**CHICAGO TRAFFIC CRASH ANALYSIS AND MODELING**

**University of Maryland, Baltimore County**

**Department of Data Science**

**Data 606: Capstone Project**

**Professor Ozgur Ozturk**

**By Srikanth Siddamsetty and Sai Charan Kolluru**

**May 13, 2024**

|  |  |
| --- | --- |
| **CONTENT** | **PAGE NUMBER** |
| Abstract | 3 |
| Introduction | 4 |
| Dataset | 5 |
| Workflow | 6 |
| Area of Study | 8 |
| Methodology | 10 |
| Results and Analysis | 11 |
| Models | 19 |
| References | 23 |

**ABSTRACT**

This study examines a wide range of factors that affect crash frequency, severity, and trends in traffic crashes in Chicago. Temporal patterns reveal notable variations in the number of crashes over the course of years and months, suggesting long-term and seasonal trends. Higher crash rates occur on some days of the week, suggesting potential connections between human behavior and traffic flow patterns. The study also examines how the time of day affects the likelihood of collisions, highlighting different patterns in the morning, afternoon, evening, and night.

Geographical analysis pinpoints particular locations, addresses, beats, or regions with elevated crash rates, suggesting potential hotspots for focused intervention. An analysis of stated speed limits, types of roadways, and directions reveals relationships with crash frequency, underscoring the significance of infrastructure design and regulation in accident avoidance. The frequency and severity of collisions are influenced by a number of factors, including traffic, road conditions, weather, traffic control devices, and alignment. In addition, the study examines demographic correlations, collision characteristics, injury patterns, and the relationship between traffic volume and speed limits. These comprehensive analyses provide valuable insights for enhancing road safety and reducing accident-related injuries and fatalities in Chicago.

**INTRODUCTION**

Road traffic accidents are a major public health and safety problem affecting individuals, communities and urban infrastructure. Understanding the dynamics of traffic accidents is critical to successful law enforcement, urban planning, and safety programs in Chicago, as in many other large cities. The purpose of this study is to provide a complete analysis of traffic crashes in Chicago, focusing on the many characteristics that influence crash rates, severity, and trends. In Chicago, a busy metropolitan center, there is a complex interplay of factors contributing to traffic accidents. Temporal patterns in crash data show variations between years and months, suggesting seasonality and long-term trends that require complex responses. Additionally, different days of the week had higher accident rates, suggesting possible relationships with traffic volumes, commuting patterns, and behavioral variables. Understanding this temporal dynamic is crucial for creating targeted tactics to reduce the risk of accidents and improve overall road safety.

The purpose of this study is to investigate collision characteristics, injury patterns, demographic correlations, and the relationship between traffic volume and speed restrictions in order to give actionable insights for stakeholders involved in traffic safety management. Lastly, the study's conclusions can influence the development of evidence-based policies that will enhance traffic safety, reduce fatalities and injuries, and advance sustainable transportation options in Chicago and other cities.

**DATASET**

The dataset that is being examined is an extensive collection of traffic crash data that contains specific information about traffic incidents that have been recorded by law enforcement. This dataset, which has a size of 429 MB and 832,299 rows, is intended to support many types of research related to traffic safety and urban planning.

A CRASH\_RECORD\_ID, which also acts as a connection to relevant vehicle and individual data in associated datasets, uniquely identifies each record in the dataset. The dataset has a wide range of characteristics that capture different facets of every collision incident, including as:

**Temporal details**: The precise time and date of the collision, as well as deduced parameters such the day of the week, month, and hour, are essential for examining the temporal trends in traffic accidents.

**Spatial details**: Geographical coordinates (latitude and longitude) and a thorough crash site, allowing for the identification of accident hotspots and the evaluation of the efficacy of traffic management strategies through spatial analysis.

**Environmental and situational context**: The weather, illumination, and road conditions at the time of the collision are important factors to consider when analyzing how the environment affects traffic safety.

**Crash specifics**: Details on the kind of collision, the state of traffic control devices, the kind and state of the road, and involvement in work zones might provide light on the causes that lead to accidents.

**Human factors:** Information about the quantity of units involved, the kinds of injuries received, and the primary and secondary contributing factors to the collision provide a thorough analysis of the human factor in traffic accidents.

