New York City Traffic Accidents

Jared Goroski, Eric Lin, Brodie McCarthy, Sung Park

Abstract

Traffic accidents pose significant challenges to public safety. To understand traffic accidents, we focused on various aspects such as trends, contributing factors, safety measures, machine learning models, and injury risk assessment. The research found the primary cause of New York City traffic accidents were driver inattention/distraction, failure to yield right-of-way, and speeding. Machine learning modeling techniques, including decision tree models, random forest classifiers, and neural networks offer insights into accident severity and severity probabilities. The findings highlight the importance of pedestrian safety, vehicle types, and contributing factors like speed and distraction in determining accident outcomes. It is crucial to identify problematic traffic areas and accident features in New York City, enabling the implementation of new safety systems or improvement to existing ones.

1 Introduction

With a population of over 8 million inhabitants [3], New York City has a complex traffic environment marked by a high volume of vehicles, diverse transportation modes, and a dense urban landscape. With nearly 100,000 car accidents a year [2] the factors contributing to the challenging traffic conditions include heavy congestion, intricate intersections, and the constant flow of commuters. The city's streets are often prone to traffic accidents, involving cars, pedestrians, and cyclists and considered the worst city in the country to drive in on multiple metrics [16].

Dating back all the way to the Commissioners' Plan of 1811, New York City was set to receive a grid system that creating an orderly arrangement of streets and avenues due to the rapidly growing nature of the city. The grid system made navigation and way-finding in Manhattan and later the rest of the New York City boroughs remarkably easy, contributing to the city's overall functionality This was a large part in determining the layout that we see today and while it remains easy to navigate in terms of directions, it has only gotten worse when it comes to the actual transportation around the city. Buses, subways, and ferries have all been implemented as successful solutions for congestion and traffic fatalities [19] but they've been temporary solutions to the issue. Safety interventions such as left turn bays, high visibility crosswalks, and speed humps have shown promise in reducing crashes while bus lanes had a negative impact when it came to traffic accidents [5].

There are many factors at play for the cause of accidents such as driver inattention, distraction, unsafe speed, failure to yield, following too closely, a disregard to traffic control, unsafe lane changing, and more [2, 18]. To combat traffic fatalities, New York City implemented a Vision Zero policy in 2014 that has brought traffic deaths to historic lows. The goal of the Vision Zero policy is to eliminate death and serious injuries from preventable traffic incidents through engineering, enforcement, and education [11]. The policy has lower New York City's default speed limit from 30 MPH to 25 MPH which has reduced the number of accidents across New York City. Through continuous improvement Vision Zero has fostered numerous additional countermeasures since its inception. These include enhanced street lighting, improved pedestrian crossings, expanded bike infrastructure, and targeted interventions such as automated speed and red light violation technology in high-risk areas [13]. While New York City's Vision Zero policy has undeniably made significant strides in reducing traffic fatalities, it's essential to recognize that achieving zero deaths and serious injuries on the roads remains a challenge.

Through our research, we will find and identify any specific boroughs, zip codes, and hot spots in New York City exhibiting a high frequency of collisions. We will then analyze traffic accident crashes to see what features have the highest influence on injury and death and find if the severity of an accident can be determined using machine learning models. By identifying problematic accident features and traffic areas of New York City we hope to identify areas to implement new safety systems and safety measures or improve existing ones.

2 Methods

2.1 Geospatial Exploration

Given the subject matter, it was necessary to explore how accidents were related to their locations. This meant looking at accidents as classified by their location via their reported borough as well as latitude and longitude. Not every accident had a zip codes or borough name that came along with their coordinates, this meant scrubbing through and classifying these latitude and longitude variables by the geographic bounds of each zip codes and borough to try and more accurately examine these trends.

2.2 Feature Analysis

2.2.1 Crash Data

Crash data provides essential information about the circumstances and conditions surrounding each accident. This includes factors such as the location (borough, zip code), time, injury, death, and contributing factors like driver behavior and vehicle maneuvers. Analyzing crash data allowed us to identify patterns and trends in accident occurrence, determine high-risk areas, and assess the effectiveness of safety interventions.

2.2.2 Person Data

Person-specific data offers insights into the individuals involved in each accident, including their roles (driver, passenger, pedestrian, cyclist), position in vehicle, use of safety equipment, age, and gender. This type of data helps in understanding how different groups are affected by traffic accidents. Analyzing person-specific data enabled us to assess the impact of factors such as age, gender, and mode of transportation on injury outcomes.

2.2.3 Vehicle Data

Vehicle data provides information about the vehicles involved in each accident, including their characteristics, such as vehicle type, make, and year, state registration, pre-crash conditions, point of impact, and vehicle damage. Analyzing vehicle data allowed us to assess the role of vehicle attributes in accident severity and identify potential safety improvements related to vehicle design or technology.

2.3 Decision Tree

Within our data, we wanted to predict the severity of collisions. This meant whether collisions would result in an injury or death. We thus wanted to use classification models to accomplish this.

The first of these models we made was a decision tree model made through Scikit-Learn. This was done mostly to create create some sort of readable or easily understandable decision tree that someone could look at a specific accident and understand if someone would be injured or not to a reasonable accuracy level. We wanted this to act as a layperson's guide that could help people make decisions when it comes to safety or better understand the importance of certain aspects of an accident. These decision trees typically lack the nuance or accuracy of other models but it still remained a reasonably effective model.

2.4 Random Forest Classifier

Following suit with trying to predict the severity of a car accident, we next made a Random Forest Classifier. Random Forest Classifiers was one of the main models used for actual accuracy. This is due to their overall success as some of the most commonly used classifiers in not just normal literature, but also in specific literature that we've seen used on this dataset before in the past. Random Forest Classifiers work very similar to Decision Tree Models but are less legible due to their vast number of individual branches and nodes. They can also usually handle various qualitative columns better due to their number of estimators.

2.5 Neural Network Predictor

We employed a neural network predictor designed to forecast the outcomes of traffic accidents, categorizing them into three potential results: uninjured, injured, or fatal. The model utilizes a comprehensive dataset that includes attributes such as crash date, location coordinates, and various personal and vehicular factors involved in each incident. Trained using TensorFlow and Keras, the neural network is able to identify complex patterns and make predictions efficiently. Its ability to handle multi-dimensional and varied data makes it highly effective for applications that demand high accuracy, such as traffic management and urban planning, where reliable decision-making is critical.

2.6 Linear Regression Predictor

We developed linear regression models to predict the number of injuries from accidents based on the time of day, using Scikit-Learn for model implementation. Initially, we employed a simple model that used the crash hour as the sole predictor, grouping the total number of injuries in each hour. We also incorporated pre-crash conditions and categorized injury outcomes as uninjured, injured, or death. Additionally, we experimented with multiple variables and interaction terms in an attempt to capture more complex patterns and improve the model's predictive accuracy.

