EXAMINING PATTERNS AND PREVENTIVE STRATEGIES FOR VEHICLE CRASHES

Meghana Kota, Harshad Gupta Pasumarthy

Advanced Data Analytics

ADTA 5940 – Analytics Capstone Experience

Dr. Jamie Humphries

04 May 2025

Introduction

Vehicle crashes are a significant public safety concern, leading to injuries, fatalities, and substantial economic losses. Traffic is very high in all the metropolitan areas, and diverse road conditions exist. Therefore, understanding the factors influencing crash occurrences is crucial for improving public road safety. Analyzing crash data can help us understand the patterns and factors that affect accidents and identify strategies to reduce them and improve transportation.

Data related to crashes is gathered from both the transportation department and law enforcement officials. Police officers document all crash incidents immediately following the event, noting details such as location, time, severity, and any fatalities that may have occurred. This information pertains to traffic crashes in Chicago as recorded by the Chicago Police Department (CPD). The crash data is available on the Chicago data portal, primarily adhering to the data elements in the format SR1050. CPD logs every traffic crash event reported, irrespective of the statute of limitations. Chicago crash data serves as a basis for analyzing crashes, identifying trends, and developing preventive measures. Analyzing this dataset can help identify various risk factors and reduce the frequency of crashes by implementing policies and strategies to decrease fatal accidents.

Analyzing Chicago's traffic crash data reveals significant trends, challenges, and opportunities to improve urban transportation safety. Day by day, the number of accidents in the Chicago region is increasing due to higher traffic volume, driver behavior, and ongoing urban development. Vehicle type, speed, weather conditions, reckless driving, and other factors are a few that affect crashes. Right now, there are few policies, rules, and regulations to reduce the occurrence of crashes. Efforts to enhance traffic safety include the incorporation of advanced safety features in vehicles. Modern vehicles are increasingly equipped with technologies such as Automatic Emergency Braking (AEB), Lane-Keeping Assist, Adaptive Cruise Control, and Blind Spot Monitoring. These systems help mitigate human error, which is a leading cause of crashes. There are a few public awareness sessions to educate the public on safe driving practices and the enforcement of traffic laws to reduce reckless driving behaviors. Smart infrastructure solutions are being adopted in urban areas, such as Chicago, including intelligent traffic signals that optimize traffic flow based on real-time conditions, and connected vehicle technology that enables vehicles to communicate with the infrastructure.

Significant research has been conducted to reduce the number of crashes and implement preventive measures. The primary goal of our study is to identify the factors that influence crashes and their severity using the Multinomial Logistic Regression and XGBoosting Classifier. By using these predictive models, we can find targeted interventions to reduce accidents. Recent advancements in road safety, with autonomous vehicles equipped with sensors and AI algorithms, can eliminate human error. We have reviewed various articles that support our study. Ultimately, we aim to offer recommendations and policies to the government on minimizing crashes by using modeling techniques such as the Multinomial Logistic Regression and XGBoosting Classifier.

Our analysis examines prior research on crash analysis, statistical modeling approaches, and best practices for mitigating road accidents, with a focus on studies that explore accident modeling, crash severity factors, and predictive safety measures. To further improve road safety in urban areas like Chicago, future research should focus on several key areas. Conducting deeper analyses of how modern vehicle safety systems interact with urban infrastructure could provide valuable insights into optimizing safety measures. Developing predictive models that incorporate real-time data from IoT devices can enhance safety measures. By combining insights from crash data analysis with advancements in vehicle technology and smart infrastructure, policymakers can implement strategies to minimize accidents and improve public safety on roads.

Research Questions

Traffic crashes are a persistent concern in major urban centers, and Chicago is no exception. With thousands of incidents reported annually, these crashes lead to injuries, fatalities, and significant economic losses. The Chicago crash data from 2020 to 2024 provides a valuable opportunity to examine patterns and identify factors that contribute to crash severity. This study aims to leverage machine learning and exploratory data analysis to gain deeper insights into crashes in Chicago. The research is guided by several key questions: What factors contribute most to severe crashes in Chicago? How do weather conditions affect crash severity? How do crash frequencies vary by time of day? Have recent traffic safety measures, such as speed limit reductions, led to fewer severe crashes? And finally, how does road lighting, whether daylight, dusk, or darkness, impact crash risk? These questions form the foundation for both the descriptive and predictive components of the analysis and help frame the study's relevance to traffic safety and policy planning.

Literature Review

Understanding the factors that contribute to crash severity is essential for improving road safety, particularly in metropolitan cities like Chicago. Researchers have conducted numerous studies to determine the variables that affect crash severity by analyzing human behavior, weather conditions, lighting levels, time factors, and safety performance. A review of vital studies and models serves as the basis for analyzing crash data in Chicago.

(Abdel-Aty & Radwan, 2000) provided a foundational model that links crash severity to driver-related factors, such as age, risk-taking behavior, and experience. Their study emphasized that environmental factors, like weather and road geometry, also significantly influence crash outcomes. This informs our analysis by reinforcing the importance of including variables such as driver behavior and roadway conditions in severity modeling. (Chang & Wang, 2006) used Classification and Regression Trees (CART) to explore complex interactions among crash-related variables, showing that non-linear models can outperform traditional methods. Their approach supports our use of Multinomial Logistic Regression and XGBoosting Classifier, which are modern tree-based models suitable for large datasets like ours. Multiple studies have investigated the impact of weather on crash outcomes. (Ahmed et al., 2011) applied a Bayesian hierarchical model to Colorado freeway data and found that adverse weather conditions, combined with road alignment and traffic volume, significantly increase crash risk. Although their study focused on the Colorado region, their methodology supports incorporating weather as a variable in our predictive modeling and accounting for interaction effects.

The research by Lee et al. (2023) examined Florida crash data before and after COVID-19 showed that decreased traffic volumes led to more severe crashes because of driver's risky driving behavior as roads became less busy. This highlights the need to examine weather effects in conjunction with behavioral transformations and environmental changes, as weather effects often manifest independently of these factors.

Temporal patterns in crash data are critical for identifying high-risk periods. (H. Chen et al., 2012) used logistic regression to find that the time of day, particularly nighttime hours, correlates with higher crash severity at intersections. Their findings underscore the importance of evaluating lighting conditions in tandem with crash timing, which helped in our analysis. The research by Doucette et al. (2021) studied crash patterns during the COVID-19 lockdowns and found significant shifts in crash frequencies across different times of day. This suggests that behavioral changes, including travel times, can influence when and how crashes occur. The research employs time-related factors CRASH_HOUR and CRASH_DAY_OF_WEEK, to determine peak risk durations in Chicago.

Widespread traffic safety interventions, including speed limits, signage, and enforcement, help prevent road accidents. High-risk areas benefit from reduced speed limits according to (Abdel-Aty & Radwan, 2000). The research of Goel et al. (2024) carried out a systematic review, which revealed convincing evidence showing the effectiveness of speed cameras and their implementation. However, they also noted research gaps in the effectiveness of newer technologies in lower-income regions. This encourages us to evaluate how posted speed limits correlate with crash outcomes in Chicago and identify areas where safety measures could be strengthened.

(Lahausse et al., 2010) demonstrated that public opinion about speed limits determines their success as a fatal crash prevention measure. Our study considers the posted speed limit as a key variable to explore whether current policies align with real-world outcomes. According to (H. Chen et al., 2012), crash risk shows a direct correlation to different light conditions. Intersections with adequate lighting exhibit fewer severe crashes, particularly during nighttime hours. (Nitsche et al., 2017) explored pre-crash scenarios at road junctions, using clustering to assess the role of lighting and visibility in severe crashes. Their findings support our inclusion of LIGHTING_CONDITION as a key predictor in our model.

Moreover, Chen et al. (2016) investigated injury severity in rollover crashes using Support Vector Machines and highlighted the interaction between lighting and driver behavior. Their model shows that advanced machine learning techniques can improve predictive accuracy for injury outcomes, aligning with our methodological approach.

