and post revision to show whether there is a necessity for further imputation or adjustment. The preprocessing steps performed on the dataset will be discussed in detail in the following section. Note that the minimum, mean, maximum values, and count of observations mentioned in the subsequent analysis are based on the preprocessed data, which may show minor discrepancies compared to the descriptive statistics summary of the raw dataset.

Table 1: Relevant Features Introduction

Variable Name	**Description**	**Mean**	**Min**	**Max**	**Notes**
number_of_vehicles	Number of vehicles involved in the accidents	1.81	1	17	
number_of_casualties	Number of casualties in the accidents	1.27	1	19	
day_of_week	The date of the accident	4.11	1	7	Represents the respective day of week

Continued on next page

Table 1 Continued from previous page

Variable Name	Description	Mean	Min	Max	Notes
nearest_hour	The exact time when the accident happened	13.95	0	24	The hours are to be entered in the first two boxes and transformed to the nearest hour
road_type	The type of road where the accident took place	5.31	1	9	1. Roundabout 2. One way street 3. Dual carriageway 6. Single carriageway 7. Slip Road 9. Unknown
speed_limit	The speed limit at the location of the accident	35.67	0	70	
junction_control	The type of control mechanism at the junction	3.75	1	4	1. Authorized person 2. Auto traffic signal 3. Stop sign 4. Uncontrolled
junction_detail	The type of junction within a 20-meter radius	2.40	0	9	0. Not near junction 1. Roundabout 2. Mini roundabout 3. T junction 5. Slip road 6. Crossroads 7. More than four arms 8. Private drive 9. Other
Pedestrian_crossing_human_control	Human intervention at pedestrian crossing areas	0.03	0	2	0. None within 50 meters 1. School crossing patrol 2. Other authorized person
light_conditions	The illumination level at the accident location	1.93	1	7	1. Daylight 4. Street lights lit 5. Street lights unlit 6. No street lighting 7. Lighting unknown

Continued on next page

Table 1 Continued from previous page

Variable Name	Description	Mean	Min	Max	Notes
weather_conditions	The weather condition at the accident location	1.64	1	9	1. Fine no high winds 2. Raining no high winds 3. Snowing no high winds 4. Fine high winds 5. Raining high winds 6. Snowing high winds 7. Fog or mist 8. Other 9. Unknown
road_surface_conditions	The road surface condition at the accident location	1.28	1	5	1. Dry 2. Wet/Damp 3. Snow 4. Frost/Ice 5. Flooded (over 3 cm deep)
did_police_officer_attend_scene_of_collision	Police presence at the collision scene	1.53	1	3	1. Yes 2. No 3. Self-reporting form used

Target Features and Imputation Firstly, we integrate the `'casualty_class'` variable from the provisional road casualty statistics report, part of the same 2023 Road Accident and Safety Statistics. This data provided detailed accounts of each accident, including cases where collision references were reported multiple times by different individuals involved. To manage this, we create a new dummy variable to determine whether any of the involved individuals is a pedestrian. We then merge this information into our main collision dataset using the collision reference number, ensuring that each collision is recorded only once. Importantly, our collision data contains unique reports for each incident, so there are no multiple entries for a single collision to contend with.

From the original distribution plot in Figure 1, we observe significant class imbalance in our target features, with a disproportionate focus on the majority class. For the Severity Level feature, Our particular concern is the minority class, which includes only 678 fatal cases and even fewer (194) pedestrian-related fatalities. This limited data poses challenges for the credibility and robustness of any predictive modeling focused solely on fatal outcomes. To address this, we consider expanding our target feature to include both serious and fatal cases. Combining these categories would not only provide a larger sample size but also maintain a focus on the more serious outcomes, which are of greater interest in safety studies.