3 Results

3.1 Geospatial Exploration

During our Exploratory Data Analysis we found that Brooklyn stood out as the borough with the highest number of traffic accidents, while Staten Island had the lowest number of accidents 1.

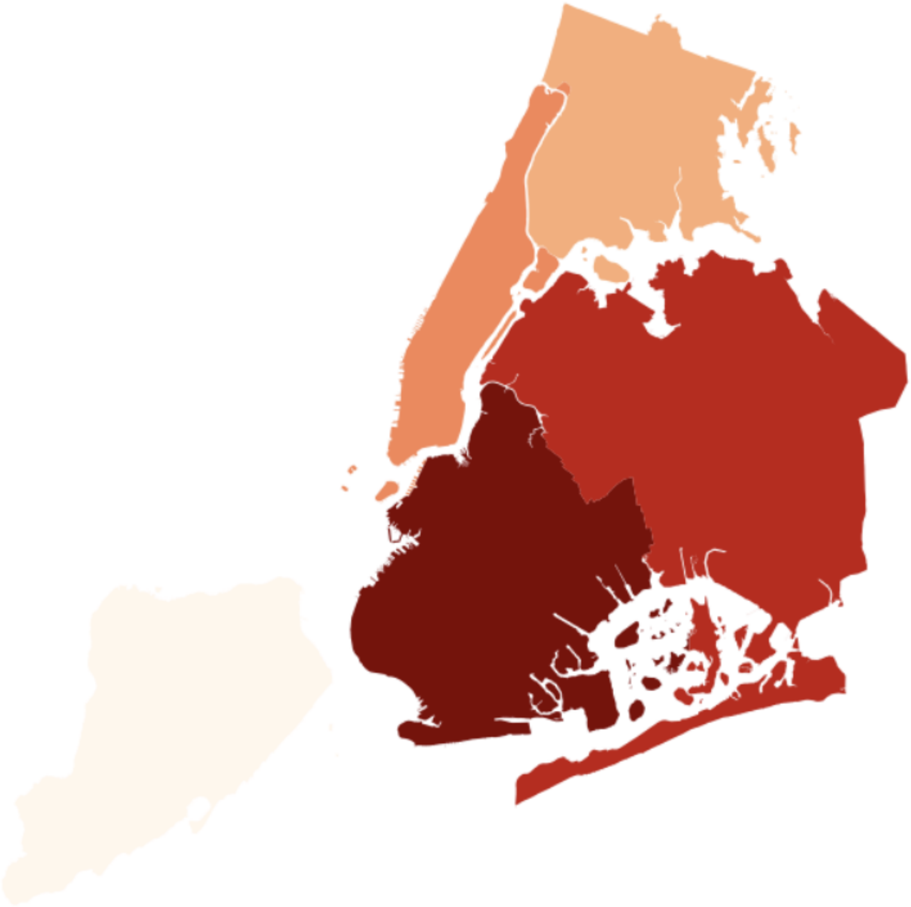

Figure 1: NYC Borough Collision Counts

When normalizing zip code with respect to population, we obtain a choropleth map of New York City zip codes in each borough with the highest rate of accidents 2.

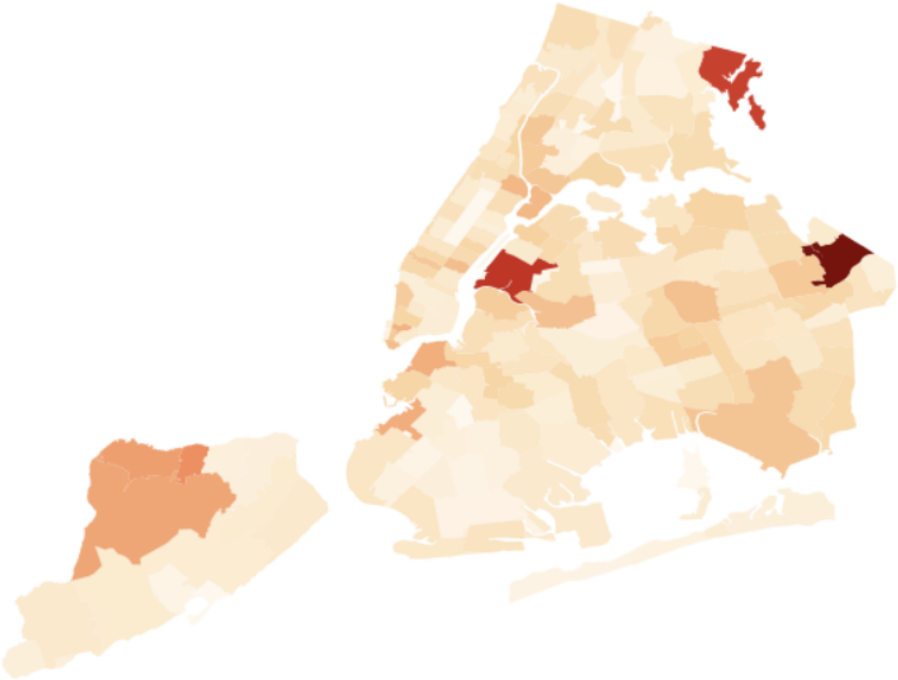

Figure 2: Collision Rate by Zip Code Normalized with Respect to Population

We find that the borough with the zip code with the highest rate of accidents is Queens, followed by Brooklyn and the Bronx, and lastly Staten Island and Manhattan. However most zipcodes had a relatively equal percentage when compared to their population size. The few that stood out did so due to a very low population count due to being just an area used for transit from one area to another or just a business center. The zipcodes on the boundaries mostly served as entry points into New York City with little population living in them but had a lot of traffic due to commuters.

3.2 Accident Factors and Vehicles

During our Exploratory Data Analysis, we found the factors that caused the most number of injuries were driver inattention/distraction, failure to yield right-of-way, and following too closely. The factors that caused the most number of deaths were unsafe speed, inattention/distraction, and failure to yield right-of-way 3.