(Durbin et al., 2015) emphasized the limitations of relying solely on police crash reports, advocating for the integration of multiple data sources (e.g., hospital records, surveys) to enhance crash data quality. Although our dataset is sourced from the City of Chicago's crash portal, this insight guides our understanding of potential data limitations and encourages careful interpretation of missing or incomplete data. (Lord & Mannering, 2010) Conducted a thorough evaluation of crash frequency modeling, which reviewed advanced statistical approaches together with their benefits and drawbacks. The insights from their research align with our decision to adopt methodologies that overcome basic regression constraints and tackle non-linear effects between variables.

The reviewed literature provides a strong foundation for our analysis of Chicago crash data. Prior research validates the selection of key variables such as weather, lighting, speed limits, and time of crash to support the use of advanced machine learning models like Multinomial Logistic Regression and XGBoosting Classifier. Studies on crash origins have revealed key information, but researchers now emphasize the importance of using modern analytical methods and real-world policy development strategies. This research extends previous work by finding crucial risk components that guide the development of practical traffic safety enhancement methods for Chicago.

Methodology

The data was collected from official sources, including the Chicago Police Department (CPD) and the Chicago Data Portal, which is available through Data.gov. The dataset includes records of vehicle crashes, with details on severity, location, time, weather conditions, road lighting, and traffic safety measures. Chicago crash data serves as a basis for analyzing crashes, identifying trends, and developing preventive measures. Analyzing this dataset can help identify various risk factors and reduce the frequency of crashes by implementing policies and strategies aimed at decreasing fatal accidents.

The dataset contains detailed information about crashes from 2020 to February 2025. The dataset had 540032 rows and 48 columns. However, the data for 2025 was found to be incomplete, covering only a partial period and lacking consistent reporting across key variables. A significant portion of the records for 2025 contained missing or null values, which could compromise the accuracy of the analysis. To maintain the reliability of the results, data from 2025 was excluded from the analysis. However, we have forecasted the number of crashes for 2025 based on historical trends. Therefore, the final dataset used in this study consists of records from 2020 to 2024. The key variable in our analysis is 'MOST_SEVERE_INJURY', which indicates the severity of the crash. The independent variables of Chicago crash data are POSTED_SPEED_LIMIT, DEVICE_CONDITION, WEATHER_CONDITION, LIGHTING_CONDITION, CRASH_HOUR, and more.

Before starting the analysis, we cleaned the dataset to make it ready for modeling. This included dropping columns with more than 5% missing data, replacing missing values in categorical columns with the label 'Unknown,' filling missing numerical values with the mean of that column, checking and correcting data types, and converting some categorical variables into numerical values for analysis. After data cleaning, we had 532037 records and 37 columns.

We then performed descriptive statistics and exploratory data analysis (EDA) to understand general patterns in the data. EDA helped us identify trends and visualize relationships between variables. For predictive modeling, we used two machine learning algorithms, such as Multinomial Logistic Regression and XGBoost Classifier. These models helped us identify which factors are associated with the severity of crashes. We have also created visualizations like bar charts and line graphs to better understand how different variables are related. Data preparation and analysis were done using Python, using libraries like pandas, matplotlib, seaborn, and scikit-learn.

Exploratory Data Analysis

To gain a comprehensive understanding of the crash data, exploratory data analysis (EDA) was performed as a preliminary step. This involved examining distributions, identifying patterns, and uncovering relationships between key variables such as crash severity, weather, lighting conditions, and time of occurrence. By visualizing and summarizing the data from 2020 to 2024, we aimed to detect trends and inform the modeling approach. The insights gained through EDA provided a foundation for feature selection.

Table 1

Descriptive Statistics

Variable	**Count**	**Mean**	**Min**	**25%**	**50%**	**75%**	**Max**	**Std Dev**
POSTED_SPEED_LIMIT	540,032	28.54	0	30	30	30	70	5.66
NUM_UNITS	540,032	2.03	1	2	2	2	18	0.47
INJURIES_TOTAL	540,032	1	1	1	1	1	1	0
INJURIES_FATAL	540,032	1	1	1	1	1	1	0
INJURIES_INCAPACITATING	540,032	1	1	1	1	1	1	0
INJURIES_NON_INCAPACITATING	540,032	1	1	1	1	1	1	0
INJURIES_REPORTED_NOT_EVIDENT	540,032	1	1	1	1	1	1	0
INJURIES_NO_INDICATION	540,032	2	2	2	2	2	2	0
INJURIES_UNKNOWN	540,032	0	0	0	0	0	0	0
CRASH_HOUR	540,032	13.28	0	9	14	17	23	5.65
CRASH_DAY_OF_WEEK	540,032	4.12	1	2	4	6	7	1.98
CRASH_MONTH	540,032	6.54	1	4	7	9	12	3.41

The descriptive statistics provide an overview of the key variables in the Chicago crash dataset. The average posted speed limit was approximately 28.5 mph, with most roads having a limit of 30 mph. The average crash hours were around 2:00 PM, 9:00 AM, and 5:00 PM. Crashes are evenly distributed throughout the week. Monthly trends show a fairly even spread as well, with the average crash occurring around June or July.

Figure 1

Number of Crashes per Year (2020–2024)

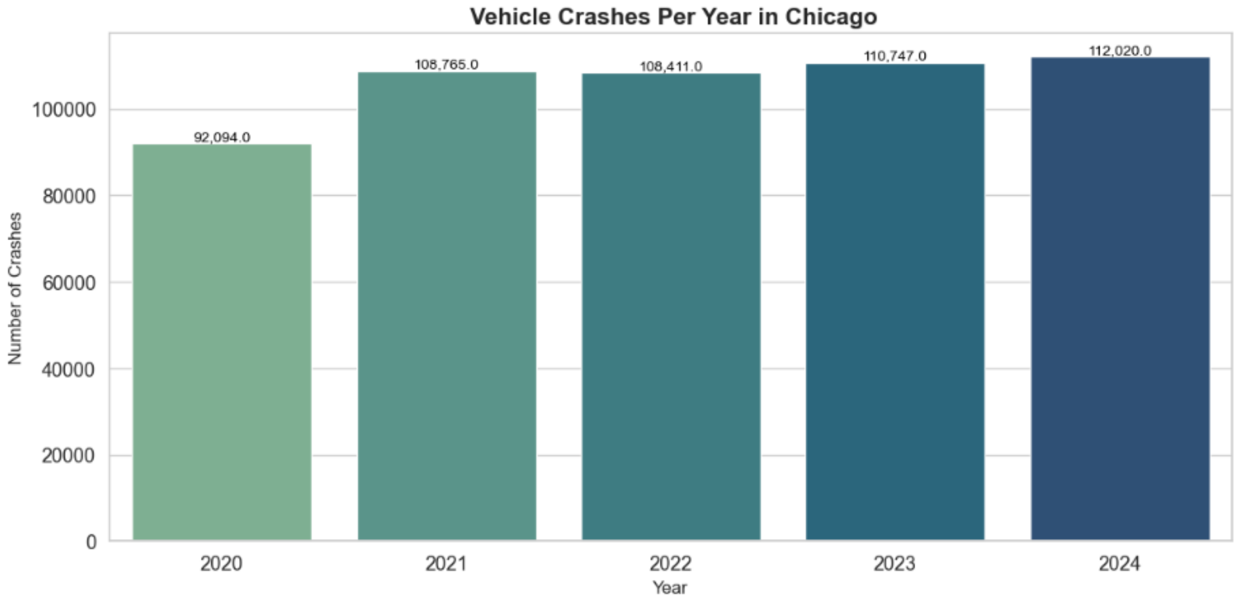


Figure 1 represents the number of crashes per year in Chicago, providing insights into yearly trends and variations in crash frequency. The x-axis denotes the years from 2020 to 2024, while the y-axis represents the number of crashes each year. The distribution of crashes appears fairly consistent from 2020 to 2024, with values around 100,000 crashes per year. The number of crashes has been increasing slightly from 2020 to 2024.

The number of crashes in 2020 is comparatively low, which may be due to policy changes, traffic regulations, or broader societal factors such as the COVID-19 pandemic. Major transportation infrastructure developments, changes in speed limits, or the introduction of new safety measures can also contribute to variations in crash rates.

