IRF World Congress 2024

Enhancing cyclist safety: Predictive analysis of injury severity and advocacy for evidence-based interventions

Virginia Petraki[1*], Stella Roussou[1], Apostolos Ziakopoulos[1], George Yannis[1]

[1]National Technical University of Athens, Department of Transportation Planning and Engineering, 5 Heroon Polytechniou str., Athens, GR-15773, Greece

*Corresponding author and presenter: vpetraki@mail.ntua.gr

Abstract

Cycling has emerged as a popular mode of transportation, promoting health and sustainability. However, cyclists face various risks, including crashes involving motor vehicles. Moreover, the growing prevalence of urban cycling as a response to traffic congestion and air pollution underscores the urgent need to address safety concerns and enhance infrastructure for cyclists sharing the road with motorized vehicles (Karanikola P, Cycling as a Smart and Green Mode of Transport in Small Touristic Cities., 2018). Understanding the factors contributing to cyclist injuries is crucial for enhancing road safety and promoting sustainable mobility. In this study, data from Bicycle Crashes in Great Britain from 1979 to 2018 are analysed to investigate the relationship between several factors and the severity of cyclist injuries. Specifically, the focus is on critical factors such as age group, day of the week, speed limits (categorized as under 30 km/h, between 30-50 km/h, and above 50 km/h), number of vehicles involved, weather conditions, road type, light conditions, road conditions, gender, and number of casualties. The supervised machine learning algorithm CatBoost is used to predict cyclist injury severity based on these factors. The feature importance analysis revealed that age group is the most significant predictor, followed by the day of the week and speed limits. Other important factors include the number of vehicles involved, weather conditions, road type, and light conditions. Lower speed limits, particularly under 30 km/h, are associated with reduced cyclist injuries, while adverse weather and poor road conditions significantly increase the severity of injuries. These findings underscore the importance of comprehensive safety strategies that consider various factors influencing crash severity. The implications of the study extend beyond academic research, informing policymakers, urban planners, and transportation authorities about the critical role of speed limit regulations in promoting cyclist safety. Advocacy for evidence-based measures, such as age-specific safety programs, improved infrastructure, and effective speed management strategies, is essential for enhancing cyclist protection. In conclusion, this analysis underscores the significance of integrating speed management measures into broader road safety strategies to create safer environments for cyclists and promote sustainable urban mobility.

Keywords: Cyclist injuries, Speed limits, Supervised machine learning, Speed management strategies, Vulnerable road users.

1. Introduction

Understanding and mitigating the severity of cyclist crash incidents is a complex task that involves extensive research and data analysis. Efforts to reduce crash severity focus on identifying and analyzing the factors that contribute to serious injuries and fatalities in road crashes are being explored. This includes the study of various elements such as road conditions, vehicle characteristics, and driver behavior. By leveraging advanced analytical techniques and machine learning models, researchers aim to predict and mitigate crash outcomes, thereby enhancing road safety and reducing the impact of traffic incidents on public health.

The phenomenon of crash injuries, particularly among vulnerable road users like cyclists, has attracted significant attention in recent years. As cycling becomes an increasingly popular mode of transportation due to its health and environmental benefits, the need to address cyclist safety has become paramount. Cyclists are at a higher risk of severe injuries due to their lack of physical protection compared to motor vehicle occupants. Factors such as vehicle speed, road conditions, weather, and visibility play crucial roles in the severity of cyclist injuries. In addition, the interaction between cyclists and motor vehicles in urban settings introduces further complexity, necessitating comprehensive safety measures.

The objective of this paper is to examine the determinants influencing the severity of cyclist injuries in road accidents within Great Britain. By utilizing data collected from 1979 to 2018, this study applies the CatBoost machine learning algorithm to evaluate the significance of various contributory factors. The identification of the most critical predictors of injury severity aims to provide valuable insights that can guide policy formulation and enhance safety measures for cyclists. This research contributes to the academic understanding of road safety and offers practical recommendations for policymakers, urban planners, and transportation authorities to strengthen cyclist protection and foster sustainable mobility.

The paper is structured as follows: Section 2 reviews the existing literature on crash severity and cyclist safety. Section 3 describes the methodological background. Section 4 details the data collection, preprocessing steps and the data analysis process. Section 5 presents the application of CatBoost algorithms and the results of the analysis, highlighting the most significant factors influencing injury severity. Section 6 discusses the implications of the findings and provides recommendations for improving cyclist safety. Finally, Section 7 concludes the paper and suggests directions for future research.

2. Literature Review

Cyclist safety and injury severity have been critical areas of research in transportation and urban planning. Numerous studies have explored various factors influencing cyclist injuries and the effectiveness of interventions aimed at mitigating crash severity. This review synthesizes key findings from existing literature to provide a comprehensive understanding of the determinants of cyclist injury severity and the methodologies used to study this phenomenon.

A study using a binary regression model to analyze factors affecting the severity of cycling crashes has identified that road conditions, weather, and lighting significantly influence the likelihood of severe injuries. Their findings emphasize the need for improved infrastructure and better visibility measures to enhance cyclist safety (Jaber et al. 2021)

Focus on the vulnerability of cyclists on the road, particularly in relation to the type of vehicle involved and driver culpability is very important (García-Herrero et al., 2019). Their probabilistic analysis highlighted that heavy vehicles, such as trucks and buses, pose a higher risk of causing severe injuries to cyclists. The study also found that drivers are more likely to be held responsible for accidents involving cyclists, suggesting the need for stricter regulations and training for drivers of larger vehicles (García-Herrero et al., 2019).

A case-control study examined cycling injury risks in London, identifying critical factors such as road design and traffic density (Mustafa Ekmekci, 2024). They found that intersections and areas with high traffic volumes are hotspots for cycling accidents, indicating the importance of targeted safety measures in these areas. This study also underscored the role of urban planning in mitigating cycling risks by incorporating cyclist-friendly infrastructure (Aldred R, 2018).