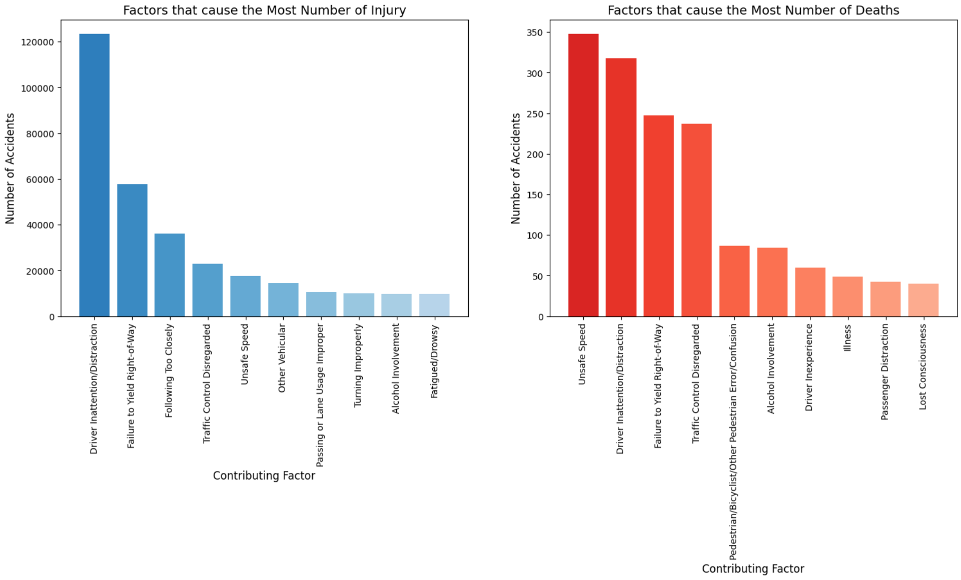

Figure 3: Common Accident Factors Count

When looking at the vehicles involved in accidents, we found that sedans, station wagons/sport utility vehicles, and passenger vehicles were involved in the highest number of accidents across New York City. However, when we look at the rate of injury and death, those vehicles are not found see figure 4.

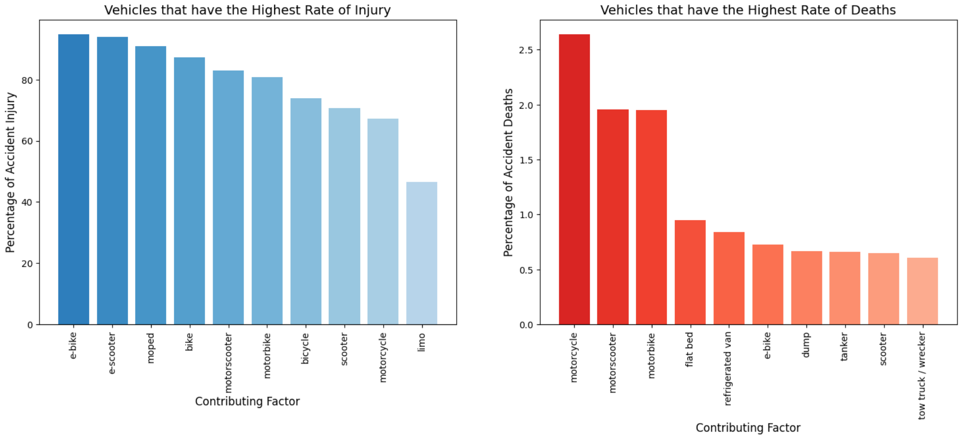

Figure 4: Vehicles Injury and Death Rates

Although sedans, station wagons/sport utility vehicles, and passenger vehicles were involved in the most accidents, two-wheel vehicles such as bikes, scooters, mopeds, and motorcycles have a 70 to 80 percent rate of injury, meaning that people on two-wheeled vehicles involved in an accident were more likely to come out injured. When looking at the rate of death, we find that two-wheeled vehicles like motorcycles, motorscooters, and motorbikes have the highest rate of death when involved in an accident, as well as large vehicles such as vans and trucks.

3.3 Injury Outcome Based on Age and Mode of Transport

When analyzing the factors influencing the severity of injuries sustained in traffic accidents, age emerges as a significant variable. The relationship between age and injury rates may initially seem to be a reflection of the frequency of accidents within each age group.

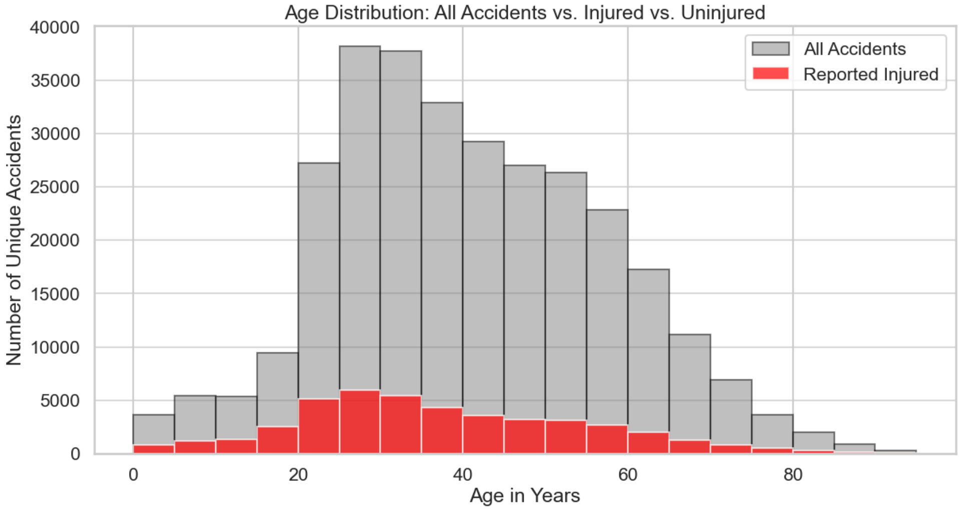

Figure 5: Injury result by ages

The histogram in Figure 5 visualizes the age distribution of individuals involved in reported accidents, distinctly emphasizing those who sustained injuries. The figure exhibits a concentration of accidents and associated injuries among the middle-aged demographic around 25-35, followed by a progressive decrease in these incidents with advancing age. This trend suggests a potential correlation between age and the likelihood of being involved in an accident. It also raises questions about the role that factors such as experience, risk exposure, and reflexes, which are typically age-related, play in the propensity for accidents and injury severity. This pattern could lead to potential targeted campaigns or preventative measures to try and lower the number of accidents between 25-35 as they contribute to the majority of accidents.