Figure 2:

Speed Limits Where Crashes Occurred

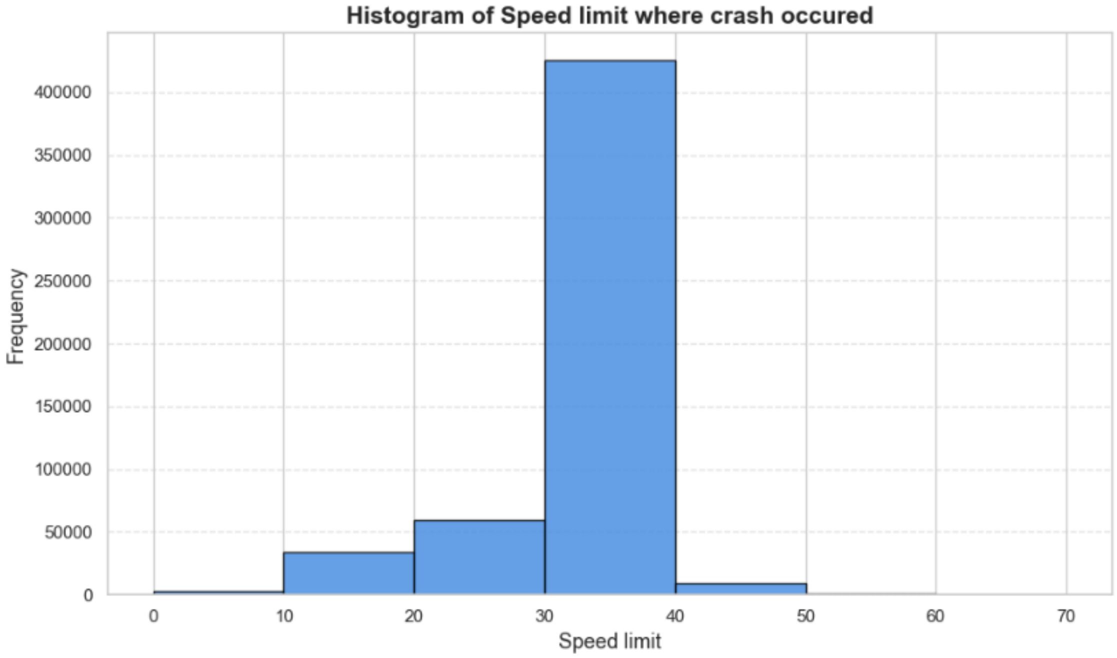


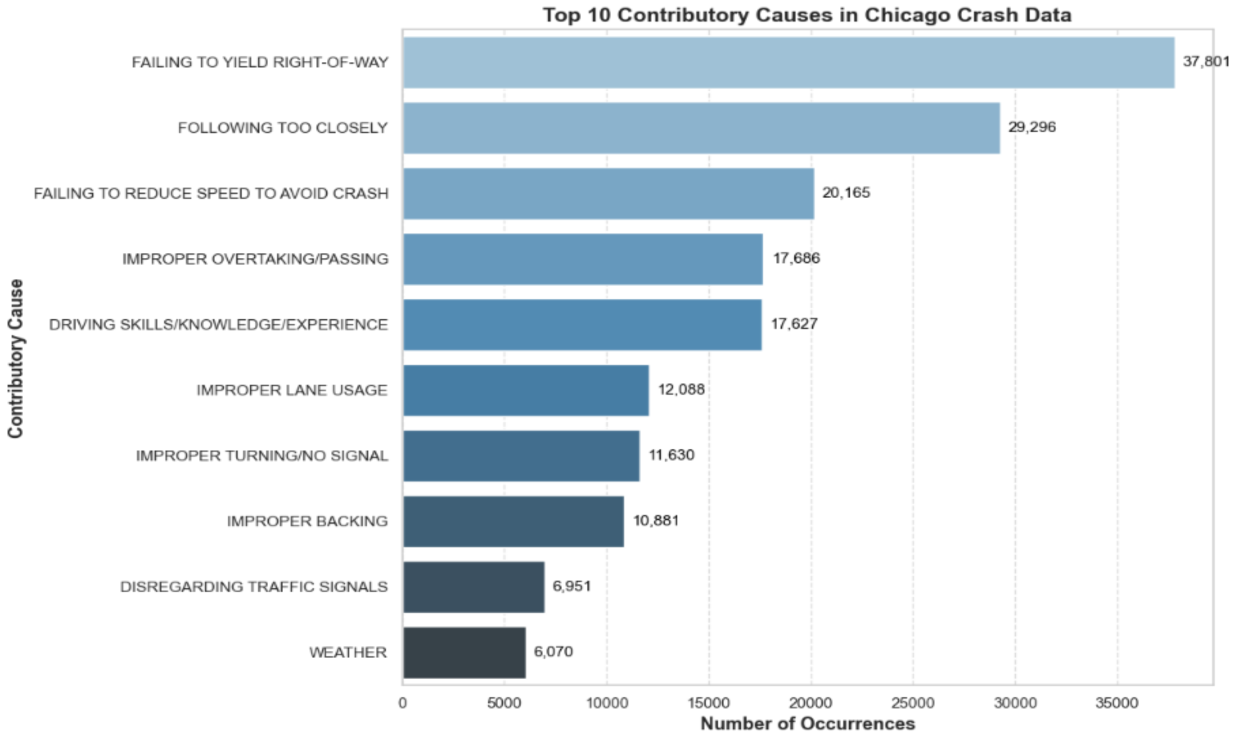
Figure 2 illustrates the distribution of speed limits where the crash occurred in Chicago, highlighting the frequency of different speed limits. The x-axis represents the speed limit, while the y-axis denotes the frequency of occurrences for each speed category. The distribution is highly left-skewed, with a significant concentration within the 30-40 speed range.

The presence of lower frequency values at the extreme ends of the histogram indicates that crashes at speeds of 0-10 and 60-70 are relatively rare. Factors influencing these variations could include traffic conditions, weather impacts, or infrastructure differences across different routes in Chicago. The peak around the 30-40 mph speed limit indicates that most crashes occur at this speed. Therefore, preventive measures should be taken to reduce the number of crashes.

Understanding the distribution of delivery speeds is crucial for optimizing logistics and ensuring timely postal services. If delays or inefficiencies are observed, further analysis into the factors influencing speed variability could help improve operational strategies. Additionally, investigating whether external factors such as road congestion, package volume, or transportation methods impact these speeds could provide insights into enhancing delivery efficiency.

Figure 3

Top 10 Contributory Causes of Crashes



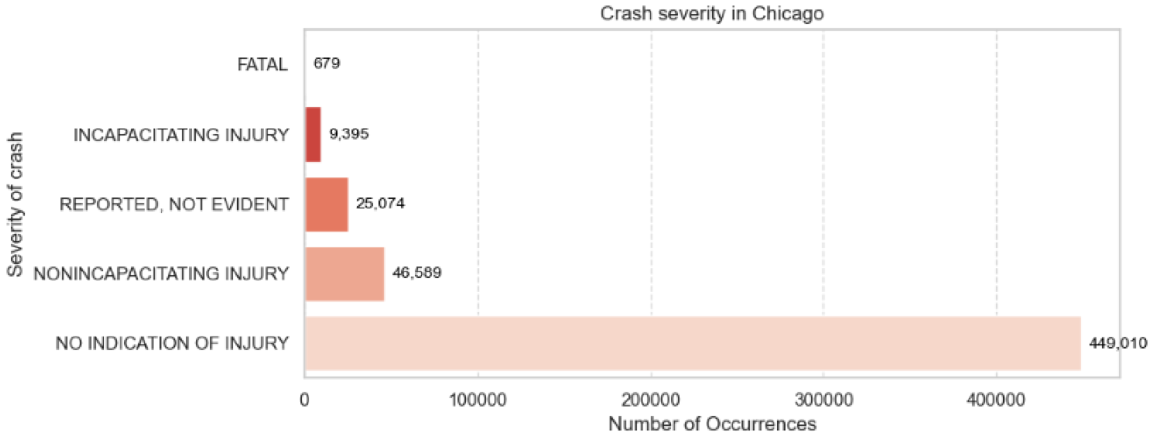
The graph above shows all the primary and secondary causes of crashes along with their frequency of occurrence. The main causes of crashes include not following traffic rules and regulations, such as failing to yield, following other vehicles too closely, failing to reduce speed, improper overtaking, and several others. Additionally, there are other contributing factors, such as driver behavior, weather conditions, and more.