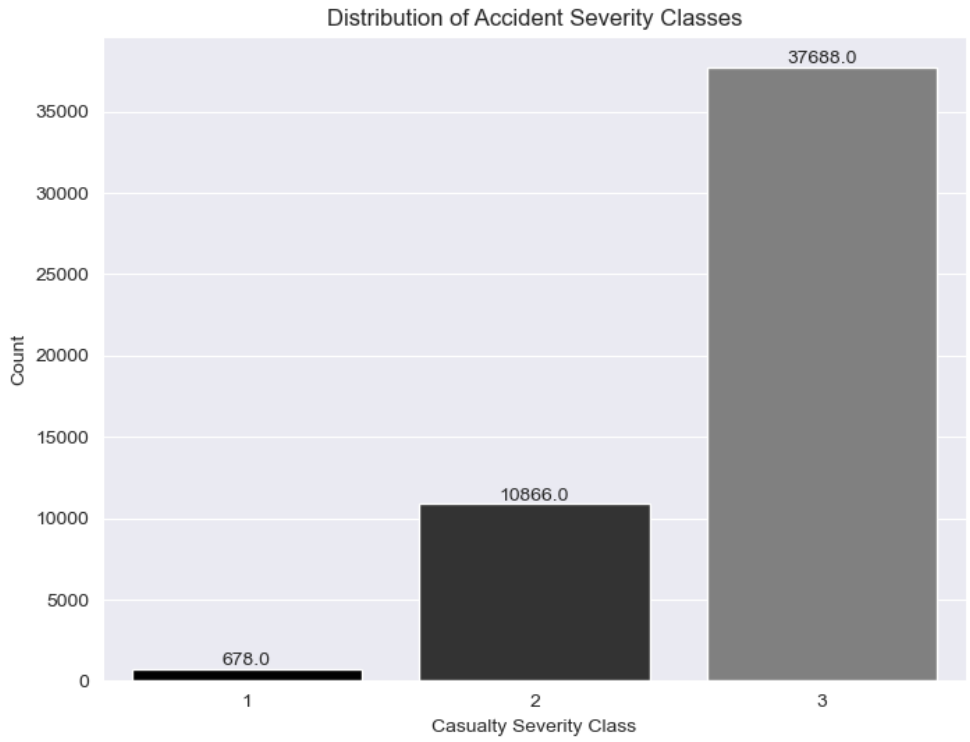

Figure 1: Distribution of Accident Severity Classes

Table 2: Target Feature Introduction

Variable Name	**Description**	**Observations**	**Changes for the Project**
casualty_pedestrian	Indicates the class of the casualty (0. Not pedestrian involved: Including driver or rider, vehicle or pillion passenger 1. Pedestrian involved)	Non pedestrian-involved 40351 Pedestrian-involved 8881	Created a dummy variable for whether pedestrians were involved
casualty_severity	Indicates the severity level of the casualty (1. Fatal 2. Serious 3. Slight)	Fatal 678 Serious 10866 Slight 37688	Created a dummy variable for whether the severity level of the casualty is beyond serious or not
casualty_over_serious	Indicates whether the severity level of the casualty is more than serious or not	Not serious 37668 Serious 11544	

Continued on next page

Table 2 Continued from previous page

Variable Name	Description	Observations	Changes for the Project
pedestrian_over_serious	Indicates whether the casualty includes pedestrian and is serious	False 46409 True 2823	Generated by multiplication of `casualty_pedestrian` and `casualty_over_serious` to make an interaction term

Descriptives

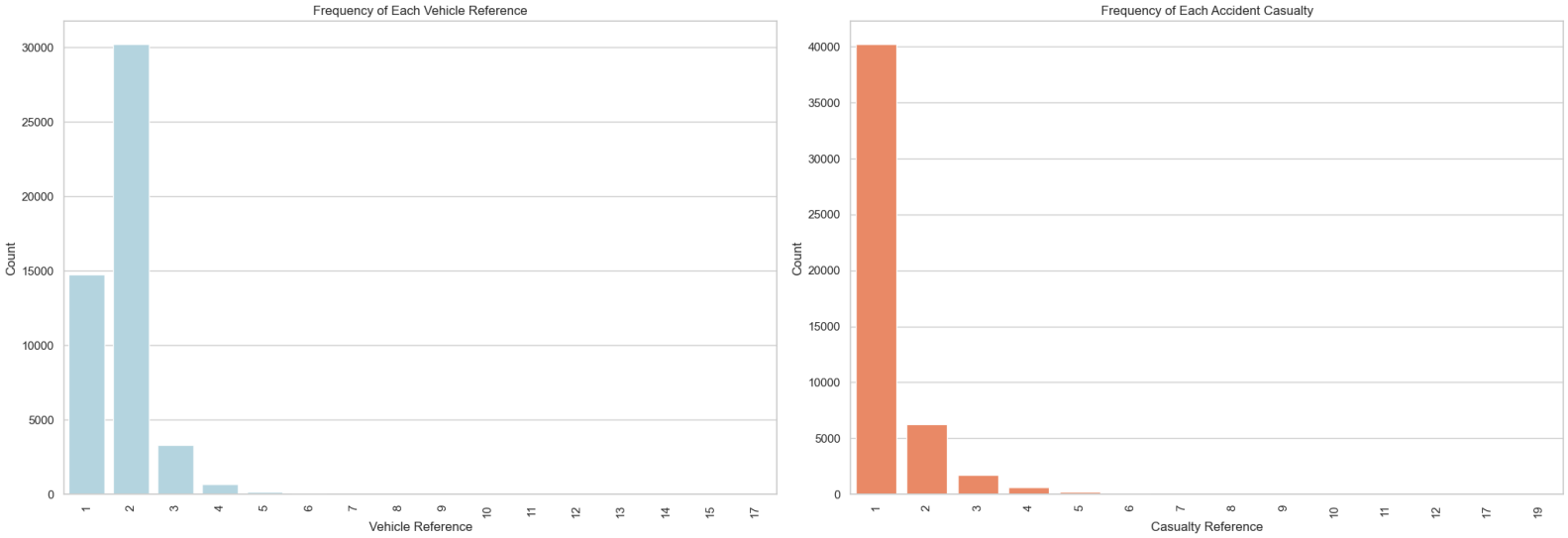

Figure 2: Distribution of Number of Vehicles/Casualties

Based on the provided plots and summary statistics before preprocessing, several key insights can be drawn about the collision data. The mean value demonstrated for `'number_of_vehicles'` indicates that, on average, there are between one and two vehicles involved in each collision. This distribution is highly left-skewed, with the majority of collisions involving just one or two vehicles, highlighting the commonality of smaller incidents over more complex multi-vehicle collisions. There are observations with a high value of 17, which, while uncommon, are considered within normal bounds for our analysis as they do not represent statistical outliers based on our criteria of frequency and impact on model accuracy.