This dataset provides insights that can inform policy decisions and focused initiatives meant to lower the number of traffic accidents on the road and improve public safety. It is a fundamental instrument for empirical research in the field of traffic safety.

**WORKFLOW**  
The analysis of traffic crashes in Chicago was conducted using a dataset that included extensive data on crash occurrences, including temporal, spatial, and demographic characteristics. Python programming and tools like Pandas and Matplotlib have made data processing, purification, and visualization easier. Relevant factors such as crash year, month, day of week, hour, location, street direction, trafficway type, weather, injuries, and deaths were initially filtered out before the dataset was constructed. Descriptive statistics were used to understand the distribution of crash data over a number of temporal and spatial variables.

We examined the annual and monthly crash statistics to investigate the temporal patterns in traffic accidents in Chicago. To visualize these counts and look for any repeating trends or changes over time, we employed bar charts. Furthermore, we looked at crash counts according to the day of the week and time of day to find patterns in the frequency of crashes over time. In order to identify high crash rates, or "crash hotspots," we conducted a geographical analysis that entailed mapping crash counts according to location, street direction, and kind of trafficway. We used scatterplots to efficiently display these geographical patterns, which allowed us to comprehend the distribution of crash locations, injuries, and fatalities around Chicago.By examining crash data on an annual and monthly basis, we were also able to examine temporal trends on a finer scale and understand changes and patterns in crash frequency over time. In addition, we employed line plots to examine trends in the number of crashes resulting in injuries and fatalities broken down by crash month and time of day. This method offered insightful information about the daily and monthly variations in crash rates.

The process of gathering and preparing data—that is, creating a dataset with complete crash information—is the first step in the workflow for examining temporal trends in traffic crashes in Chicago. Monthly and annual crash counts are computed to identify patterns over time, and bar charts are used to clearly show such patterns. The analysis includes crash counts broken down by time of day and day of the week, offering insights into daily and weekly fluctuations in crash probability. Scatterplots are used in geographic analysis to visualize crash locations and severity spatially. Crash hotspots are mapped according to location, trafficway type, and street direction. After that, yearly and monthly temporal trends are investigated. To capture daily and monthly variations in crash rates, line graphs representing time of day trends are used.