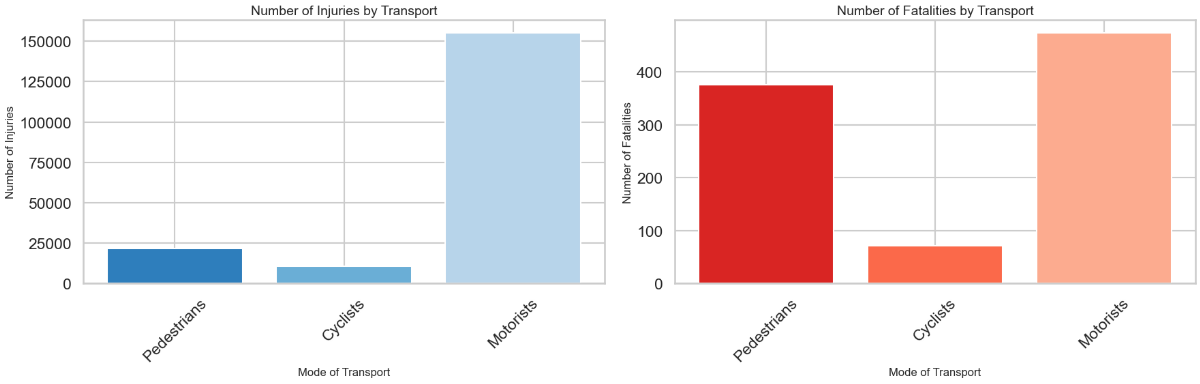

Figure 6: Injuries and Fatalities by Mode of Transport

Figure 6 reveals a stark contrast in the risk profiles for different road users. Motorists (people in enclosed vehicles), while encountering the highest number of injuries, have the most protection between them, and the accidents. On the other hand, pedestrians, have no protection measures and are exposed, resulting in a significantly higher risk of fatalities when involved in traffic accidents. The data illustrates that pedestrians suffer the most severe outcomes despite lower overall injury counts, underscoring their heightened vulnerability in traffic incidents.

3.4 Decision Tree

One of the first models we used to predict whether an individual accident would result in an injury, fatality, or just unspecified was the Decision Tree model. The decision tree model was made to primarily act as a simplistic way to visualize and report the findings rather than for pure predictive capabilities.

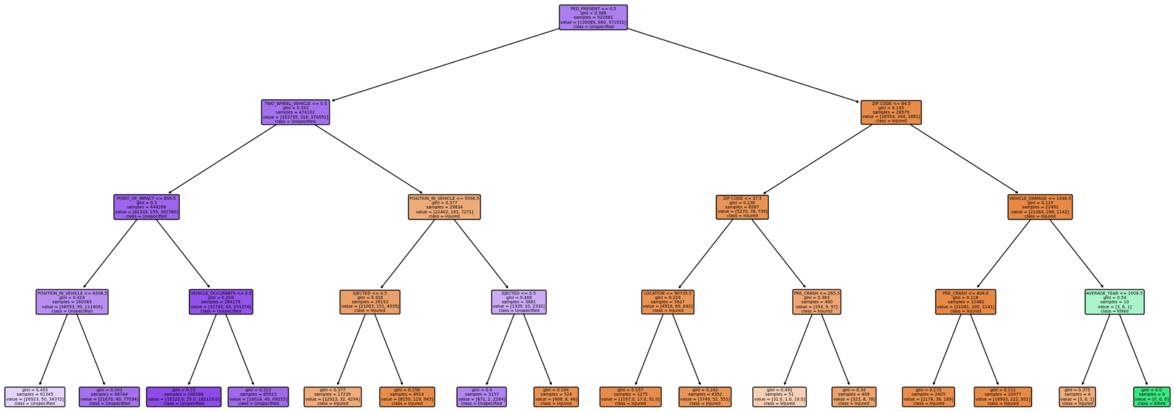

Figure 7: Earlier Decision Tree with 4 Levels

A few trees were made but our best performing one was a decision tree with 5 levels of nodes which garnered an accuracy of 75.5% in classifying. We could have gotten better results but this was done with balanced class weights. A reoccurring issue in classifying was the fact that our data at the very end of cleaning had very low percentage of accidents resulting in deaths. We had an end percentage of 74% unspecified, 25% injured, and then <1% deaths. This would lead to most models simply neglecting to include a death classification at all. We wanted our model to be more sensitive as we wanted this model to be more descriptive of the whole picture and what might lead to deaths.

When not balanced, these decision trees would often operate at an accuracy of ¿80% however they would never classify any deaths in our testing set. This set off alarms as this model was intended to provide easily understandable insights for people to gather information from.

When it came to the decision trees, the points that mattered the most for classification across all iterations were whether or not a pedestrian was present, whether someone was ejected, and if cyclists were also present in the collision. A few other ideas that popped up were the the point of impact as well as the time of day. This was something also brought up in other articles that mentioned how accidents later in the day tended to be more fatal due to the less traffic collision and thus higher traffic speeds. [18]

The purple color represents unspecified, orange represents injury, and green represents if someone was killed or not. The decision tree does vastly over represent the number of deaths compared to the data set but it does provide some valuable insight Our end tree had relatively high recall except for the number of injured with 0.40 but pretty poor precision for the number of killed with 0.02. This was to be expected as with the balanced weights we had a lot of injured classified as killed. A more legible version of the graph follows.

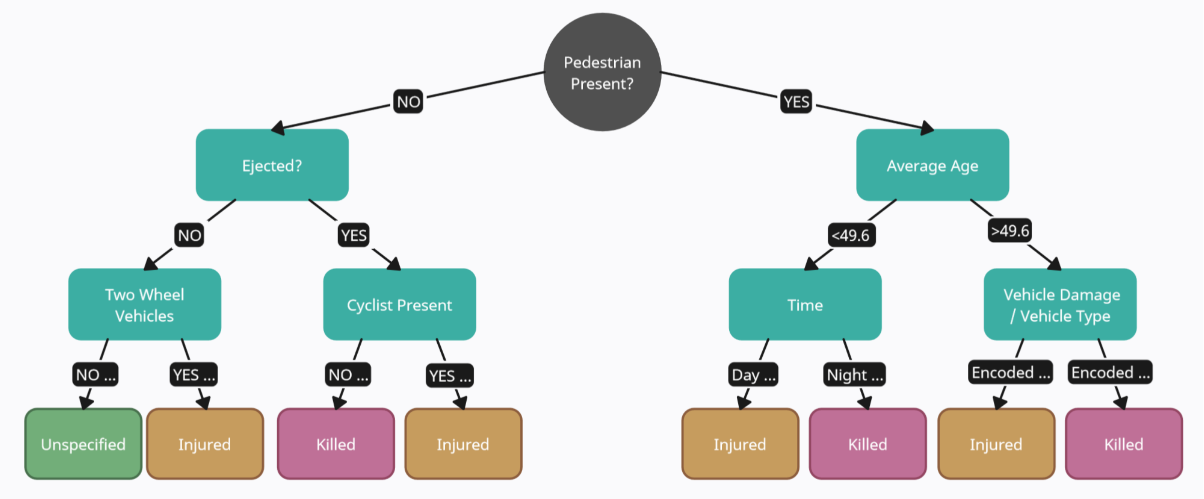

Figure 8: Decision Tree Simplified

With this we can see the most important features of the model produced. Whether or not a pedestrian was present was the biggest overarching factor where if a pedestrian was present, it was almost certain of an injury or death occurring. Without a pedestrian present, it determined if there was an ejection in the collision and if so, it required whether or not a cyclist was present to determine if there was a death or injury. Without an ejection either, it mostly relied on if an unsafe or two wheel vehicle was present as well such as a motorcycle or moped which could lead to an injury. However most accidents included none of these and were thus correctly lumped into the unspecified.