To effectively reduce the number of crashes, it is crucial to understand the underlying causes. Figure 3 depicts all the causes of crashes, ranging from high-impact to low-impact factors.

By analyzing this graph, we can recommend that the public strictly follow traffic rules and also provide training programs for drivers if necessary.

Figure 4

Crash Severity Distribution



Chicago is one of the busiest cities in the United States. As the population increases, so does the number of vehicles on the road, increasing the likelihood of crashes. Figure 4 illustrates the severity of crashes in Chicago. Severity of crash is categorized into 'NO INDICATION OF INJURY', 'NONINCAPACITATING INJURY', 'REPORTED, NOT EVIDENT', 'INCAPACITATING INJURY', and 'FATAL'. According to the graph, we can say that most of the crashes do not indicate injury. However, there are a few crashes with heavy injuries and human deaths.

Figure 5

Monthly Distribution of Crashes

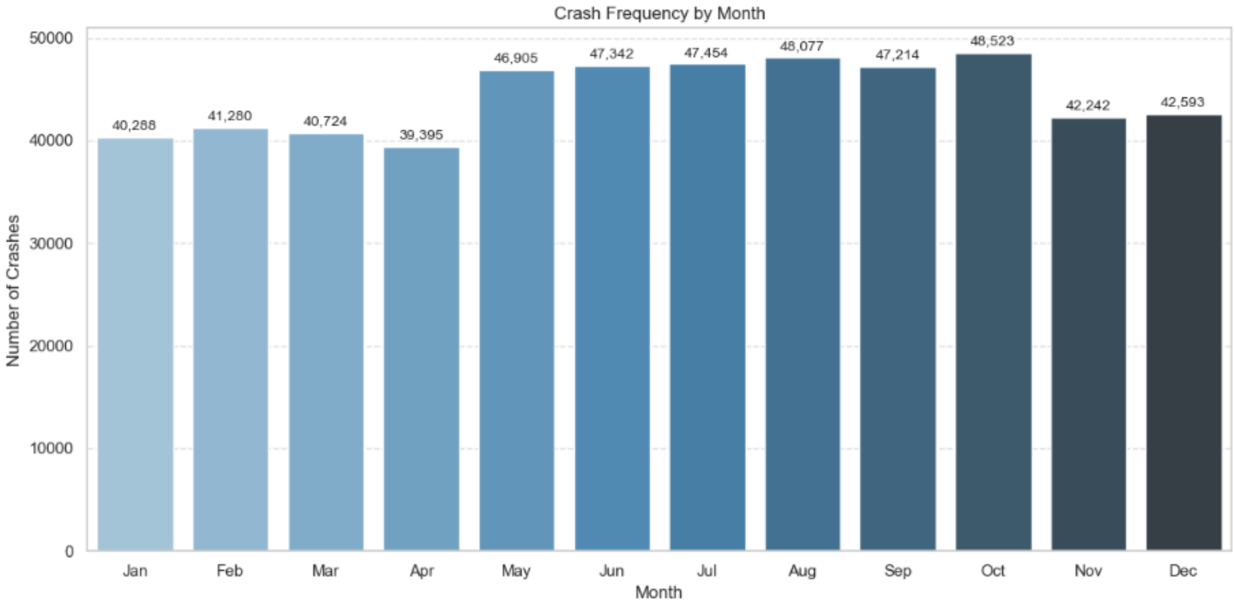
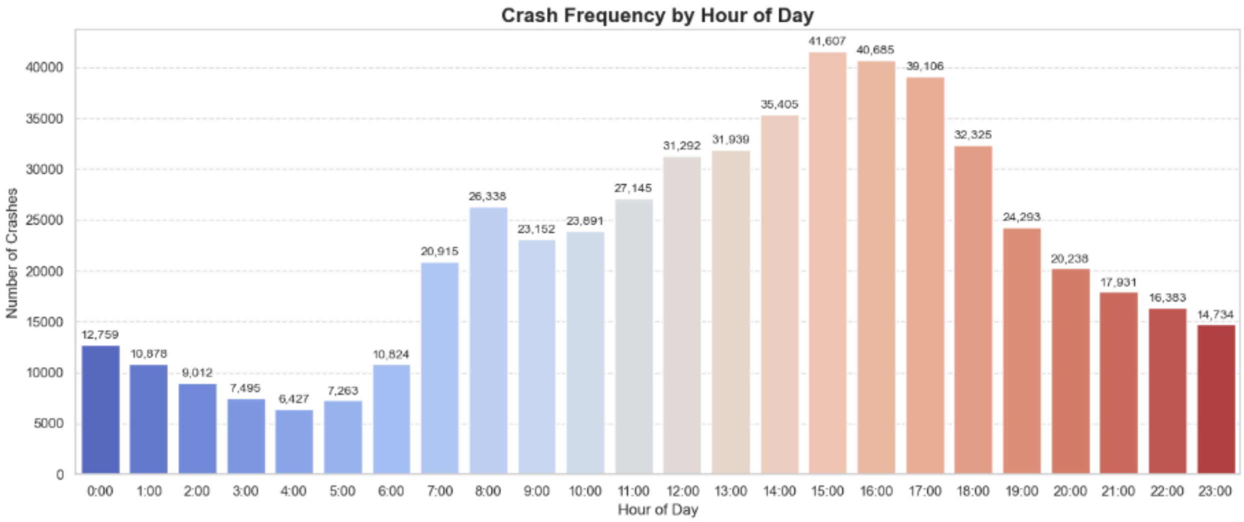


Figure 5 displays the number of crashes that occurred in each month for the year 2020 to 2024. The vertical axis represents the "Number of Crashes," ranging from 0 to 50,000. The horizontal axis shows the "Month," with each bar representing January through December.
The graph indicates a higher frequency of crashes during the warmer months, starting in May and peaking in August and October, with over 47,000 and 48,000 crashes respectively. There is a noticeable dip in crash frequency during the spring months of February, March, and April, with around 40,000 crashes respectively.

Figure 6

Hourly Distribution of Crashes



The above graph shows the occurrence of crashes at each hour of the day in a 24-hour format. The graph shows a clear pattern of crash frequency throughout the day. The number of crashes is relatively low at late night and early morning till 6 a.m. The crash frequency has increased from then, peaking significantly in the afternoon and early evening hours. The highest peak is observed between 4 PM (hour 16) and 6 PM (hour 18). From then on, the number of crashes gradually decreases throughout the evening and late night. This distribution strongly suggests a correlation between traffic volume and crash occurrences, with the busiest times of day experiencing the highest number of incidents.