A similar pattern is observed for the `'number_of_casualties'` with a left-skewed distribution, suggesting that most collisions result in a single injury. The distribution peaks sharply at one, with a significant drop-off for higher values. The maximum value recorded is 70, an extreme case that significantly deviates from the norm. Given the rarity of such incidents and their potential to disproportionately influence statistical models, we consider removing the only outlier from the dataset to ensure a more accurate and representative analysis of typical road traffic collisions.

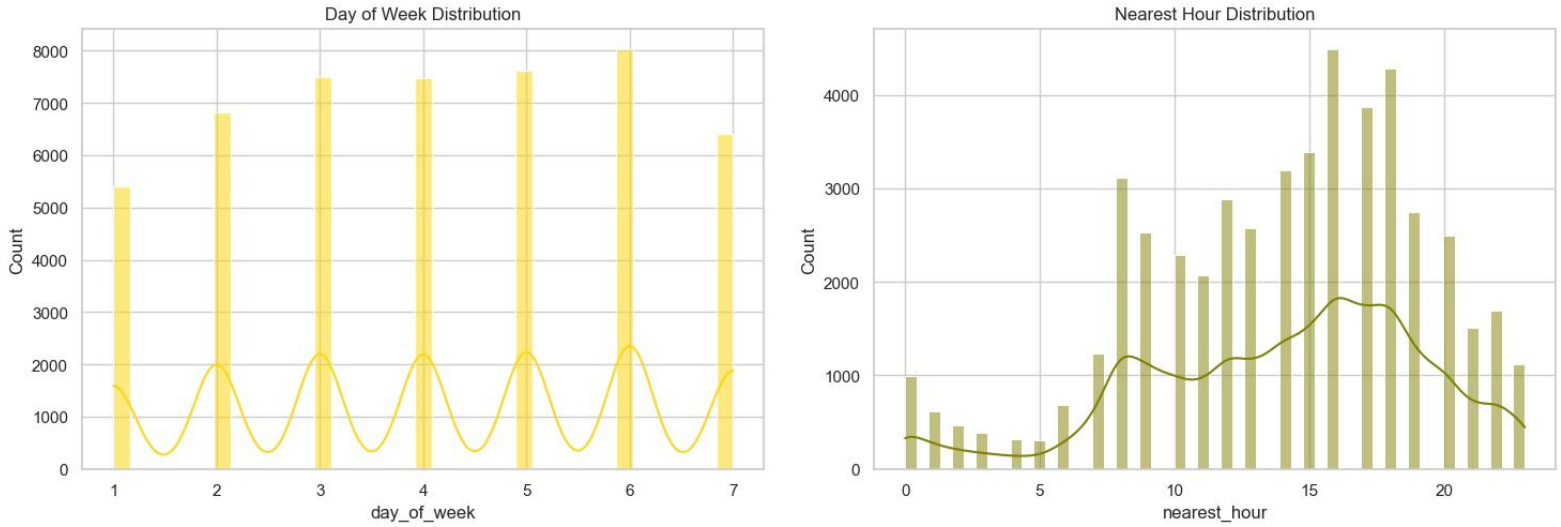

Figure 3: Distribution of Days of Week and Nearest Hour

The most frequent 30 mph speed limit often indicates urban areas with higher traffic density, increased pedestrian activity, and complex road layouts, potentially leading to a higher number of accidents. Interestingly, from Figure 3, the plot reveals that the number of accidents does not peak during extreme weather conditions or periods of poor visibility. Instead, most accidents occur in fine weather, under normal daylight conditions, and on dry roads. In terms of timing, the distribution of accidents is relatively even across the weekdays, with slight reductions on Mondays and Sundays, which might be related to lower traffic volumes at the start and end of the week. There is a noticeable peak in accidents from 10 am to 8 pm, aligning with busy daytime hours when traffic volumes are higher due to various daily activities such as commuting, working time, and school runs.

Our analysis of the variables with specific definitions reveals notable skewness in distribution across several categories, highlighting the dominance of certain road features in accident reports. A significant observation is that many accidents occur at locations not controlled by traffic signals or authorized personnel, suggesting a higher risk of collisions potentially due to increased likelihood of driver error in these uncontrolled settings. Specifically, the category *Not at or within 20 meters of a junction* from `'junction_detail'` indicates that a majority of incidents occur away from junctions. This suggests that straight roads or stretches without intersections pose higher risks, likely due to higher speeds or less attentive driving. Additionally, *T or staggered junctions* or *staggered junctions* and *Crossroads* are identified as common sites for accidents. These junction types, characterized by their complex traffic flows and multiple conflict points, require careful navigation, which may not always occur, leading to a higher incidence of collisions. The predominance of accidents on single carriageways, which typically do not have separation between opposing streams of traffic, further supports the need for targeted safety measures in these areas.[1]

Considering these insights, interventions such as the installation of clearer signals, enhanced road markings, or even traffic lights at key locations could significantly mitigate risks. Further quantitative analysis could refine these recommendations by identifying specific factors that significantly contribute to higher accident rates in these areas.