If a pedestrian was present however, we observed that it often used the average age of the people in the accident to determine some level of severity. Iterations varied on the exact ages but they often found that those accidents with older average ages (right at the junction) were more likely to die rather than be injured when compared to those collisions with younger average ages (left at the junction). In our final decision tree, it was found that an average age of 49 was the cutoff however it should be noted that these ages depended on the limited availability of deaths in the dataset, rather than a condemnation of anyone past 49. Then after that it often looked to the time of day which seemed to agree with earlier research where it said that accidents later in the day (right at the function) were more likely to be fatal. Again, the exact time of this split would change on iteration but it would find that accidents early in the morning or later than afternoon were more dangerous times. Another feature that often popped up across all branches was the importance of vehicular damage or the type of vehicle beyond if there was just a two wheeled vehicle or not. This classification helped with analysis of if there was a large amount of damage then it was more likely to result in a death or typically injury. The final decisions made with these branches were often encoded and difficult to decipher but offered insight into the importance of damage caused high rates of speed, impact of the collision, and more.

3.5 Random Forest Classifier

The main model used to try and accurately classify the severity of an accident was a random forest classifier. This performed better even with balanced class weights. With 100 estimators, our average accuracy for the model after cross validation was 81.3% with a high of 83.6% and a low of 73.8% when made into 5 separate folds for cross validation.

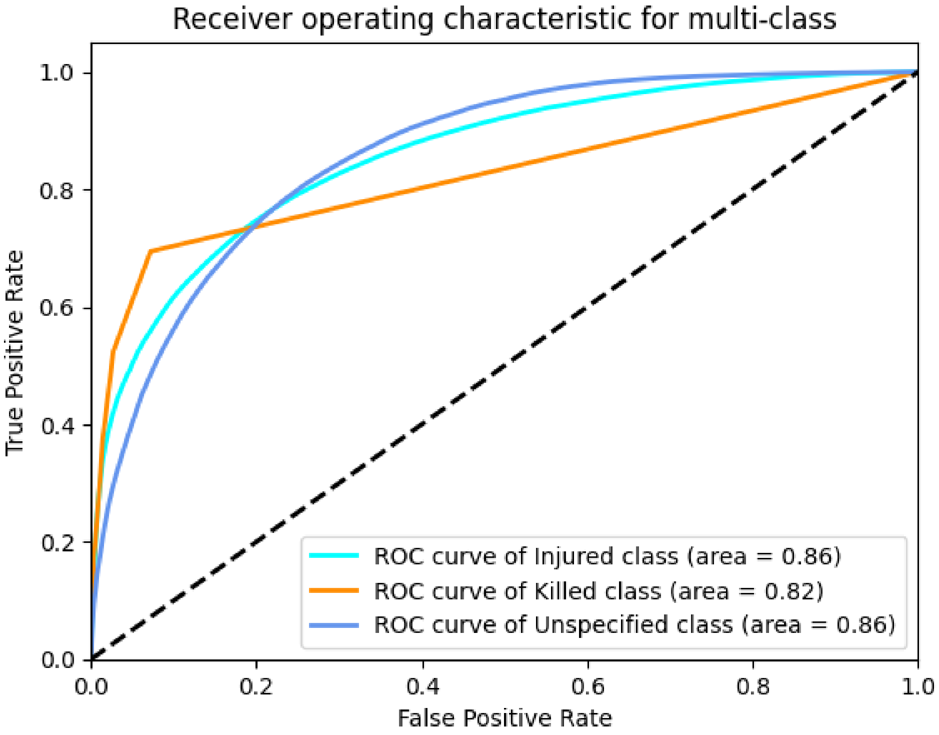

Figure 9: Random Forest ROC Curves

Given the various metrics, our ROC curves 9 also performed quite well with the area under the curve for injured, killed, and unspecified being 0.86, 0.82, and 0.86 respectively. The killed class was our hardest just due to the nature of the disproportionately low amount of deaths in the dataset.

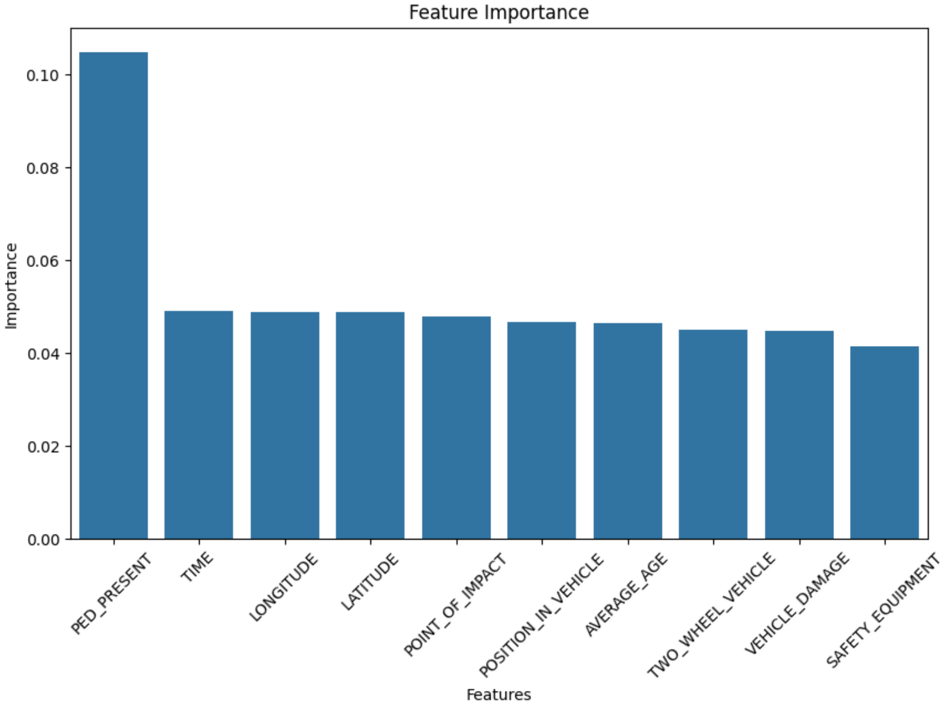

Figure 10: Random Forest Top Feature Importances

When it came to the feature importance 10, the random forest had relatively similar rates of importance as compared to the decision tree. It agreed and found that whether or not a pedestrian was present was the most important feature at nearly 10.5% on its own. It also found the importance of time of day and even the point of impact or the position in the vehicle. Another thing it found important was the presence of two wheeled vehicles like motorcycles or mopeds which was something we included due to the data exploration performed earlier.