Figure 7

Weather Conditions During Crashes

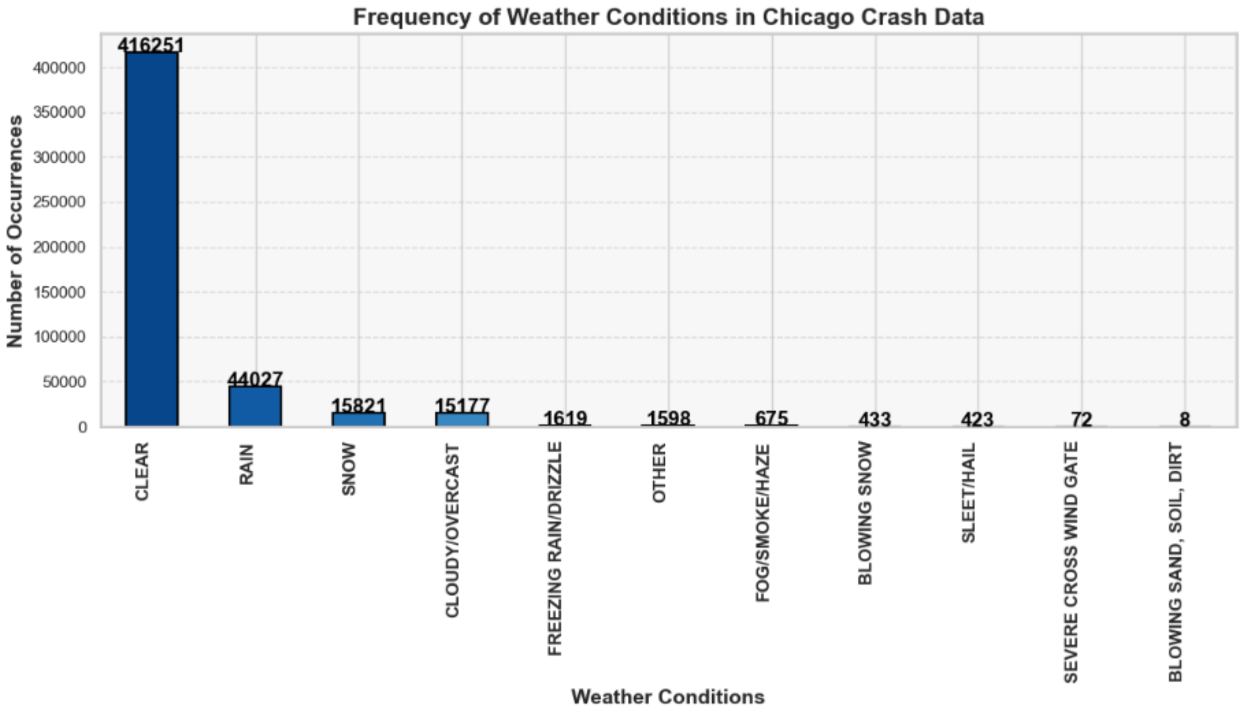


Figure 7 represents the distribution of weather conditions reported at the time of traffic crashes in Chicago. The y-axis represents the number of occurrences, and the x-axis represents different weather conditions. The most prominent weather condition during the crash is ‘CLEAR’. Hence, we can say that various other factors beyond adverse weather play a key role in the majority of crashes in Chicago. Followed by ‘CLEAR’, most of the crashes are reported during ‘RAIN’, ‘SNOW, ‘DRIZZLE’, and others.

This distribution highlights a key aspect for traffic safety research in Chicago. While adverse weather conditions do contribute to accidents, the vast majority occur under clear conditions. Therefore, investigations into the primary causes of crashes should likely focus on factors such as driver behavior, traffic, road design, and other non-weather-related variables. The relatively low frequency of crashes under more severe weather conditions might suggest effective driver adaptation or lower traffic volumes during such times.

Modeling

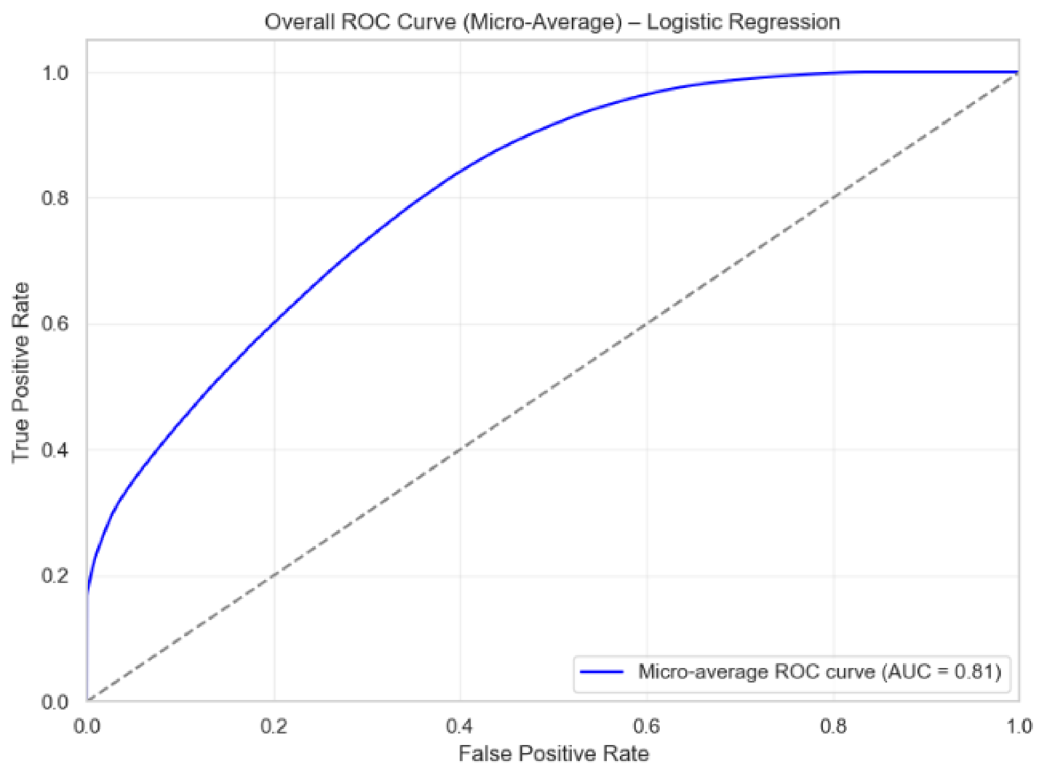
The target variable, MOST_SEVERE_INJURY, has a multiclass classification label, including no injury to fatal. We evaluated by splitting the dataset into training and test data 80-20 ratio. The target variable had a lot of imbalanced data where most crashes fall under minor or non-injury categories, so we applied the Synthetic Minority Oversampling Technique (SMOTE) to the training data. This technique generated synthetic samples of the minority classes to ensure that the model is not biased towards the majority class. We trained machine learning models with a balanced dataset, including Logistic Regression and XGBoost Classifier. Among them, the XGBoost classifier is effective due to its ability to capture complex, non-linear relationships in the data. The "Unknown" category was removed from the target variable, as it does not represent a meaningful or actionable crash severity level. Including it introduces noise and ambiguity, potentially reducing the ability to learn clear patterns among the valid severity classes.

The multinomial logistic regression model achieved an overall accuracy of 49%, indicating moderate performance in predicting the severity of crash injuries across multiple classes. The class-wise metrics show significant variability, reflecting challenges in handling the class imbalance in the dataset. The model performed best on the "No Indication of Injury," achieving F1-scores of 0.91, with relatively high precision and recall. However, its performance was notably poor in critical injury-related classes such as "Incapacitating Injury" and "Nonincapacitating Injury", with F1-scores of 0.26 and 0.24, respectively. The macro and weighted average scores for precision, recall, and F1-score of 0.49. Despite its interpretability, logistic regression was not sufficiently complex to capture the relationships among features. These findings suggest that while

SMOTE has partially addressed class imbalance, logistic regression lacks the flexibility to effectively capture the nonlinear patterns in crash data.

Figure 8

ROC Curve



To further evaluate the Multinomial Logistic Regression model, a micro-averaged Receiver Operating Characteristic (ROC) curve was plotted as shown in Figure 8. The ROC curve shows the trade-off between the true positive rate and false positive rate across different classification thresholds. In the above figure, the curve bows noticeably towards the top-left corner, indicating a relatively strong performance. The Area Under the Curve (AUC) value is 0.81, which suggests that the model possesses a good overall score, considering all classes. Despite the model's struggles with individual class recall and imbalance, the ROC-AUC score reflects its ability to distinguish between positive and negative outcomes at a global level.