[1]The remaining distribution plots for defined variables are attached at the end.

Preprocessing

Although there are no missing values found by applying `isnull()`., during the initial data review, we identify interpretations of the minimal value(-1) and maximum value (99) which are undefined for the defined variables. We firstly consider excluding rows with these -1 values to maintain data integrity; however, 6312 out of the total 49,316 observations are affected(approximately 12.8% of the data). The full removal may simplify the dataset but would reduce the dataset size significantly, leading to the missingness of valuable information and biased results. Therefore, we decided to continue with the technique of imputation with invalid values that replaces the invalid values with the median value of respective columns.

To better align with our analysis goal of analyzing pedestrian road safety, we modify two original features `'casualty_class'` and `'casualty_severity'`. We create `'casualty_pedestrian'` to isolate pedestrian cases (1 for pedestrians, 0 otherwise) and `'casualty_over_serious'` for categorizing accidents as either serious or fatal (1) versus slight (0). Severity of a casualty refers to whether the casualty was killed, seriously injured or slightly injured and usually refers to the severity of the most severely injured casualty.

Furthermore, to prioritize the focus of severe accidents involving pedestrians, we introduce an interaction term `'pedestrian_over_serious'` by multiplying the two created dummy variables. This interaction term allows us to specifically target and analyze the most critical subset of accidents, enhancing our ability to draw relevant conclusions about factors affecting pedestrian safety in severe incidents.

Research Design

In this study, we analyze road safety data from reported collisions in the UK in 2023, focusing on pedestrian-related accidents. We employe Random Forest and XGBoost models, chosen for their advantages: Random Forest for its robustness in handling complex patterns without needing feature scaling and its effectiveness in managing outliers, and XGBoost for its proficiency in addressing data imbalance through mechanisms like higher weighting of misclassified instances and regularization to prevent overfitting.

Despite their strengths, both models present limitations, particularly in scenarios with extreme class imbalances (e.g., 2823 severe to 46409 non-severe pedestrian accidents). To mitigate this, we consider adjusting the class_weight parameter and employing the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic examples from the underrepresented class, ensuring a balanced dataset and enhancing model accuracy. The computational demands of these models increase with the complexity of their hyperparameter settings, which could impact overall performance.

The analysis is designed to target three specific outcomes: (1) whether an accident involved a pedestrian, (2) whether an accident resulted in serious injuries or fatalities, and (3) whether a serious collision involved a pedestrian. For each target feature, we train Random Forest and XGBoost models, evaluating them with precision, accuracy, and ROC-AUC—metrics chosen for their ability to reflect both the correctness and discriminative power of the models aiming to identify the most accurate models for these outcomes. Additionally, we analyze the models' feature importance outputs to identify risk characteristics associated with increased pedestrian danger, providing insights to inform policy adjustments and ultimately improve road safety for pedestrians.

We divide the data into an 80% training set and a 20% test set, using the stratify parameter to ensure that class proportions in the training and testing datasets mirrored the original distribution. This strategy helps evaluate the models' ability to generalize to new data. After splitting, SMOTE is applied to the training data to address further the issue of class imbalance, ensuring the models were trained on a representative sample of the diverse scenarios present in the actual data.

Results Interpretation

Table 3: Accuracy of Models Across Different Targets

		Accuracy	
Model	Target Feature	Pre-tuning	After-tuning
Random Forest	Pedestrian	91.34%	91.71%
	Severity Level	65.24%	65.43%
	Interaction	87.80%	87.89%
XGBoost	Pedestrian	91.89%	92.06%
	Severity Level	62.46%	62.70%
	Interaction (original)	84.93%	84.53%
	Interaction (updated)	-	87.81%

Our analysis targete three crucial aspects: pedestrian involvement, severity of accidents, and severe accidents involving pedestrians. For pedestrian involvement (Target 1), both Random Forest and XGBoost demonstrate high precision and recall for non-pedestrian-involved cases (class 0), intensifying their effectiveness in identifying more frequent scenarios. Notably, post-tuning improvements in XGBoost are significant, enhancing its capability to handle imbalanced data, which lead to better precision and recall for pedestrian-involved accidents (class 1).[2]

In predicting the severity of accidents (Target 2), both models show moderate success in non-severe cases but struggle with severe cases (class 1), even though tuning provide slight enhancements. The inherent complexity contributing to severe accidents leads to lower accuracy and ROC_AUC values, suggesting the difficulty in the prediction task and the need for further refinement in feature selection of more relevant variables and updated model tuning.