These models largely agreed with previous articles on the importance of pedestrian and cyclist safety in the realm of public safety and curbing injury or deaths in traffic collisions. This goes to show how more work should be done when it comes to public safety like the institution of guard rails for pedestrians or inclusion of automatic braking systems or sensors to help with driver inattention causing accidents. And as drivers or pedestrians, we need to be more aware of our surroundings and obeying traffic laws and be more aware of the different levels of danger that pedestrians or cyclists face.

3.6 Neural Network Analysis

The construction and evaluation of a neural network to predict traffic accident outcomes yielded insightful results. The network, structured with three internal layers with 128, 64, and 32 respectively that output to the 3 potential outcomes those being uninjured, injured, or fatal. was optimized to discern patterns within a dataset, rigorously excluding direct injury indicators to focus on underlying predictive factors.

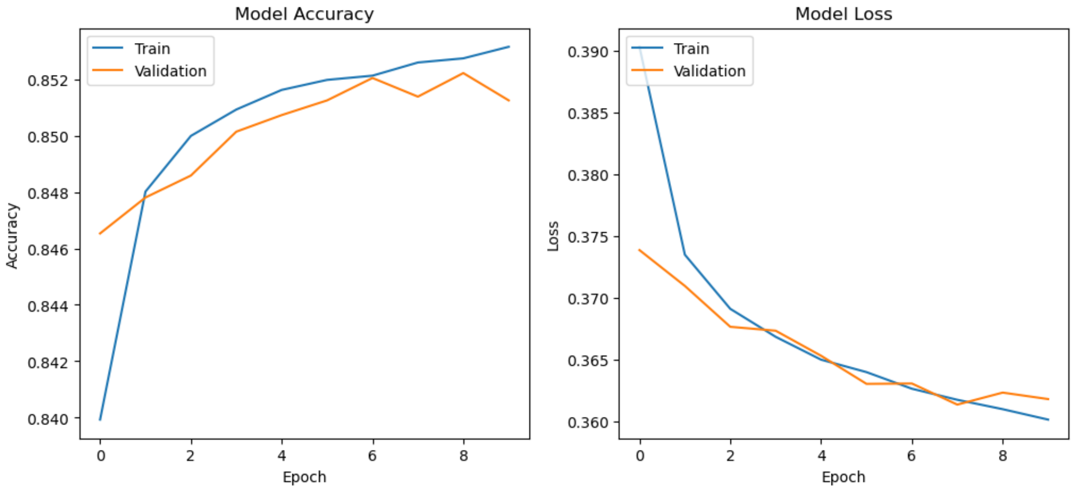

Figure 11: Model Accuracy and Loss over Training Epochs

As depicted in Figure 11, the model's accuracy on the training and validation sets displayed an upward trajectory with each epoch with a roughly 85% accuracy when running through the cross fold, indicating continuous learning and adaptation. Simultaneously, the loss—a measure of how well the model's predictions match the actual data—decreased steadily. This balance between increasing accuracy and decreasing loss is emblematic of a well-fitting model that maintains generalized without over-fitting to the training data. The loss function that was used for this model was sparse categorical cross entropy. The cross entropy error is at .393 and ended up down to .360, This loss function is particularly suited for classification problems with multiple classes where the classes are represented as integers. It measures the discrepancy between the predicted probability distribution and the true distribution, effectively penalizing the probability assigned to the incorrect class. This approach is efficient in handling large datasets with numerous classes, ensuring that the model remains robust and accurate across diverse inputs.

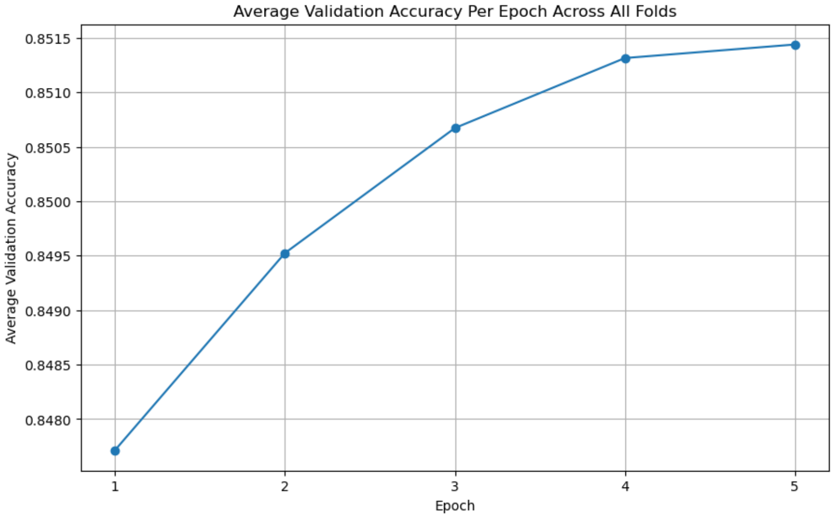

Figure 12: Average Validation Accuracy Per Epoch Across All Folds

Figure 12 illustrates the average validation accuracy obtained through a 5-fold cross-validation method. This approach not only reaffirms the model's performance but also underscores its reliability across the different subsets of data. With each fold serving as an independent test, the model demonstrated consistent accuracy, as evidenced by the steady increment seen across epochs. The average accuracy maintains a consistent nature across the different runs and a consistent improvement till the end ending at 85.13%. Across the five folds, the accuracy varied by just 0.11%, with the lowest score being 85.08% and the highest 85.19%, resulting in an average accuracy of 85.14%.

In the context of these results, accuracy refers to the proportion of total predictions that the model classified correctly. It is a primary metric for evaluating classification models, particularly when the classes are well-balanced. Loss, on the other hand, quantifies the difference between the predicted values and the actual values, providing a more granular view of model performance. The low and decreasing loss in tandem with high accuracy suggests that the model's predictions are not only correct but also made with high confidence. Together, these metrics provide a picture of the neural network's performance, indicating a high level of precision in predicting the severity of traffic accident outcomes. The consistency of these metrics across training epochs and validation folds presents a strong case for the model's deployment in practical applications, where reliable predictions can inform preventative measures and policy-making.

3.7 Linear Regression

We developed a linear regression model to predict injury and death rates in accidents, utilizing Scikit-Learn for model construction. Initially, we conducted a correlation analysis using a heatmap to examine relationships among all numerical variables with injury and death rates. This analysis revealed that the crash hour had a significant correlation; the correlation coefficients were 0.450 with injury rate and -0.160 with death rate. Other variables exhibited coefficients ranging between 0.1 and -0.1. For instance, the average age and injury ratio showed a relatively low correlation of -0.120 compared to crash hour. Furthermore, we assessed the correlation between pre-crash conditions—a quantitative variable—and injury and death rates, which yielded a weak correlation coefficient of 0.072. Consequently, crash hour was selected as the sole variable for our initial linear regression model.