The effectiveness of various safety measures for cyclists in a meta-analysis revealed that dedicated cycling lanes, traffic calming measures, and helmet use significantly reduce the severity of injuries in cycling accidents. The study advocates for comprehensive safety programs that integrate these measures to protect cyclists effectively (Ahmed Jaber, 2021).

These studies collectively highlight the important nature of cycling safety, involving factors ranging from infrastructure and vehicle types to environmental conditions and driver behavior. The consistent theme across the literature is the critical need for integrated safety measures that address the diverse risks faced by cyclists. Enhanced infrastructure, stringent vehicle regulations, and comprehensive urban planning are essential components of effective cyclist safety strategies.

3. Methodological background

The variable of interest in the present analysis is the severity of cyclist injuries. To investigate the factors affecting injury severity among cyclists in Great Britain, detailed data from various sources spanning from 1979 to 2018 were analyzed using advanced machine learning techniques. Specifically, CatBoost algorithms were employed to determine the importance of numerous contributing factors in predicting injury severity. This robust ensemble method is particularly effective in handling categorical features and missing values, making it well-suited for the complexity of traffic incident data.

3.1 CatBoost

Gradient Boosted Decision Trees (GBDT’s) are a powerful tool for classification and regression tasks in Big Data (Hancock J. K., 2020). CatBoost is a member of the family of GBDT machine learning ensemble techniques. Since its debut in late 2018, researchers have successfully used CatBoost for machine learning studies involving Big Data. Furthermore, as a Decision Tree based algorithm, CatBoost is well-suited to machine learning tasks involving categorical, heterogeneous data. Another important issue of CatBoost is its sensitivity to hyper-parameters and the importance of hyper-parameter tuning (Hancock J. K., 2020).

For the analysis, CatBoost algorithms were employed. CatBoost, short for categorical boosting, is a powerful supervised ML algorithm developed by Yandex, specifically designed to handle categorical features effectively (Hancock J. K., 2020). CatBoost is based on gradient boosting of decision trees and uses one-hot encoding to handle categorical data. Like XGBoost, CatBoost also encompasses multiple Classification and Regression Trees (CART). Its adaptability and effectiveness have made it a top performer in numerous ML competitions (Liudmila P, 2018).

CatBoost is designed for increased speed, accuracy, and ease of use (Liudmila P, 2018). The core boosting technique in CatBoost is based on the superimposition of new tree models in the errors and residuals of previous models. The tree ensemble is then combined to reach the final prediction. The loss function of CatBoost includes two terms: (i) a training loss term and (ii) a regularization term to control model complexity and prevent over-fitting (Dorogush, 2018); Li et al., 2019). In essence, CatBoost applies a mapping function between variables, where a regression tree ensemble model uses a number of functions K additively to predict y, so that (Liudmila P, 2018):

$$\hat{y} = \varphi(x_i) = \sum_{k=1}^{K} f(x_i) \qquad \text{Eq. (1)}$$

Where $\hat{y}$ is the predicted value of the original dependent (or response) variable y and $x_i$are the independent (or explanatory) n variables across i observations. The loss function expresses the distance between training data and predicted values and is defined as $l(\hat{y}_i, y_i)$. A common choice of l is the mean squared error for a set of parameters φ_i (JH, 2001):

$$l(\varphi_i) = \sum_{\iota=1}^{I} (\hat{y}_i - y_i)^2 \qquad \text{Eq. (2)}$$

A penalizing term, $\Omega(f)$, is also introduced for model complexity control such that:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|c\|^2 \qquad \text{Eq. (3)}$$

Where γ, λ are penalizing coefficients, T is the number of leaves in the regression tree. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf, creating a flowchart, and c is the weight assigned to each leaf. Having obtained the loss function, $l(\hat{y}_i, y_i)$, and the penalizing term, $\Omega(f)$, the objective function can be formulated as:

$$L(\varphi_i) = \sum_{i=1}^{I} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f) \qquad \text{Eq. (4)}$$

As with most ML methodologies, CatBoost features a number of tunable model hyperparameters that can be optimized before or during cross-validation of results, such as (CatBoost, 2022):

- Learning rate: Governs the magnitude of iterations for minimizing the cost function.
- Depth: Governs the maximum depth of the tree.
- L2_leaf_reg: L2 regularization term on weights.
- Border count: Number of splits for numerical features.
- Random strength: Controls the intensity of randomness when scoring splits.

Following good ML practices, the hyperparameters of CatBoost algorithms should be tuned initially before their final executions, and their predictions should subsequently be evaluated with model evaluation metrics. The highly non-linear, tree ensemble structure of CatBoost makes it resilient against bias from multicollinearity effects (Hancock et al., 2021). For classification algorithms, model performance is evaluated by the predictive performance of each configuration in terms of correct classifications. In binary classification, this is mainly supported by the confusion matrix of the test subset.

4. Data collection

4.1 Dataset description

The dataset used for this analysis encompasses detailed records of bicycle accidents in Great Britain over a period spanning from 1979 to 2018. This extensive dataset was sourced from Kaggle, specifically from the "Bicycle Accidents in Great Britain (1979 to 2018)" collection curated by John Harshith and accessible via Kaggle.

The dataset includes a comprehensive range of variables that provide a detailed account of each bicycle crash. These variables cover different aspects of the crashes, including the specific date and time when the crash occurred (Crash Date and Time, the severity of the crash categorized into different levels such as fatal, serious, and slight injuries (Severity), descriptions of the weather conditions at the time of the crash, which include sunny, rainy, snowy, etc. (Weather Conditions), details about the state of the road, such as dry, wet, icy, etc. (Road Conditions), information about the lighting conditions, for instance, daylight, darkness with or without street lighting (Light Conditions), the number and types of vehicles involved in the crash (Vehicle Involvement). The dataset also contains casualty information including the age group and gender of the cyclists, and the number of casualties.