The analysis of severe accidents involving pedestrians (Target 3) poses significant challenges, particularly in accurately predicting class 1 scenarios. Despite the high overall accuracy of around 90% for both models, indicating effective class differentiation, the performances remaine substandard due to the complexity of the category. The models find it difficult to achieve a balance between precision and recall, with hyperparameter optimization leading to only marginal improvements. The different hyperparameter settings for XGBoost to obtain higher accuracy further highlight the challenges in obtaining satisfactory performance, suggesting that high accuracy alone does not necessarily reflect a well-performing model in this context.

Our horizontal comparison indicate that post-tuning generally enhanced model metrics such as F1-Score, precision, and recall across both models for most classes and targets. XGBoost marginally outperforms Random Forest, particularly in managing more complex or imbalanced classes, due to its

[2]The full classification reports of total 6 models including all metrics are attached at the end.

advanced tuning capabilities and built-in mechanisms for class weight adjustment and regularization. The more elaborate and varied hyperparameter range in XGBoost contribute to its better performance, particularly in predictions for class 1 scenarios post-tuning. Under the objective for higher accuracy, the Random Forest model would be selected for most targets.

As discussed in the model justification, predictive modeling of severe pedestrian accidents face significant hurdles due to the complexity and severe imbalance in class distribution. These challenges are intensified by the low precision and recall in class 1 predictions, despite employing techniques such as SMOTE to address class imbalance.

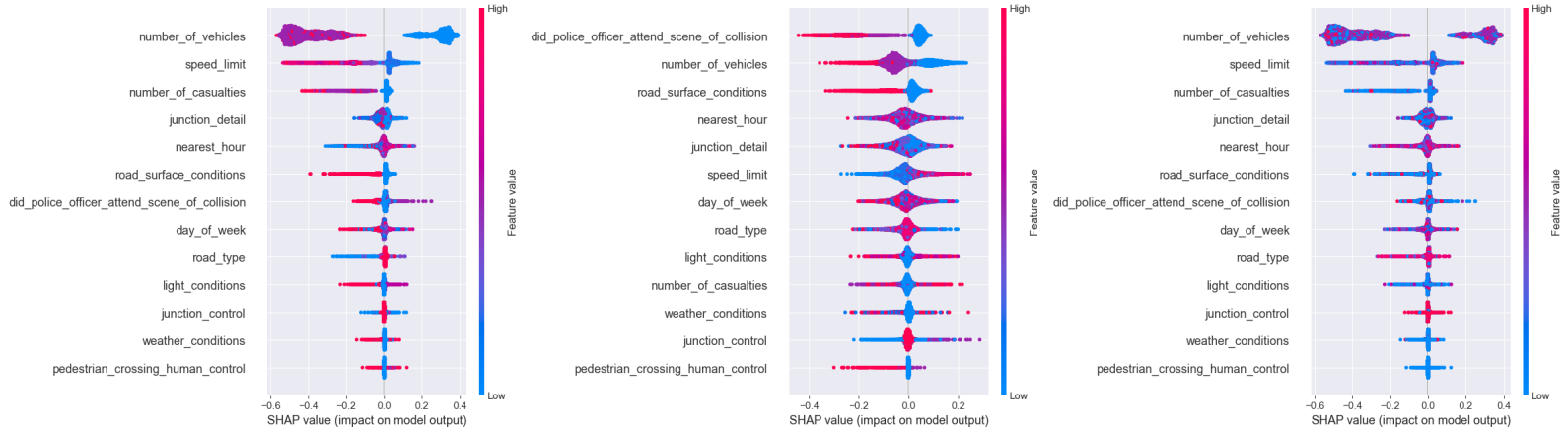

Figure 4: SHAP Plot of Random Forest Models

The SHAP plots for Random Forest and XGBoost models reveal the significant influence of features like the number of vehicles and speed limit on accident severity predictions, with increases in either leading to higher severity classifications. Notably, features such as road surface and light conditions also impact predictions, with poor conditions and nighttime settings associated with higher accident severity. XGBoost demonstrates a relatively broader spread of influential features, indicating its capacity to capture more complex interactions and a greater sensitivity to variations in feature values. For instance, the presence of police action at scene, prominently affecting XGBoost's predictions, suggests more severe incidents are often attended by law enforcement.