Figure 9

Time series Forecasting using Prophet

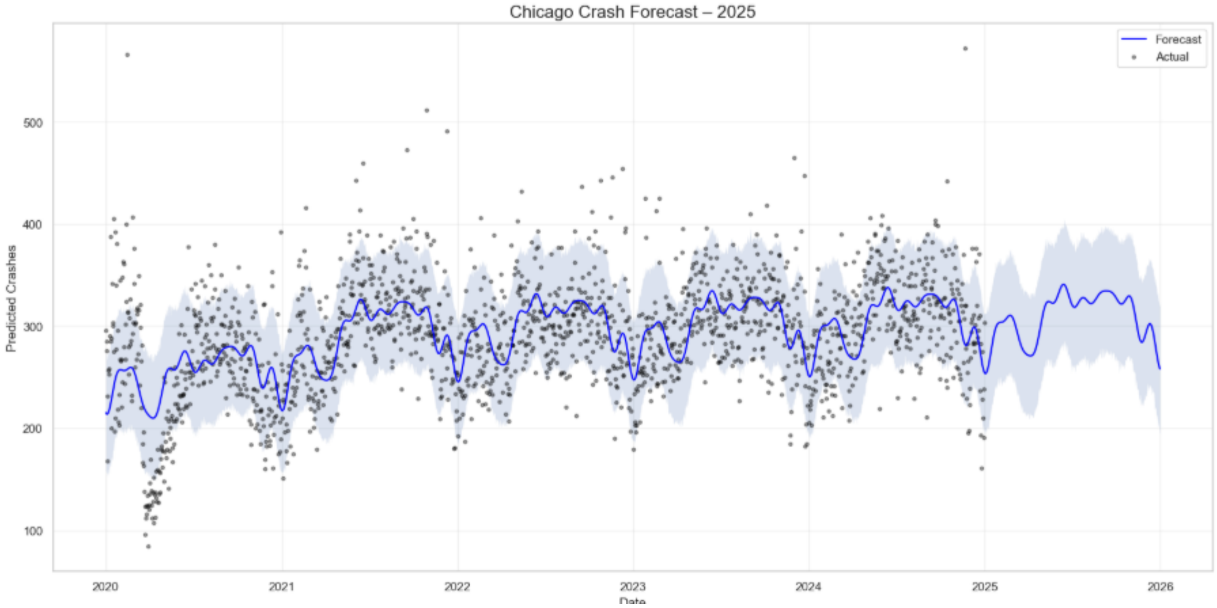
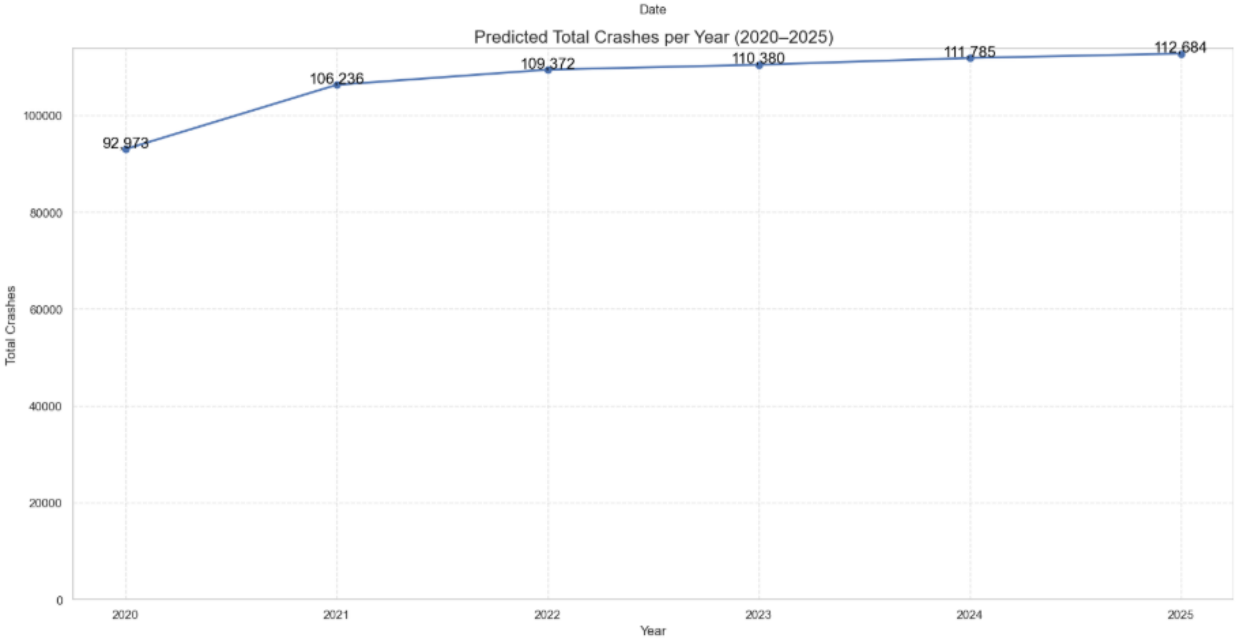


Figure 10

Time series Forecasting for 2025

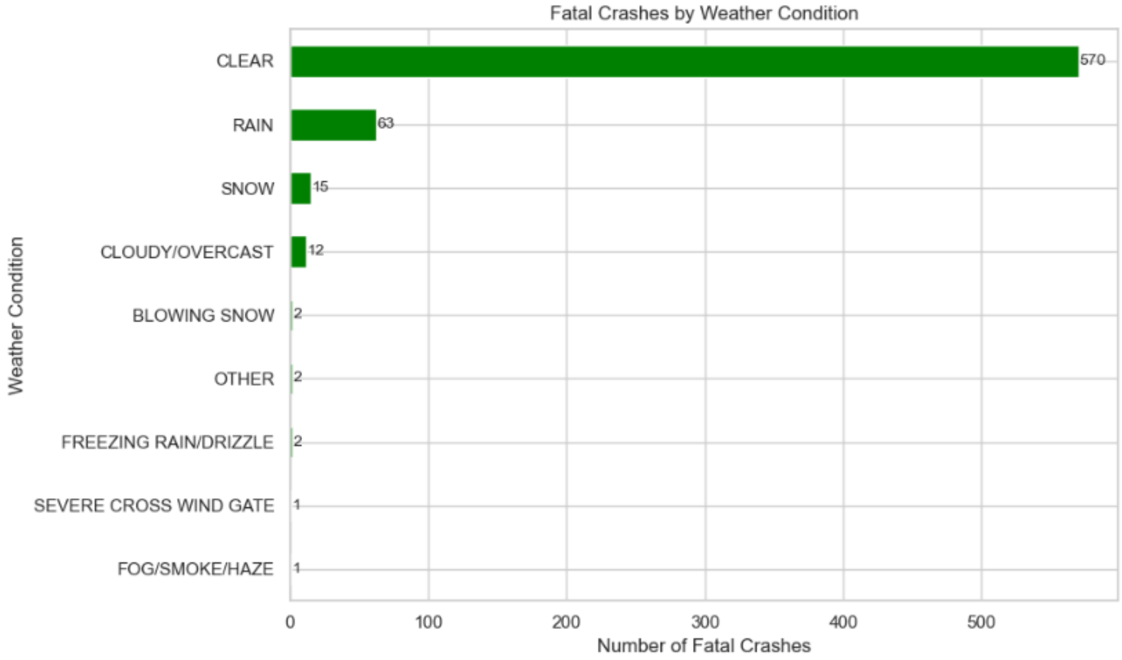


To forecast future traffic crash trends in Chicago, we used Prophet time series forecasting model. Historical crash data from 2020 to 2024 was used to train the model, and predictions were extended through the end of 2025. Figure 9 shows the predicted number of daily crashes, showing both the forecast (blue line) and the actual historical values (black dots), along with a shaded confidence interval representing uncertainty bounds. The model also displays a stable upward trend over time, indicating a slow but steady increase in crash frequency.

Figure 10 shows forecasts revealing a consistent upward trend in total crashes from 92,973 in 2020 to 112,684 in 2025. This represents an approximate 21% increase over six years, and a 0.8% increase between 2024 and 2025. This may be due to traffic volume.