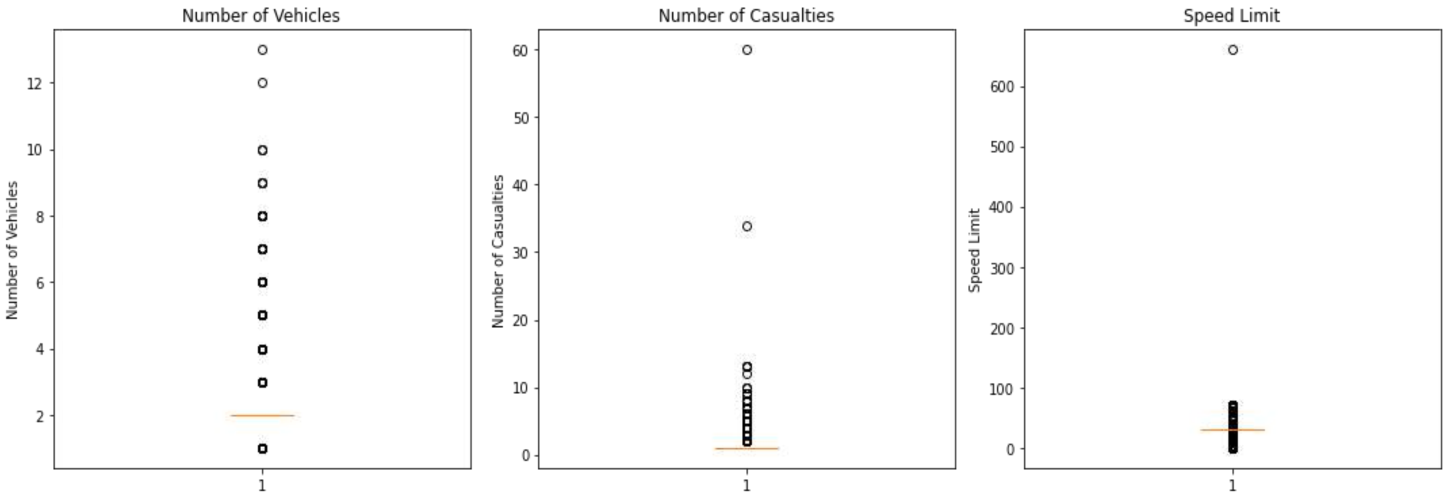
4.2 Dataset pre-processing

The data pre-processing for this research involved several critical steps to ensure the integrity, reliability, and suitability of the dataset for subsequent analysis.

The preprocessing commenced with the merging of two distinct datasets: one containing detailed records of bicycle crashes and another providing specific information about the bicyclists involved in these incidents. The datasets were merged based on a common identifier, "crash number," which uniquely identifies each crash event. This merging step was crucial for creating a comprehensive dataset that integrates both crash-specific and cyclist-specific data, providing a holistic view of each crash.

Identifying and removing outliers is essential to prevent skewed analysis and model performance. Outliers for the variables "number of vehicles," "number of casualties," and "speed limit" were identified. The outliers were defined based on the threshold indicated by the red lines in the provided box plots in the following graph 1. These data points were subsequently removed from the dataset to ensure a more accurate and representative analysis.

Figure 1: Outliers Boxplots



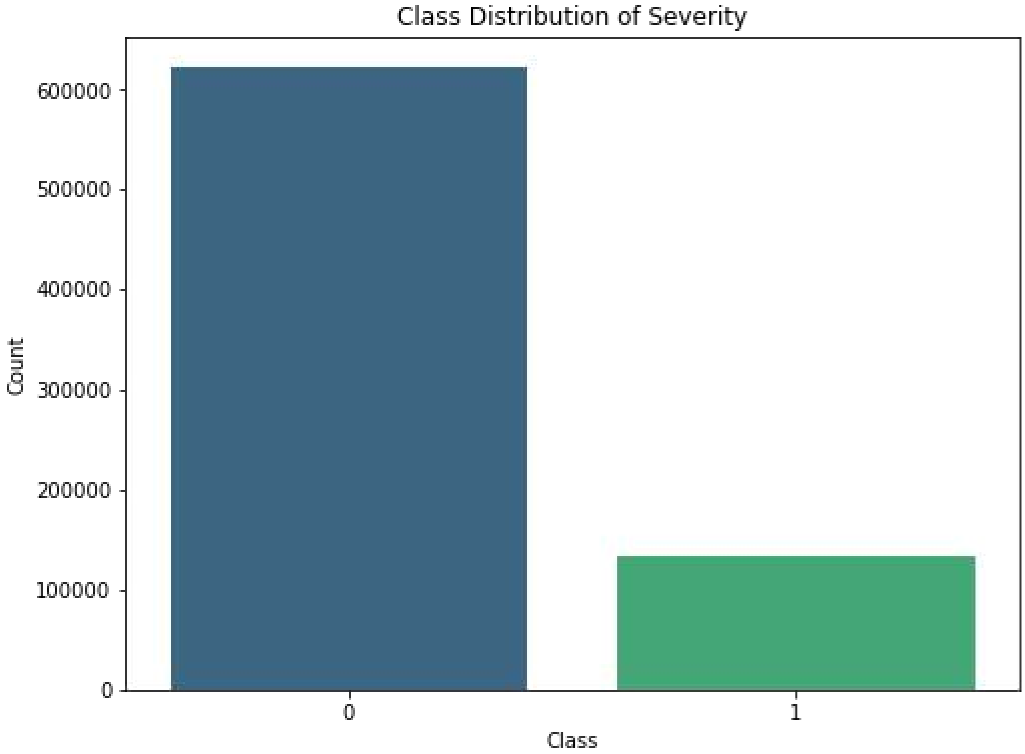
To maintain the dataset's integrity, records with missing values were dropped. This step also included the removal of records with unknown variables. By excluding these incomplete entries, the dataset's quality was enhanced, ensuring that the analysis was conducted on complete and reliable data. This process reduced the dataset from 827,861 entries to 754,636 entries.