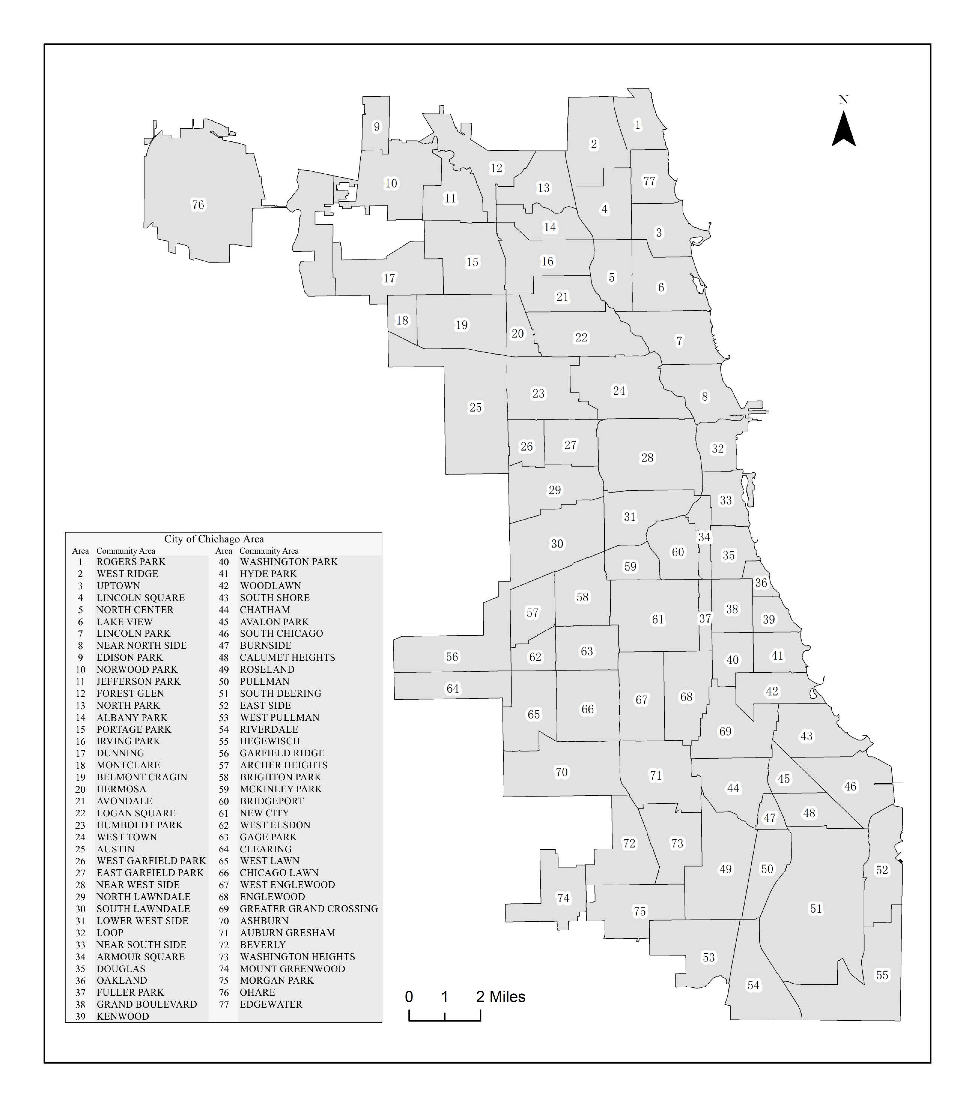
A diagram of a road with various icons

Description automatically generated with medium confidence

**AREA OF STUDY:**

An analysis of data from a comprehensive study of traffic crashes in Chicago between 2014 and 2024 reveals that a significant portion of accidents—roughly 41.7% of all recorded incidents—occur during the midday hours. At 339,924 incidents, this midday high exceeds the morning total of 241,005 (29.6%), evening total of 138,796 (17%), and nighttime total of 94,384 (11.6%), indicating a relationship between traffic volume and crash frequency. A more thorough analysis of temporal patterns shows a startling increase in fatal and overall injuries over the course of the decade. The number of fatal injuries peaked in 2023 at fifteen cases in January, a significant increase from the zero cases recorded in the same month in the years 2013 to 2015.

Given that Chicago is the third-largest city in the United States with a population of around 3 million, spatial analysis within the 77 community areas—a city division created in the 1920s for sociological study—offers insights into crash distribution throughout the city. Data indicates that certain places may be more vulnerable to both fatal and non-fatal accidents. The total number of injuries peaked at 2,362 in June 2018, suggesting that community-specific factors may have an impact on road safety. The concentration of crashes in particular areas necessitates the application of data-driven traffic safety measures as well as focused interventions. These measures might be customized to suit the particular difficulties faced by the more vulnerable locations found through the investigation, and they could include community safety education programs, enhanced traffic regulation enforcement, and infrastructure upgrades.



**METHODOLOGY:**

**Anova test and Tukey’s test**

In order to analyze the variance in crash frequencies, we used ANOVA and Tukey's Honestly Significant Difference (HSD) tests to look at the temporal trends of traffic crashes in Chicago over a period of years. With a p-value close to zero and a very significant F-statistic of 1535.181, the ANOVA findings showed significant variations in the number of crashes between various years. This implies that the average annual crash rate fluctuates, underscoring the impact of changing circumstances throughout time.

Tukey's HSD test was used to further analyze the data to find particular years, especially 2024, that significantly deviated from the rest. These differences may indicate abnormal trends or impacts resulting from outside influences like modifications to traffic laws, improvements to urban infrastructure, or noteworthy social events. These results highlight the need for continuous observation and focused interventions in years that are statistically distinct in order to improve road safety and successfully lower crash rates.

**Chi2test**  
A highly significant statistical link between collision types and key contributory reasons in traffic events was found by using the Chi-square test. The results showed a Chi-square statistic of 1,075,594.2335 with a p-value of almost zero. This implies that particular crash kinds are more likely to happen when certain contributing factors are present than would be predicted by chance alone. The research highlights the possibility for focused interventions and policy formulations that address common contributing factors to lower the incidence and severity of traffic crashes. The analysis is based on a contingency table with 663 degrees of freedom.