Figure 11

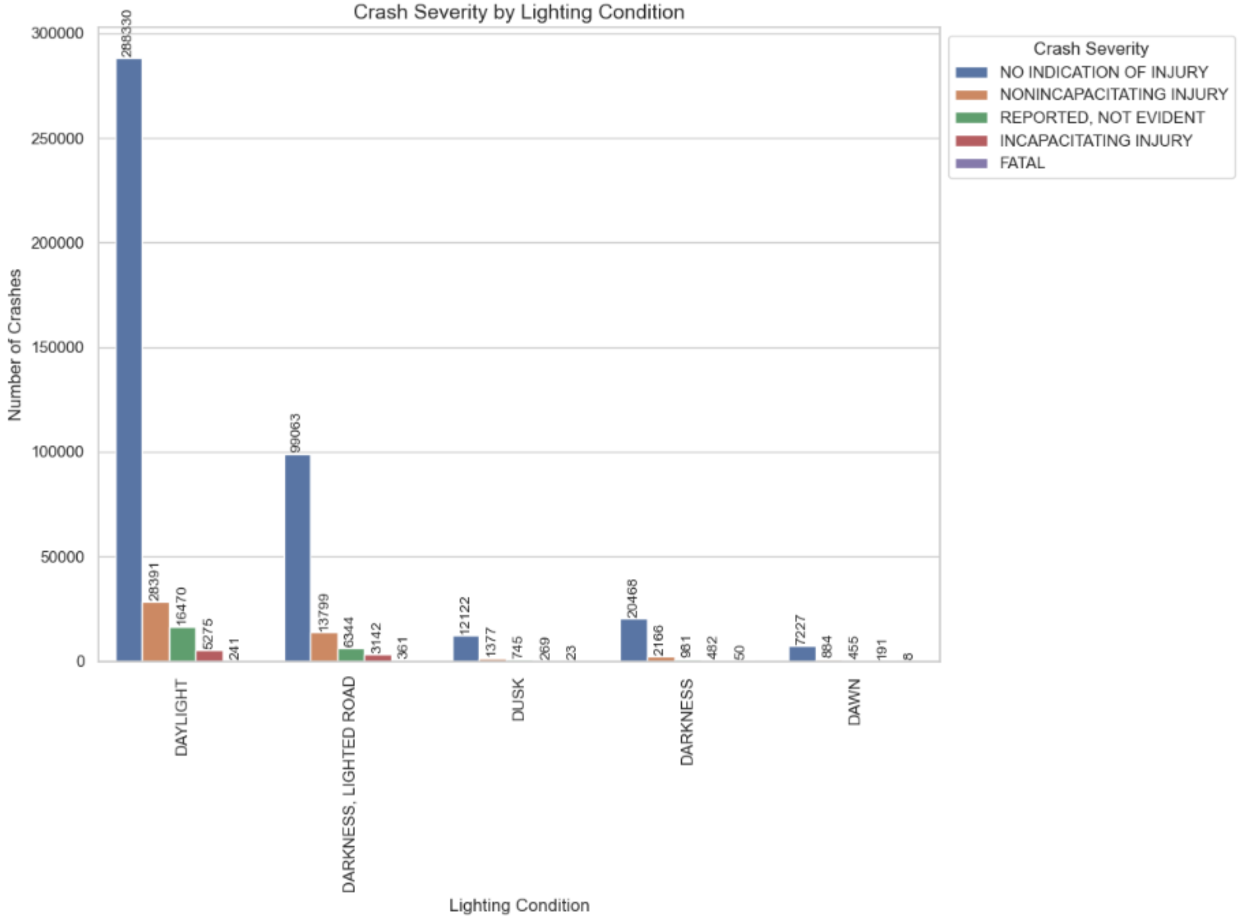
Fatal Crashes by Weather Conditions



As our main aim was to focus on identifying the factors contributing to crashes and their severity, we considered only fatal crashes and created a bar graph to analyze the impact of weather conditions. According to Figure 11, most fatal crashes occurred during clear weather, followed by rain and snow. This suggests that weather may not be the only factor influencing severe crashes, but it can still contribute to certain cases. Our analysis also revealed that a significant number of severe crashes occurred between May and October, particularly during rainy and snowy conditions.

Figure 12

Crash Severity by Lighting Conditions



In Figure 11, we plotted a bar graph differentiating crash severity by lighting conditions to understand the occurrence of crashes. From the graph, we can observe that most crashes occurred during daylight, followed by darkness and dusk. Focusing specifically on the fatal and evident injury categories, it is evident that a majority of these crashes occurred at night under lighted road conditions, followed by crashes during dusk.

We have also used an XGBoost classifier to predict crash severity for our imbalanced dataset.

Table 2:

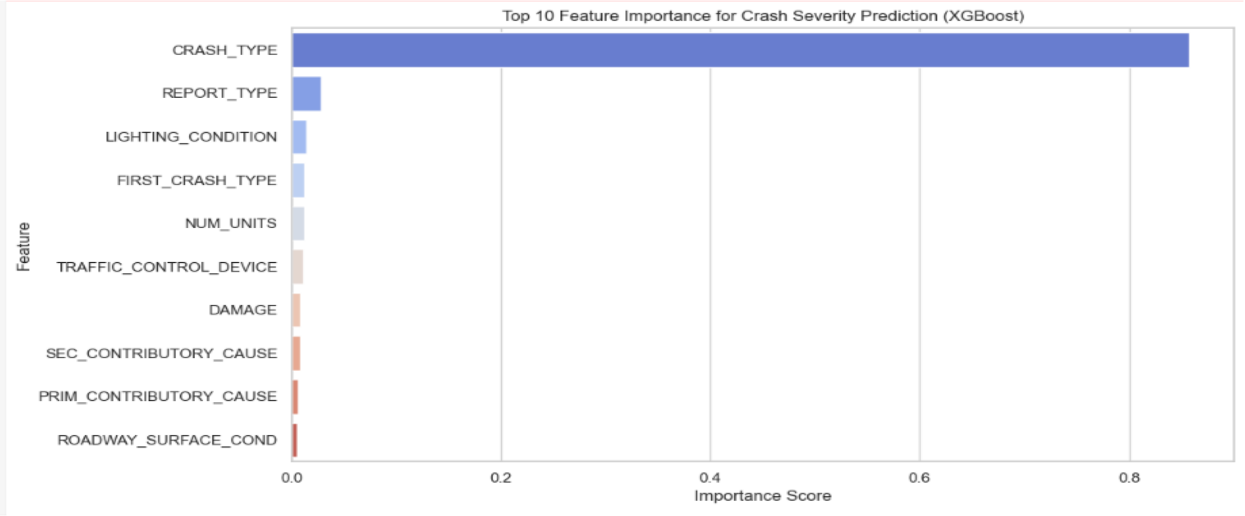
Classification Report for XGBoost Classifier

Class	Precision	Recall	F1-Score
FATAL	0.86	0.96	0.91
INCAPACITATING INJURY	0.61	0.60	0.60
NO INDICATION OF INJURY	0.92	0.95	0.94
NONINCAPACITATING INJURY	0.54	0.46	0.50
REPORTED, NOT EVIDENT	0.60	0.60	0.60

The XGBoost classifier showed improved predictions of crash severity levels compared to logistic regression. It achieved an overall accuracy of 72%, which means it correctly predicted the severity of crashes in 72 out of every 100 cases. The model also performed consistently across different severity categories, as shown by its macro and weighted F1-scores of 0.71. Among all classes, the model performed best for the “Fatal” and “No Indication of Injury” categories. However, the model had more difficulty distinguishing between the mid-range injury categories. Overall, the XGBoost model predicted high-risk and low-risk crash outcomes, especially when the crash severity is very high (Fatal) or very low (No Indication of Injury).

Figure 13

Feature Importance



To understand the factors influencing crash severity predictions, feature importance was analyzed using the trained XGBoost model. The chart above highlights the top 10 features. The most important feature is CRASH_TYPE, which had a significantly higher importance score than any other variable, even after using SMOTE. This means that the type of crash (e.g., rear-end, fatal, injuries) played a major role in determining how severe the crash was. Other important features included lighting conditions, the number of vehicles involved in the crash, the primary and secondary causes, and more. The graph shows that while several features influence crash severity, CRASH_TYPE is the dominant predictor, and the model relies heavily on it.

Results

The analysis of the Chicago crash dataset from 2020 to 2024 revealed several significant trends and insights. Descriptive statistics indicated that the average posted speed limit at crash locations was approximately 28.5 mph, with most crashes occurring at 30 - 40 mph. Crashes were relatively evenly distributed across days of the week and months, with a slight increase during the warmer months of May through October. Hourly patterns showed a peak in crash frequency between 4 PM and 6 PM, coinciding with evening rush hours. The top 10 contributory causes of crashes showed that violations of traffic rules, such as following too closely, failure to yield, or speeding, are the primary factors in crash occurrence. The majority of crashes resulted in no indication of injury, although several resulted in incapacitating or fatal outcomes.