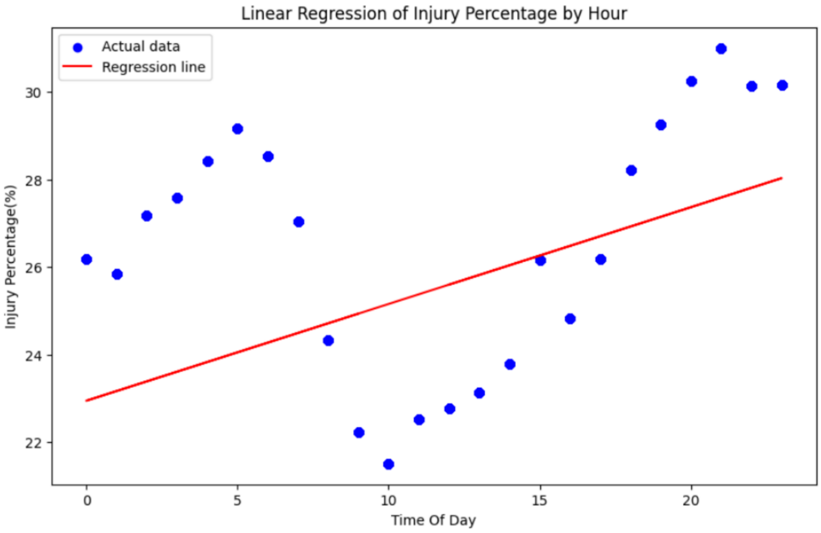

Figure 13: Linear Regression Of number of injuries by crash hour.

These models aim to predict the likelihood of an individual being injured or killed during a specific hour of the day. Our initial model, which utilized crash hour as the sole predictor, categorized injuries by hour and normalized them by the total number of injuries (as illustrated in Figure 13). This approach yielded an R-squared value of 0.2053 and an RMSE of 2.5044. Similarly, the model predicting death probability by hour achieved an R-squared value of 0.0269 and an RMSE of 0.0714.14

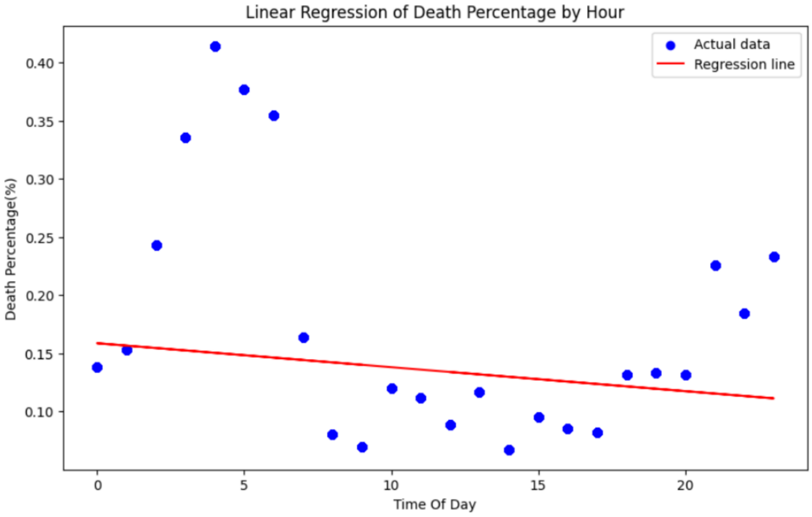

Figure 14: Predicted vs. Actual Injuries based on Pre-Crash Condition Regression Model

However, the limited correlation observed in these models underscores the inadequacy of relying solely on crash hour for accurate predictions of injury and death. In subsequent experiments, we incorporated multiple variables and interaction terms into the regression models. Despite these enhancements, the increase in predictive power was minimal, with an R-squared improvement of less than 0.001. This marginal enhancement indicates that the added complexity from multiple variables and interaction terms had little impact on refining the model's accuracy. Compared to other modeling approaches, our linear regression models underperformed, highlighting the limitations of this methodology for the given dataset.

4 Discussion

Throughout the project we identified major locations in New York City that exhibited a high frequency of collisions, primary factors that caused crashes and injury, and effective machine learning models.

Traffic accident trends across multiple studies and our own Exploratory Data Analysis, demonstrated noteworthy patterns and contributing factors. Among the New York City's boroughs, Brooklyn stood out with the highest documented crash occurrences at 32% of the total crashes, followed by Queens at 27%, then Manhattan at 22%, then the Bronx at 15%, and finally Staten Island at 4% [17] 1. We also found that when looking at New York City collisions per capita the two ZIP Codes that ranked the highest resided in Queens in Flushing, ZIP Code 11362 and Long Island City, ZIP Code 11101. It is then followed by a ZIP Code in the Bronx, 10464 and two in Staten Island 10303 and 10302.

Across multiple studies and in our analysis 3, primary factors like driver distraction, failure to yield right-of-way, and following too closely emerged as the primary factors leading to traffic accidents in New York City [17]. Our analysis on vehicle injury 4 found that people on two-wheeled vehicles that were involved in accidents were more likely to come out injured than not which was also found in a report in 2018, where 18,718 reported accidents involving cars colliding with bicycles, resulted in a significant portion of injured cyclists [15].

Both the random forest classifier and the neural network models demonstrated a strong capability of predicting the results of a traffic accident. The neural network achieved a commendable 85% accuracy in determining injury status, slightly outperforming the random forest model, which secured an 82% accuracy rate.

Neural networks, validated through cross-validation identified critical variables in our research, predicted probabilities of a collision occurring given a set of initial conditions and accident severity and chance of fatality similar to those found by Wijewickrama and Rezaeiahari and Shakil et al. [20, 1]. It demonstrated the potential of neural networks as a reliable tool for forecasting traffic accident outcomes.

The Random Forest Classifier performed well with a higher accuracy than the non-information rate while still retaining a higher sensitivity compared to the decision tree. Much of the data points to helping prioritize pedestrian safety when it comes to making accidents less severe. This could include ideas as discussed like automatic braking systems [4], protected bike lanes or crosswalks [5], and other protective measures like more strictly enforced laws or better designed roads [12]. Both the random forest classifier and the neural network models could help shaping future road safety strategies and reducing the impact of traffic incidents on public health.

4.1 Challenges

The project involved a significant transition from three large datasets (Crashes [7], Person [9], Vehicles [10]) that contained 11.6 million rows and 75 columns to a comprehensive master dataset with 600 thousand rows and 45 columns. The original datasets had issues such as extensive missing data, encoded or unreadable entries, inconsistent data entries, and different data formats. The machine learning phase faced challenges in addressing the imbalanced nature of ground truth labels, with over 800 deaths, 162,000 injuries, and 464,000 unspecified outcomes. Random forest models struggled to accurately predict fatalities, often prioritizing injury or unspecified outcomes to optimize overall accuracy. Balancing predictive accuracy while predicting all outcomes proved complex and an ongoing challenge for the project. With 44 input columns available for predictive modeling, selecting the most relevant features and optimizing model performance presented additional challenges. Balancing model complexity, readability, and accuracy required frequent experimentation and fine-tuning of machine learning algorithms. Navigating the intricacies of data cleaning, imbalanced ground truths, and model optimization was crucial for the project's success.