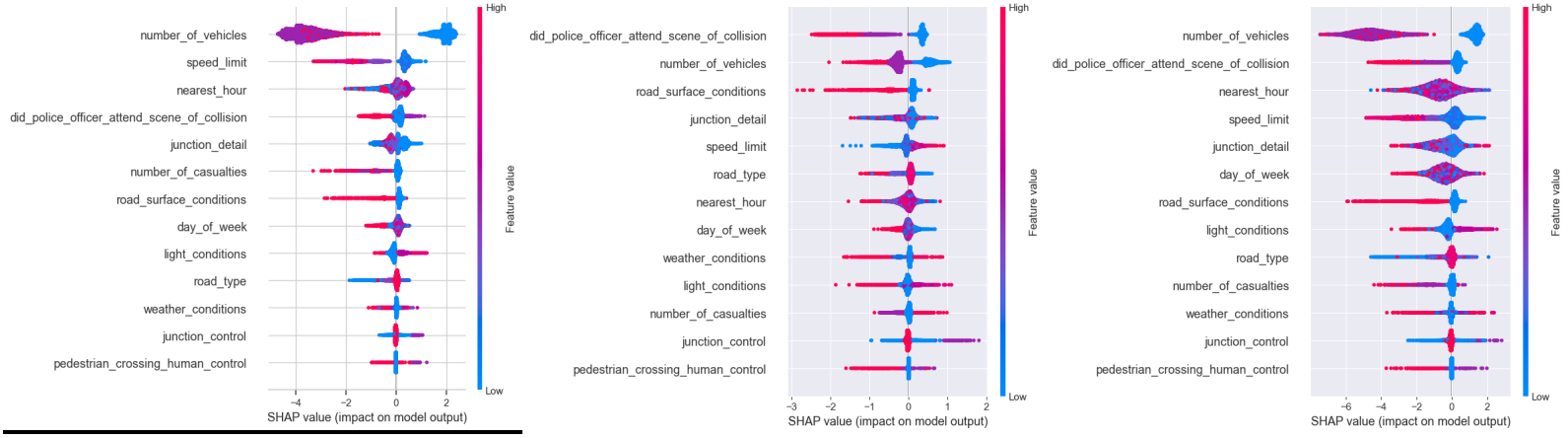

Figure 5: SHAP Plot of XGBoost Models

The insights from this study highlight the critical importance of model selection and hyperparameter tuning in managing complex, real-world datasets typical in road safety analysis. XGBoost, with its complex tuning options and its ability to handle imbalanced classes, proved to be slightly more effective than Random Forest, particularly following hyperparameter adjustments to obtain better model performances. These findings underline the necessity of employing advanced machine learning techniques and continuous model optimization to enhance predictive accuracy for rare but critical outcomes like severe pedestrian accidents. This study also emphasizes the ongoing need for further research and methodological advancements to improve the capabilities of predictive models in traffic safety applications, ensuring a balance between precision and recall with more comprehensive hyperparameter configurations. Maintaining this balance is crucial in the context of road safety, where inaccuracies in predicting severe accidents can lead to significant resource misallocation—either by over-predicting, which could waste valuable emergency response resources, or under-predicting, which might fail to prevent serious harm, especially when targeting collision severity level, the police enforcement is considered to be the most significant factor here. Therefore, optimizing this balance through comprehensive hyperparameter configurations is essential for effective and efficient road safety management.

Conclusions

This study of 2023 Road Safety Data from the UK underscores the critical importance of model selection and hyperparameter tuning in managing complex, real-world datasets typical in road safety analysis. XGBoost, with its sophisticated tuning options and capability to handle imbalanced classes, proved slightly more effective than Random Forest, particularly after hyperparameter adjustments. The SHAP analysis highlighted that certain features such as the number of vehicles, speed limits, road surface conditions, and lighting significantly influence accident severity predictions. These insights can be instrumental in developing targeted safety measures and policy adjustments, particularly for pedestrian safety.

Implications for Policy Adjustments

The findings offer actionable insights for enhancing pedestrian safety through specific policy measures. For instance, implementing stricter speed regulations in high-accident areas and improving road infrastructure, such as better lighting and road surface maintenance, could significantly reduce accident severity. Furthermore, prioritizing resource allocation during high-risk periods, as identified by the models, could enhance preventive measures and emergency response effectiveness.

Further Research and Directions

While this study provides foundational insights, further research is needed to explore additional variables that could affect road safety, such as temporal variations in pedestrian behavior and environmental factors. For example, studies like that of Gerogiannis and Bode, which examined pedestrian behavior at unmarked crossings, suggest that local conditions and behavioral patterns play a crucial role in accident rates and should be considered in road safety designs.

The analysis by Marchant, Hale, and Sadler, which found an association between increased lighting and higher accident rates, calls for a nuanced understanding of how lighting and other interventions might interact to influence road safety. This contradiction highlights the complexity of road safety issues and suggests that a multifaceted approach, combining policy adjustments, technological enhancements, and public health strategies, is essential.

Furthermore, the critique by Staton et al. (2022) regarding the lack of progress in reducing road deaths in the UK points to the need for robust leadership and comprehensive strategies at both national and municipal levels. It emphasizes that improving road safety requires a holistic approach that includes not only data-driven insights but also considerations of governance and community engagement.

Limitations and Challenges

Hyperparameter tuning is a critical aspect of optimizing machine learning models, but it presents several challenges. As we acknowledged before the utilization of Random Forest model and XGBoost model, the process of selecting the best combination of hyperparameters is time-consuming and computationally complex. In the analysis, the hyperparameter settings used may not have been exhaustive enough to achieve the highest possible accuracy. We observed instances where the tuned model exhibited lower accuracy compared to the original model, which is not an anticipated outcome. To address this issue and validate our findings, we conducted additional experiments with different hyperparameter ranges with gridsearch. While the approach yielded slight improvements, the results were not consistent across all settings. Furthermore, when optimizing models for different target variables to achieve our goal for better accuracy, we employed different thresholds, which introduces challenges in comparing the models' performance between each other.