The speed limits were grouped into three categories to simplify the analysis and improve the model's performance. The speed limits were categorized Low Speed (Speeds less than 30 mph), Medium Speed (Speeds between 30 mph and 50 mph) and High Speed (Speeds greater than 50 mph).

Categorical variables were encoded to prepare the data for machine learning algorithms. Additionally, the severity labels were adjusted to simplify the classification task. The original severity categories were recoded such that serious and fatal crashes (FSI) were coded as 1, and light injury crashed were coded as 0.

Given the observed class imbalance in the severity of injuries, as depicted in Figure 2, SMOTE (Synthetic Minority Over-sampling Technique) was applied. SMOTE generates synthetic examples for the minority class, thus balancing the class distribution and ensuring that the machine learning model is trained on a balanced dataset. This step is critical for improving the model's ability to accurately predict severe injuries, as it prevents the model from being biased towards the majority class.

Figure 2: Class Distribution of Severity



To address this, while training the CatBoost model, the class weights were adjusted to give more importance to the minority class. Specifically, class weights were set as follows: class_weights: [{0: 1, 1: 1.15}]. This adjustment helps balance the influence of each class during model training without modifying the original data distribution.

After cleaning, the dataset was partitioned into training and test sets, with a 80-20 split, to ensure robust model training and validation. This split was performed while preserving the original distribution of variables using the train_test_split function from the scikit-learn library in Python. This approach ensures that the training and test sets maintain the original distribution of the target variable. Studies such as those by Kohavi (1995) and Sivanandam et al. (2006) have highlighted the importance of an adequate split to prevent overfitting and underfitting, ensuring that the model performs well on unseen data. Kohavi's work on cross-validation techniques emphasizes the need for robust validation methods, with the 80-20 split being a simple yet effective strategy in many practical scenarios.

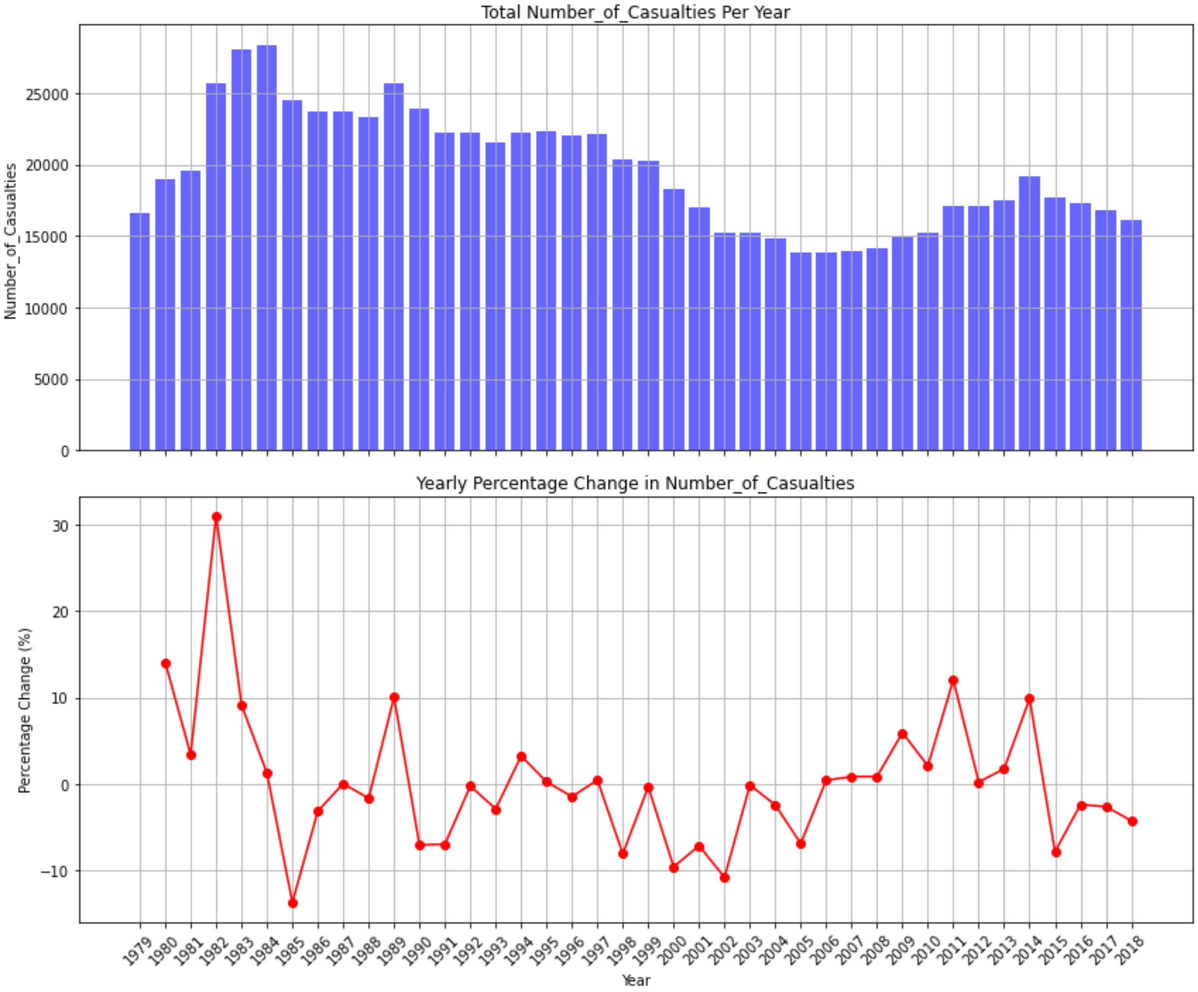
4.3 Dataset Statistics

The following graph presents two plots, where the top plot shows the total number of casualties per year, and the bottom plot depicts the yearly percentage change in the number of casualties.