4.2 Future Work

4.2.1 Weather

Integrating weather data into our analysis could provide valuable insights into the correlation between weather conditions and the frequency or severity of traffic accidents. By examining factors such as precipitation, temperature, and visibility, we can better understand how adverse weather conditions contribute to accidents and potentially develop predictive models to mitigate risks. Weather data can be obtained from OpenWeatherMap's One Call API [14] for each New York City location or borough for each day in the last 40 years, which is more than enough for our work.

4.2.2 Vehicle Distribution

Integrating data on the spatial distribution of vehicles across New York City boroughs with traffic volume analysis can provide a comprehensive understanding of traffic flow and its impact on accident rates. By examining traffic patterns, infrastructure, road conditions, and congestion levels, we can identify high-risk areas and reveal opportunities for targeted traffic interventions or infrastructure improvements to enhance road safety, reduce risk, and increase traffic efficiency.

4.2.3 Pothole

Examining pothole data alongside traffic accident data can also provide valuable context regarding the role of infrastructure conditions in contributing to accidents. By identifying correlations between the location of potholes and accident hotspots, we can reveal locations for targeted maintenance efforts and infrastructure investments to enhance road safety. An example dataset can be obtained from NYC Open Data through their 311 Service Request from 2010 to the Present [6] or a community-created dataset on Potholes using the same dataset [8].

4.2.4 Person Race

While race data may present ethical challenges, its inclusion in our analysis could offer insights into disparities in traffic accident rates among different racial or ethnic groups. However, it's crucial to approach this aspect with sensitivity, ensuring that our methodologies adhere to ethical guidelines and respect privacy concerns.

By leveraging additional datasets on weather data, borough car distribution, traffic volume information, pothole data, and race data (with appropriate ethical considerations) we can make a more thorough analysis of NYC traffic accidents. These future steps hold the potential to uncover new insights, inform evidence-based interventions, and ultimately contribute to the improvement of road safety and transportation infrastructure in New York City.

5 Conclusion

This study has presented the critical need for targeted traffic safety interventions, particularly for vulnerable road users such as pedestrians and bicyclists. Our analysis revealed that pedestrians and cyclists are disproportionately at risk of injury or fatality in traffic accidents, with a significant number of incidents attributable to key factors such as distracted driving, failure to yield, and speeding. The geographical analysis indicates that Brooklyn and Queens are hotspots for traffic accidents, suggesting a focused approach to safety improvements in these areas could be particularly effective.

In terms of predictive modeling, our findings demonstrate the effectiveness of machine learning techniques in understanding and predicting traffic accident outcomes. The neural network, with an accuracy of 85.14%, proved slightly more effective than the random forest model, which had an accuracy of 83.6%. This suggests that both of these models are effective predictor models because both neural networks and random forests offer valuable insights for traffic safety analysis. Both models show potential for further refinement through increased complexity and enhanced data integration.

The linear regression models under-performed, indicated by low R-squared values, reaffirms the need for more sophisticated analytical tools in this domain. Therefore, we recommend that future policy efforts not only focus on engineering and enforcement solutions but also incorporate advanced data analytics to identify risk patterns and evaluate the effectiveness of traffic safety interventions.

In light of these findings, we advocate for implementing protective measures such as dedicated bike lanes and high-visibility crosswalks, especially in areas with high accident rates. By leveraging detailed predictive insights and focusing on high-risk factors and locations, policymakers and traffic safety officials can more effectively reduce the incidence of traffic-related injuries and fatalities.

6 Code Appendix

N.Y.C. Traffic Accident GitHub

References

[1] Shakil Ahmed, Md Akbar Hossain, Sayan Kumar Ray, Md Mafijul Islam Bhuiyan, and Saifur Rahman Sabuj. A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance. *Transportation Research Interdisciplinary Perspectives*, 19:100814, 2023.

[2] Christy Bieber. Nyc Car Accident Statistics In 2024. *Forbes*, June 30 2023.

[3] U.S. Census Bureau. U.S. Census Bureau QuickFacts: New York city, New York, July 1 2022.

[4] François Char and Thierry Serre. Analysis of pre-crash characteristics of passenger car to cyclist accidents for the development of advanced drivers assistance systems. *Accident Analysis & Prevention*, 136:105408, 2020.

[5] Li Chen, Cynthia Chen, Reid Ewing, Claire E. McKnight, Raghavan Srinivasan, and Matthew Roe. Safety countermeasures and crash reduction in new york city—experience and lessons learned. *Accident Analysis & Prevention*, 50:312–322, 2013.

[6] NYC Open Data. 311 service requests from 2010 to present, Oct 2011.

[7] NYC Open Data. Motor Vehicle Collisions - Crashes. https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data, May 7 2014. [Online; accessed 2024-02-18].

[8] NYC Open Data. Potholes, Mar 2014.

[9] NYC Open Data. Motor Vehicle Collisions - Person. https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Person/f55k-p6yu/about_data, Nov. 11 2019. [Online; accessed 2024-02-18].

[10] NYC Open Data. Motor Vehicle Collisions - Vehicles. https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Vehicles/bm4k-52h4/about_data, Nov. 11 2019. [Online; accessed 2024-02-18].

[11] NYC DOT. Nyc DOT.

[12] Ibraheem M. Karaye, Temitope Olokunlade, Alyssa Cevetello, Kameron Farhadi, and Corinne M. Kyriacou. Examining the Trends in Motor Vehicle Traffic Deaths in New York City, 1999–2020. *Journal of Community Health*, 48(4):634–639, mar 7 2023.

[13] City of New York. Vision Zero.

[14] OpenWeatherMap. New york city weather forecast.

[15] PeoplePoweredMovement. Nyc bicycle safety overview: Infrastructure & crash stats.

[16] Heather Reinblatt. The Worst Cities for Driving in America. *Circuit*, Nov. 16 2023.

[17] Abhishek Saxena and Stefan A. Robila. Analysis of the new york city's vehicle crash open data. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 6017–6019, 2021.

[18] Khaled Shaaban and Mohamed Ibrahim. Analysis and identification of contributing factors of traffic crashes in new york city. *Transportation Research Procedia*, 55:1696–1703, 2021. 14th International scientific conference on sustainable, modern and safe transport.

[19] Robert Steuteville. Big safety benefits from transit, sep 14 2016.

[20] Ishani R. Wijewickrama and Mandana Rezaeiahari. Predicting motor vehicle accidents in new york city. *IIE Annual Conference.Proceedings*, pages 503–508, 2018.