Another significant limitation of this study comes from the nature of the dataset used. The collision and casualty data, obtained from the Department of Transportation of the UK government, only include personal-injury collisions reported to the police. This data collection method may lead to under-reporting and under-recording of accidents due to various factors. For example, some collisions may not be reported if no injury occurs or if the parties involved reach a compromise. Additionally, some drivers may fail to report collisions due to ignorance of legal requirements or reluctance, particularly if they were under the influence of alcohol. These factors can result in an incomplete dataset, potentially affecting the accuracy and generalizability of the analysis.

Moreover, the characteristics of the collision data pose challenges in maintaining a comprehensive set of related variables. Unlike the casualty data, which contains repeated collision references, each person involved in a collision provides a separate report. This makes it difficult to aggregate sociodemographic factors, such as age and gender, for each case. The absence of complete sociodemographic information may limit the depth of the analysis and the ability to draw insights based on these factors. To address these limitations, future research could explore alternative data sources complementing the existing dataset. Additionally, employing advanced data imputation techniques or conducting sensitivity analyses could help mitigate the impact of missing or incomplete data. Furthermore, researchers could investigate the use of more advanced hyperparameter tuning techniques, such as Bayesian optimization to improve the range and effectiveness of the current tuning process.

References

[1] Choudhary, J. K., Rayala, N., Kiasari, A. E., & Jafari, F. (2023). *Road Safety in Great Britain: An Exploratory Data Analysis*. International Journal of Transport and Vehicle Engineering, 17(7), 273-274.

[2] Department for Transport. (2023). Road Safety Data - Vehicles Collision Provisional Unvalidated Mid-year 2023 [Data set]. Retrieved from https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

[3] Department for Transport. (2023). Road Safety Data - Casualty Provisional Unvalidated Mid-year 2023 [Data set]. Retrieved from https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

[4] Gerogiannis, A., & Bode, N. W. F. (2024). *Analysis of Long-term Observational Data on Pedestrian Road Crossings at Unmarked Locations*. Safety Science, 172, 106420. https://doi.org/10.1016/j.ssci.2024.106420

[5] Marchant, P., Hale, J. D., & Sadler, J. P. (2020). *Does Changing to Brighter Road Lighting Improve Road Safety? Multilevel Longitudinal Analysis of Road Traffic Collision Frequency During the Relighting of a UK City*. Journal of Epidemiology and Community Health, 74(5), 467–472. https://doi.org/10.1136/jech-2019-212208

[6] Roberts, G. (2024, January 29). Safety concerns over Highway Code changes two years after being introduced. *Fleet News*. Retrieved from https://www.fleetnews.co.uk/news/safety-concerns-over-highway-code-changes-two-years-after-being-introduced

[7] Staton, M., Barnes, J., Morris, A., & Waterson, P. (2022). 'Over to You': Using a STAMP Control Structure Analysis to Probe Deeper into the Control of UK Road Safety at a Municipal Level – The Case of Cambridgeshire. *Ergonomics*, 65(3), 429–444. https://doi.org/10.1080/00140139.2021.1968033

[8] Sufian, M. A., Varadarajan, J., & Niu, M. (2024). Enhancing Prediction and Analysis of UK Road Traffic Accident Severity Using AI: Integration of Machine Learning, Econometric Techniques, and Time Series Forecasting in Public Health Research. *Heliyon*, 10(7), e28547. https://doi.org/10.1016/j.heliyon.2024.e28547

Table 4: Evaluation of Pedestrian Predictive Models

	Random Forest Model		XGBoost Model	
Metric	Pre-tuning	Post-tuning	Pre-tuning	Post-tuning
Class(0)				
Precision	0.97	0.97	0.97	0.98
Recall	0.92	0.93	0.93	0.93
F1-Score	0.95	0.95	0.95	0.95
Support	8071	8071	8071	8071
Class(1)				
Precision	0.72	0.72	0.72	0.73
Recall	0.86	0.87	0.89	0.90
F1-Score	0.78	0.79	0.80	0.80
Support	1776	1776	1776	1776
Accuracy	91.34%	91.71%	91.89%	92.10%
ROC_AUC	0.940	0.944	0.953	0.955

Table 5: Evaluation of Severity Level in Predictive Models

	Random Forest Model		XGBoost Model	
Metric	Pre-tuning	Post-tuning	Pre-tuning	Post-tuning
Class(0)				
Precision	0.80	0.80	0.83	0.83
Recall	0.73	0.73	0.64	0.64
F1-Score	0.76	0.76	0.72	0.73
Support	7538	7538	7538	7538
Class(1)				
Precision	0.31	0.32	0.32	0.33
Recall	0.41	0.41	0.56	0.57
F1-Score	0.36	0.36	0.41	0.42
Support	2309	2309	2309	2309
Accuracy	65.24%	65.43%	62.46%	62.70%
ROC_AUC	0.621	0.621	0.625	0.659