Using the Multinomial Logistic Regression model, the dataset was evaluated for crash severity classification. The model achieved an accuracy of 49%, with a particularly high F1-score (0.91) for the "No Indication of Injury" class. Despite the use of SMOTE to address class imbalance, the logistic regression model struggled with accurately classifying the minority classes. The XGBoost classifier showed better performance than logistic regression. It achieved an overall accuracy of 72% and offered more balanced performance across all crash severity classes. The XGBoost model was particularly effective in predicting both "Fatal" and "No Indication of Injury" classes. The Feature Importance analysis revealed that CRASH_TYPE was the most influential variable, followed by lighting condition, number of vehicles involved, and contributory causes in crash severity.

Time series forecasting using the Prophet model showed an upward trend in the total number of crashes occurring yearly, from 92,973 crashes in 2020 to 112,684 crashes in 2025, a 21% increase over six years. This forecast highlights the need to monitor traffic crashes closely, as they remain a significant public safety concern. Fatal crashes were most commonly associated with clear weather, followed by rain and snow. When analyzing lighting conditions, the majority of fatal and severe injury crashes occurred at night under artificial lighting, followed by dusk.

Interpretation

The analysis of Chicago's crash data reveals several key factors influencing crash severity, including weather conditions, time of day, lighting, and posted speed limits. Adverse weather conditions, particularly rain and snow, were frequently associated with fatal crashes, likely due to reduced visibility and slippery road surfaces. Similarly, lighting conditions significantly affected crash outcomes. Severe injuries and fatalities occurred during nighttime and dusk, especially in poorly lit areas. These insights emphasize the need for seasonal road safety measures, improved street lighting, and driver awareness campaigns during low-visibility conditions.

Crash occurrences were notably higher during peak traffic hours, such as late evening. Intersections and high-traffic zones with inadequate infrastructure, such as poorly maintained roads, were also identified as hot spots for severe crashes. These patterns highlight the importance of implementing infrastructure improvements, such as better-designed intersections, reflective signage, and optimized traffic signal timings, especially in high-risk areas. Measures such as speed cameras, law enforcement, and lower speed limits in vulnerable zones could significantly mitigate crash impact. Future research incorporating AI-driven traffic monitoring and autonomous vehicle technologies could offer a more comprehensive approach. Nonetheless, these findings provide a strong foundation for data-driven policymaking and can guide other cities in designing targeted interventions to improve road safety and reduce traffic-related fatalities.

Conclusion

Using machine learning models such as Multinomial Logistic Regression and XGBoost for Chicago crash data, we found that weather conditions, time of day, lighting, and speed limits significantly impact the severity of crashes. The XGBoost model outperformed logistic regression, demonstrating strong predictive power, particularly for fatal crashes. Time series forecasting further indicated a steady rise in crash occurrences through 2025, emphasizing the need for proactive safety interventions. Our findings support the implementation of targeted measures such as improved lighting, stricter speed regulations, infrastructure upgrades, and data-informed traffic policies. By leveraging these results, policymakers and transportation authorities can work towards building safer, smarter cities.

Limitations

While this study offers valuable insights into the factors influencing crash severity in Chicago, it has a few limitations. First, the dataset lacks critical behavioral and contextual information such as driver distraction, alcohol or drug influence, seatbelt usage, and the driver's age. These factors can significantly influence crash outcomes, but were not captured in the Chicago crash data.

Second, although we addressed class imbalance using SMOTE, this technique may introduce synthetic patterns that do not fully reflect real-world scenarios, potentially impacting model generalizability. Additionally, the exclusion of the "Unknown" category from the target variable may have removed potentially informative instances. Including such sensor data could enhance model accuracy and support the development of dynamic, real-time traffic safety systems. Future research should explore integrating diverse data sources and emerging technologies such as AI-powered traffic surveillance and autonomous vehicle data to improve predictive modeling and policy effectiveness.

Future Recommendations

Based on the findings, a few recommendations exist to improve traffic safety and reduce crash severity in Chicago and other urban areas. First, there is a high-risk time specifically late afternoons and early evening hours, and a few areas are identified as crash-prone hotspots. Increasing police presence, public awareness campaigns, and innovative traffic control measures during these times could reduce the number of crashes. Additionally, while lighting was not among the top predictors in the model, its visual correlation with injury severity suggests that improving road lighting, especially at intersections and unlit zones, could help reduce serious crashes.

Secondly, Researchers and policymakers should incorporate proper sampling techniques. Policymakers should also consider expanding the dataset to include behavioral variables, possibly through integration with hospital or insurance records, to gain more insights into the crashes. Finally, investment in real-time data collection technologies, such as connected vehicle infrastructure and traffic sensors, can enhance the timeliness of crash data, allowing for more dynamic safety interventions. Incorporating these can guide city planners and traffic safety authorities in developing proactive strategies to make Chicago's roads are safer for all users.

References

Abdel-Aty, M. A., & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, *32*(5), 633–642. https://doi.org/10.1016/S0001-4575(99)00094-9

Ahmed, M., Huang, H., Abdel-Aty, M., & Guevara, B. (2011). Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. *Accident Analysis & Prevention*, *43*(4), 1581–1589. https://doi.org/10.1016/j.aap.2011.03.021

Chang, L.-Y., & Wang, H.-W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention*, *38*(5), 1019–1027. https://doi.org/10.1016/j.aap.2006.04.009

Chen, C., Zhang, G., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention*, *90*, 128–139. https://doi.org/10.1016/j.aap.2016.02.011

Chen, H., Cao, L., & Logan, D. B. (2012). Analysis of Risk Factors Affecting the Severity of Intersection Crashes by Logistic Regression. *Traffic Injury Prevention*, *13*(3), 300–307. https://doi.org/10.1080/15389588.2011.653841

Durbin, D. R., Myers, R. K., Curry, A. E., Zonfrillo, M. R., & Arbogast, K. B. (2015). Extending the value of police crash reports for traffic safety research: collecting supplemental data via surveys of drivers. *Injury Prevention*, *21*(e1), e36–e42. https://doi.org/10.1136/injuryprev-2014-041155

Goel, R., Tiwari, G., Varghese, M., Bhalla, K., Agrawal, G., Saini, G., Jha, A., John, D., Saran, A., White, H., & Mohan, D. (2024). Effectiveness of road safety interventions: An evidence and gap map. *Campbell Systematic Reviews*, *20*(1). https://doi.org/10.1002/cl2.1367

Lahausse, J. A., van Nes, N., Fildes, B. N., & Keall, M. D. (2010). Attitudes towards current and lowered speed limits in Australia. *Accident Analysis & Prevention*, *42*(6), 2108–2116. https://doi.org/10.1016/j.aap.2010.06.024

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, *44*(5), 291–305. https://doi.org/10.1016/j.tra.2010.02.001

Nitsche, P., Thomas, P., Stuetz, R., & Welsh, R. (2017). Pre-crash scenarios at road junctions: A clustering method for car crash data. *Accident Analysis & Prevention*, *107*, 137–151. https://doi.org/10.1016/j.aap.2017.07.011

Stevenson, M., Harris, A., Mortimer, D., Wijnands, J. S., Tapp, A., Peppard, F., & Buckis, S. (2018). The effects of feedback and incentive-based insurance on driving behaviours: study approach and protocols. *Injury Prevention*, *24*(1), 89–93. https://doi.org/10.1136/injuryprev-2016-042280

Publisher data.cityofchicago.org. (2025, March 1). City of Chicago - traffic crashes - crashes. Catalog. https://catalog.data.gov/dataset/traffic-crashes-crashes

Doucette, M. L., Tucker, A., Auguste, M. E., Gates, J. D., Shapiro, D., Ehsani, J. P., & Borrup, K. T. (2021). Evaluation of motor vehicle crash rates during and after the COVID-19-associated stay-at-home order in Connecticut. Accident Analysis & Prevention, 162, 106399. https://doi.org/10.1016/j.aap.2021.106399

Lee, J., Liu, H., & Abdel-Aty, M. (2023). Changes in traffic crash patterns: Before and after the outbreak of COVID-19 in Florida. Accident Analysis & Prevention, 190, 107187. https://doi.org/10.1016/j.aap.2023.107187