The top plot, representing the total number of casualties per year, shows a significant initial rise, peaking around the early years of the observed period. This peak indicates a period of high casualties, potentially due to various factors such as increased traffic, insufficient safety measures, or other external influences. Following this peak, there is a noticeable decline in the number of casualties, suggesting improvements in safety measures, traffic management, or policy changes that effectively reduced casualties. However, the decline is not entirely smooth; there are fluctuations indicating periods of both increases and decreases in casualties. In recent years, there is a slight upward trend, suggesting that while overall safety might have improved compared to the peak years, certain factors are causing an increase in casualties again.

The bottom plot, illustrating the yearly percentage change in casualties, shows significant volatility in the early years. This high level of fluctuation suggests that the factors affecting casualties were highly variable, potentially due to rapid changes in traffic patterns, policy implementations, or other situational factors. After the initial volatility, the percentage changes become less extreme, indicating a period of stabilization where year-to-year changes were more consistent and less dramatic. However, there are still notable fluctuations in recent years, which could be due to new influences affecting traffic safety, such as changes in road conditions, enforcement of traffic laws, or other relevant factors.

Figure 3: Number and Percentage Change of Casualties Per Year

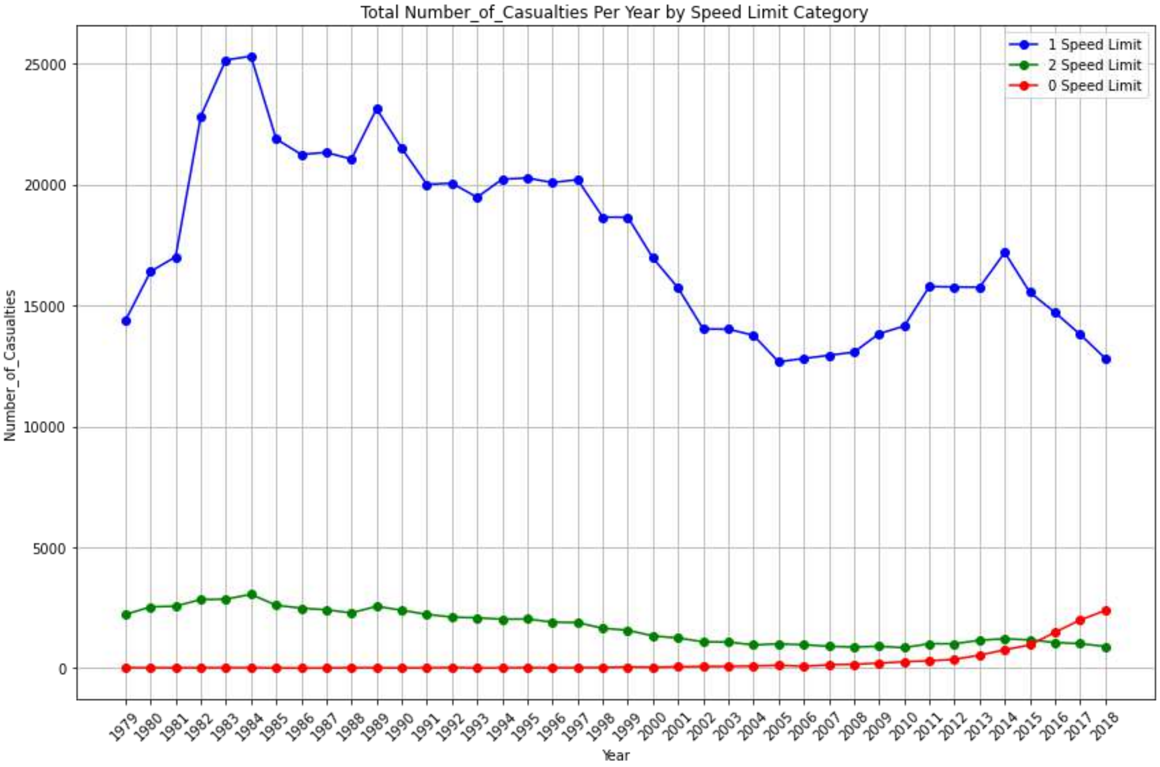


Overall, the initial rise and peak in casualties highlight a period of increased risk, followed by a general decline that points to successful interventions in improving road safety. The recent fluctuations and slight upward trend in both total casualties and percentage changes suggest that while past measures have been effective, ongoing efforts are required to address new and emerging factors contributing to casualties. Targeted interventions, continued monitoring, and adapting to changing conditions will be crucial in sustaining and further improving road safety.

The figure 4 presents the annual number of casualties segmented by three different speed limit categories (labeled 0, 1, and 2). In speed limit category 1 (blue line), the number of casualties shows a sharp increase in the early years, reaching a peak around the middle of the observed period. Following this peak, there is a noticeable decline in the number of casualties, indicating possible improvements in safety measures or changes in traffic policies. Although the overall trend is a decline, there are fluctuations in recent years, suggesting varying factors influencing the number of casualties.

For speed limit category 2 (green line), the trend is more moderate and consistent compared to category 1. The data shows a relatively stable trend with a slight decline over the years, indicating steady improvements or consistent enforcement of safety regulations in this speed limit category. There are fewer fluctuations in this category, suggesting that the factors affecting casualties might be more controlled or predictable. In speed limit category 0 (red line), the number of casualties starts very low, significantly lower than the other two categories. However, in recent years, there is a notable increase in the number of casualties in this category. This increase could be due to various reasons such as changes in traffic patterns, increased reporting, or other external factors. Despite the recent increase, the total number of casualties in this category remains much lower compared to categories 1 and 2.