Table 6: Evaluation of Interaction Models

	Random Forest Model		XGBoost Model	
Metric	Pre-tuning	Post-tuning	Pre-tuning	Post-tuning
Class(0)				
Precision	0.97	0.97	0.98	0.96
Recall	0.90	0.90	0.86	0.90
F1-Score	0.93	0.93	0.91	0.93
Support	9282	9282	9282	9282
Class(1)				
Precision	0.23	0.23	0.23	0.22
Recall	0.47	0.49	0.67	0.46
F1-Score	0.31	0.32	0.34	0.30
Support	565	565	565	565
Accuracy	87.80%	87.89%	84.93%	87.81%
ROC_AUC	0.849	0.859	0.866	0.838

Table 7: Comparison between Different Hyperparameter Settings

	XGBoost (Original setting)	**XGBoost (New setting)**
Class(0)		
Precision	0.98	0.96
Recall	0.86	0.90
F1-Score	0.91	0.93
Support	9282	9282
Class(1)		
Precision	0.22	0.22
Recall	0.68	0.46
F1-Score	0.33	0.30
Support	565	565
Accuracy	84.53%	87.81%
ROC_AUC	0.8672%	0.8382%

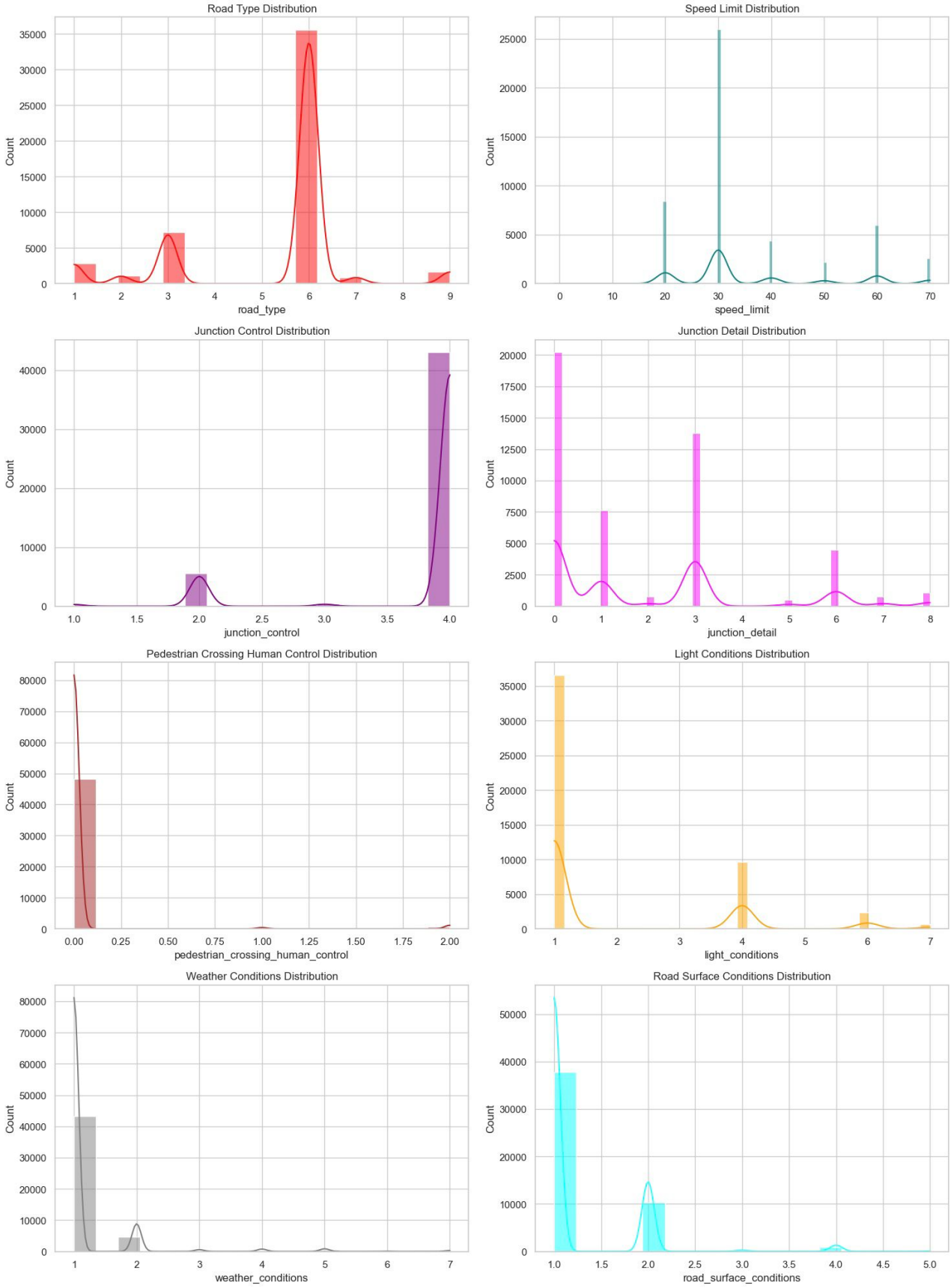

Figure 6: Distribution Plots for Remaining Variables