Figure 4: Number of Casualties Per Year based on Speed Limit Categories



Overall, speed limit category 1 consistently has the highest number of casualties, indicating that areas with this speed limit might require targeted interventions to reduce casualties. The lower number of casualties in speed limit category 2 suggests better safety or more effective control measures in these areas. The increasing trend in category 0 is a cause for concern and warrants further investigation to identify and address the underlying causes.

5. Results

The UK Cyclist Crash dataset was analyzed using CatBoost algorithms, with the results presented in this section. The analyses were conducted using Python, with the main framework following the guidelines provided by the CatBoost development team (2018) and Prokhorenkova et al. (2018). The dataset was split into training and test subsets with an 80-20 ratio, maintaining similar class distributions for the dependent variable. Given the observed class imbalance in the dataset, SMOTE (Synthetic Minority Over-sampling Technique) was applied to ensure that the model had an equal number of samples from each class during training and the class weights were adjusted to improve predictive performance. Specifically, class weights were set as follows: {0: 1, 1: 1.15}.

5.1 CatBoost Results

Hyperparameter tuning with 5-fold cross-validation was carried out to mitigate overfitting and enhance the model's performance. The objective was to determine the internal model configuration that provided the highest classification accuracy. A total of 7 different hyperparameter combinations were tested, randomly chosen from the ranges listed in Table 4. The entire process of selecting, running, and comparing these combinations was automated using Python code. RandomizedSearchCV was used to perform hyperparameter tuning with cross-validation. The search was conducted over a specified parameter distribution, including learning rate, iterations, depth, l2_leaf_reg, border_count, and class weights. The following best parameters were identified through the search. The final optimized hyperparameters are presented in the rightmost column of Table 4.

Table 4: Hyperparameter Tuning Results for CatBoost Model

Hyperparameter	Examined range	Optimized Value
Learning rate	0.01 - 0.29	0.29
Iterations	150 - 450	450
Depth	3 - 11	11
L2 Leaf Reg	1 - 10	10
Border Count	32 - 128	32
Class Weights	{0: 0.5, 1: 50}	{0: 1, 1: 1.15}
Random State	Fixed at 42	42

The manual fine-tuning and optimization were conducted to ensure the best possible performance of the CatBoost model in predicting the severity of road crashes. Once the optimal model was determined, feature importance metrics were extracted for the contributing variables and are shown on Table 5.

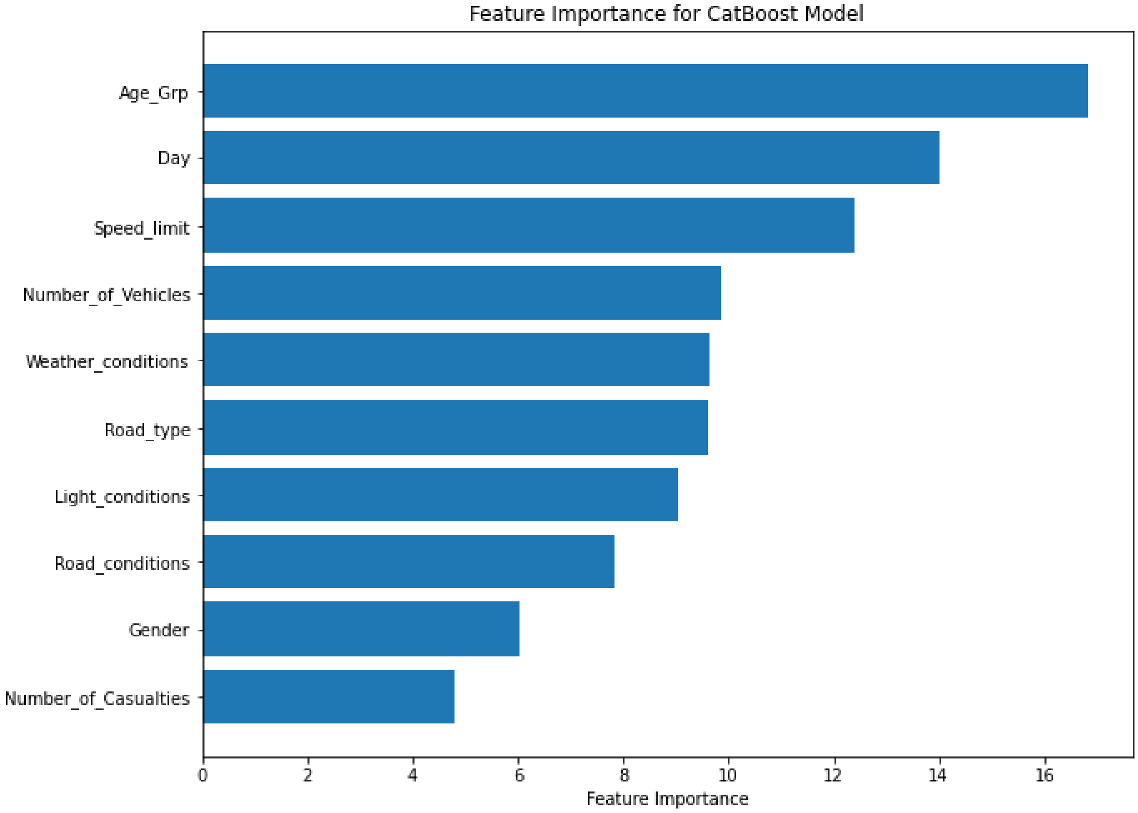
The feature importance analysis provides insights into which variables have the most significant impact on predicting the severity of road crashes. The gain metric, used here, measures the contribution of each feature in improving the model's performance. A higher gain value indicates a more substantial impact on the model's predictions.

Table 5: CatBoost optimized model feature importance

No.	Feature	Gain
1	Age_Grp	16.826
2	Day	13.993
3	Speed_limit	12.386
4	Number_of_Vehicles	9.866
5	Weather_conditions	9.623
6	Road_type	9.593
7	Light_conditions	9.052
8	Road_conditions	7.838
9	Gender	6.026
10	Number_of_Casualties	4.792

The bar plot below represents the feature importance scores from the CatBoost model trained to predict the severity of cyclist injuries. Feature importance measures how much each feature contributes to the model's predictive power. Higher scores indicate that the feature has a more significant impact on the model's performance.

Figure 5: CatBoost Feature Importance Plot



The graph presents the importance of various features in predicting the severity of cyclist injuries using the CatBoost model. The most significant factor is the age group as different age groups have varying levels of vulnerability, with younger and older cyclists being potentially more at risk of severe injuries. The number of casualties is also highly significant, suggesting that incidents involving more casualties tend to be more severe. The speed limit at the crash location is a critical factor, which directly impacts the severity of injuries due to the increased impact force associated with higher speeds. The day of the week is also important, reflecting differences in traffic patterns and cycling activities between weekdays and weekends, which can influence injury severity.

Other significant features include weather conditions, which affect the likelihood and severity of injuries, with adverse conditions like rain or snow increasing the risk. Light conditions also play a crucial role, as poor visibility during nighttime can lead to more severe injuries. The type of road where the crash occurs, such as highways, urban, or rural roads, impacts the severity of injuries due to differing risk levels. Road conditions, such as whether the road is wet, dry, or icy, further affect the severity of crashes. The number of vehicles involved in an incident is another key factor, with multi-vehicle crashes often resulting in more severe injuries due to the complex dynamics and multiple impacts. Lastly, the gender of the cyclist, while lower in importance, still contributes to the model, indicating potential differences in risk exposure or injury outcomes based on gender.

5.2 Model Performance Evaluation

The model's performance on the test set is evaluated using precision, recall, and F1-score metrics for each class, along with overall accuracy. These metrics provide a comprehensive understanding of how well the model distinguishes between the two classes of severity in cyclist injuries.

For class 0 (representing fatal and serious injuries), the model achieves a precision of 0.59, indicating that 59% of the instances predicted as class 0 are correctly identified. The recall for class 0 is 0.60, meaning the model correctly identifies 60% of all actual class 0 instances. The F1-score, which balances precision and recall, is 0.59 for class 0, demonstrating a moderate level of accuracy in predicting severe cyclist injuries.

For class 1 (representing non-injury and light injuries), the precision is 0.59, indicating that 59% of the instances predicted as class 1 are correctly identified. The recall for class 1 is 0.59, meaning the model correctly identifies 59% of all actual class 1 instances. The F1-score for class 1 is 0.59, reflecting balanced performance in predicting less severe injuries.

The overall accuracy of the model is 0.59, as showcased in Table 6, suggesting that the model correctly predicts the severity of cyclist injuries 59% of the time. The macro average for precision, recall, and F1-score is 0.59, providing an unweighted average of the model's performance across both classes. The weighted average indicates that the model maintains a precision of 0.59, recall of 0.59, and F1-score of 0.59 across the dataset.

Table 6: Model Performance Metrics

Metric	**Class 0 (FSI)**	**Class 1 (Non-injury/Light)**
Precision	0.59	0.59
Recall	0.60	0.59

F1-score	0.59	0.59
Accuracy		
Macro avg	0.59	0.59
Weighted avg	0.59	0.59

The confusion matrix, as presented in Table 7, provides a comprehensive overview of the model's performance in predicting the severity of cyclist injuries. The model accurately identifies 74,095 instances where the model correctly predicted class 0 (true positives) and 72,865 instances where the model correctly predicted class 1 (true negatives). However, the model also incorrectly predicted class 1 instead of class 0 in 50,291 instances (false positives) and class 0 instead of class 1 in 51,374 instances (false negatives).

Table 7: Confusion Matrix

	Predicted Class 0	Predicted Class 1
Actual Class 0	74095	50291
Actual Class 1	51374	72865

6. Discussion

The findings from this study underscore several critical factors that significantly influence the severity of cyclist injuries in road accidents. These insights are pivotal for shaping policies and urban planning efforts aimed at enhancing cyclist safety and promoting sustainable urban mobility.

The analysis highlights the crucial role of speed limits in determining injury severity. Lower speed limits, especially those under 30 mph, are associated with a significant reduction in severe injuries among cyclists. This aligns with the established literature, which consistently advocates for stringent speed management as a fundamental aspect of road safety (Elvik and Mysen, 1999). Policymakers should prioritize the implementation and strict enforcement of lower speed limits in urban areas, particularly in zones with high cyclist activity. Such measures not only reduce the impact force in collisions but also enhance overall traffic safety.

The age group of cyclists emerged as another vital factor, indicating that younger and older cyclists are particularly vulnerable to severe injuries. This finding suggests the necessity for age-specific safety interventions. For instance, targeted educational programs and campaigns can be designed to raise awareness about safe cycling practices and the importance of protective gear. Implementing such programs could effectively mitigate the heightened risks faced by these vulnerable age groups. Similar recommendations have been supported by studies emphasizing the varying physical and perceptual capabilities across different age demographics (García-Herrero et al., 2019).

Environmental conditions, including weather and lighting, also play a significant role in influencing injury severity. Adverse weather conditions, such as rain or snow, and poor visibility during nighttime substantially increase the risk of severe injuries. Therefore, improving cycling infrastructure to withstand various weather conditions is imperative. Enhancing street lighting to ensure better visibility can further safeguard cyclists. These infrastructural improvements are critical in mitigating the risks associated with environmental factors, as supported by prior research on the impact of weather and lighting on cyclist safety (Cossalter et al., 2018).

In addition to environmental and demographic factors, road type and conditions were found to significantly impact injury severity. Highways and poorly maintained roads pose higher risks to cyclists, often resulting in severe injuries due to higher vehicle speeds and complex dynamics in multi-vehicle crashes (Fraboni et al., 2019). Thus, investing in the maintenance and improvement of road quality, as well as implementing dedicated cycling lanes on both highways and urban roads, are essential measures. These interventions can substantially reduce the likelihood of severe injuries among cyclists.

To ensure the effectiveness of these recommendations, continuous monitoring and policy adjustments are crucial. Policymakers should regularly review cyclist safety data and the outcomes of implemented measures to refine strategies and address emerging challenges. By adopting an evidence-based approach, transportation authorities can create safer environments for cyclists, encouraging more sustainable and healthy modes of transportation. This study provides a robust foundation for such interventions, offering practical insights that can inform the development of comprehensive road safety strategies.

7. Conclusion

This study has provided valuable insights into the factors influencing the severity of cyclist injuries in road accidents in Great Britain. Leveraging the CatBoost machine learning algorithm identified key predictors of injury severity, including speed limits, age group, day of the week, weather conditions, light conditions, road type, and road conditions. These findings have significant implications for policy-making and urban planning, emphasizing the need for targeted interventions to enhance cyclist safety.

The critical role of speed limits in reducing injury severity underscores the importance of implementing and enforcing lower speed limits in urban areas. Policymakers should advocate for speed management strategies that prioritize the safety of vulnerable road users such as cyclists. Additionally, identifying age-specific vulnerabilities highlights the necessity for targeted educational programs and safety campaigns tailored to younger and older cyclists. These interventions can effectively mitigate the heightened risks faced by these age groups and contribute to overall road safety.

Improving cycling infrastructure is another crucial recommendation arising from this study. Enhancing road quality, ensuring better street lighting, and maintaining weather-resistant infrastructure are essential measures to protect cyclists from environmental hazards. Urban planners and transportation authorities must prioritize these improvements to create safer cycling environments. The significance of these infrastructural changes is supported by existing literature, which emphasizes the need for comprehensive safety strategies that address various risk factors.

Despite the valuable findings of this study, there are limitations that warrant further research. The data used spans several decades, during which time there have been numerous changes in road safety policies, infrastructure, and vehicle technology. Future research could benefit from more recent data to capture the current trends and effectiveness of contemporary safety measures. Additionally, exploring the impact of emerging technologies, such as advanced driver assistance systems (ADAS) and smart city initiatives, on cyclist safety could provide new insights.

Future research should also consider the behavioral aspects of both cyclists and drivers. Understanding the interactions between different road users and how behavioral factors influence crash outcomes can lead to more effective safety interventions. Moreover, investigating the socioeconomic factors that may affect cyclist safety, such as access to protective gear and cycling education, can provide a more holistic understanding of the determinants of injury severity.

In conclusion, this study has highlighted the multifaceted nature of cyclist injury severity and provided actionable recommendations for enhancing cyclist safety. By implementing targeted interventions based on the identified predictors, policymakers, urban planners, and transportation authorities can create safer environments for cyclists. Continued research in this area is essential to address emerging challenges and to ensure that safety measures evolve in response to changing conditions and technologies. Through sustained efforts and evidence-based policies, safer, more sustainable urban mobility for all road users can be promoted.

Bibliography

1. Ahmed Jaber, J. J. (2021). An Analysis of Factors Affecting the Severity of Cycling Crashes Using Binary Regression Model. Sustainability , pp. 6945, https://doi.org/10.3390/su13126945.

2. Aldred R, G. A. (2018, Aug). Cycling injury risk in London: A case-control study exploring the impact of cycle volumes, motor vehicle volumes, and road characteristics including speed limits. . Accid Anal Prev. , p. 10.1016/j.aap.2018.03.003.

3. Bentéjac, C. C.-M. (2021). A comparative analysis of gradient boosting algorithms. Artif Intell Rev, pp. https://doi.org/10.1007/s10462-020-09896-5.

4. CatBoost. (2022). Retrieved from CatBoost: https://catboost.ai/en/docs/concepts/parameter-tuning

5. Dorogush, A. V. (2018). CatBoost: gradient boosting with categorical features support. p. arXiv preprint arXiv:1810.11363.

6. García-Herrero, S., Aldred, R., Anaya-Boig, E., & Mariscal, M. A. (2019). Vulnerability of cyclists on the road. A probabilistic analysis of the database of traffic injuries in Spain focusing on type of involved vehicle and driver culpability. Proceedings of the 29th European Safety and Reliability Conference. Research Publishing, Singapore.

7. Hancock, J. T. (2020). CatBoost for big data: an interdisciplinary review. . Journal of Big Data, pp. 7(94). https://doi.org/10.1186/s40537-020-00369-8.

8. Jaber, A., Juhász, J., & Csonka, B. (2021). An Analysis of Factors Affectingthe Severity of Cycling Crashes UsingBinary Regression Model. Sustainability , pp. 13,6945 https://doi.org/10.3390/su13126945.

9. JH, F. (2001). Greedy function approximation: a gradient boosting machine. pp. 1189–232.

10. https://www.kaggle.com/datasets/johnharshith/bicycle-accidents-in-great-britain-1979-to-2018

11. Karanikola P, P. T. (2018). Cycling as a Smart and Green Mode of Transport in Small Touristic Cities. Sustainability, p. 10(1):268. https://doi.org/10.3390/su10010268.

12. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1137-1143.

13. Liudmila P, G. G. (2018). Catboost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems 31, pp. 6638–6648.

14. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems,, p. 30.

15. Molnar, C. (2018). A guide for making black box models explainable. Retrieved from https://christophm. github. io/interpretable-ml-book.

16. Mustafa Ekmekci, N. D. (2024, May). Assessing the Impact of Low-Speed Limit Zones' Policy Implications on Cyclist Safety: Evidence from the UK. Transport Policy, p. http://dx.doi.org/10.1016/j.tranpol.2024.04.014.

17. Sivanandam, S. N. (2006). Introduction to Neural Networks Using Matlab 6.0. McGraw-Hill Education.