**RESULTS AND ANALYSIS**:

**Temporal Trends**

**Is there a significant difference in the number of crashes across different years or months?**

From 2013 to 2018, there was a discernible increase of crashes, which peaked in that year and then gradually decreased in the years that followed. Nevertheless, no discernible seasonal trend is seen when we look at the data on a monthly basis. There is no discernible trend from month to month, with some months—like September and October—exhibiting greater crash counts and others—like April—showing lower counts.

A graph of a crash

Description automatically generated

### Yearly Crash Counts:

* The number of crashes seems to decrease slightly from 2016 to 2020.
* There is a noticeable decrease in crash counts from 2016 to 2017, followed by a relatively stable trend until 2020.

### Monthly Crash Counts:

* There are fluctuations in crash counts across months, but there is no clear increasing or decreasing trend.
* The highest number of crashes seems to occur in the early months of the year, particularly in January and February.
* There is a slight decrease in crash counts during the summer months (June, July, August).

Overall, the data suggests a slight decrease in crashes over the years, while monthly crash counts fluctuate without a clear trend.

**Do certain days of the week have a higher frequency of crashes compared to others?**

A graph with orange bars

Description automatically generated

### Crash Counts by Day of Week:

* The highest number of crashes occurs on Fridays, followed by Thursdays and Saturdays.
* Sundays and Mondays have relatively lower crash counts compared to other days of the week.
* The difference in crash counts across different days of the week is noticeable, with Friday being the day with the highest number of crashes.

**Does the time of day (morning, afternoon, evening, night) impact the likelihood of a crash?**

A graph of a crash

Description automatically generated

### Crash Counts by Time of Day:

* The data shows that the highest number of crashes occurs during late afternoon and early evening hours, particularly between 3:00 PM and 6:00 PM.
* There is a noticeable decrease in crash counts during the late night and early morning hours, with the lowest number of crashes occurring between 2:00 AM and 6:00 AM.
* Crash counts start to rise again from 6:00 AM, reaching a peak in the late afternoon hours.

This suggests that late afternoon and early evening hours are associated with a higher likelihood of crashes, while late night and early morning hours see fewer crashes.

**Geographical Analysis:**

**Are there specific locations (addresses, beats, or areas) where crashes are more prevalent?**

A screenshot of a computer

Description automatically generated

A graph of a crash

Description automatically generated

### 

### Crash Counts by Location (Top 10):

* The plot shows the top 10 locations with the highest crash counts.
* Location 3 has the highest number of crashes, followed by Location 5 and Location 1.
* There is a noticeable variation in crash counts across different locations, with the top 10 locations accounting for a significant portion of the total crashes.

**Do crashes occur more frequently in certain street directions or types of streets?**

A screenshot of a computer

Description automatically generated

A graph of a crash

Description automatically generated

### Crash Counts by Street Direction:

* The plot shows the distribution of crash counts based on street direction.
* Street direction "N" has the highest number of crashes, followed by "W" and "S".
* There is a variation in crash counts across different street directions, with some directions experiencing higher crash frequencies compared to others.

**What is the distribution of crash counts based on street type?**

A graph of crash counts

Description automatically generated

**MODELS**

We used a number of classification techniques, such as SGD Classifier, ADA Boost, Gradient Boost, Random Forest, and Bagging. We also looked into using PCA and kernel approximation to improve the performance of the models. Out of all the methods that were examined, Gradient Boost had the greatest accuracy.   
With an F1 score of 0.40, the Gradient Boost model obtained an accuracy score of 0.89 on the test set. Similar results are obtained for the model with an F1 score of 0.40 and a balanced accuracy of 0.63 on the validation set.

**Gradient Boost Results:**

Best Hyperparameters:

Number of Estimators: 400

Best Performance Metrics (Validation Set:

- F1 Score: 0.40

- Recall: 0.28

- Precision: 0.71

- Balanced Accuracy: 0.63

- ROC AUC: 0.82

Test Set Results:

- Accuracy Score: 0.89

- F1 Score: 0.40

- Recall Score: 0.28

- Precision Score: 0.72

- Balanced Accuracy Score: 0.63

- ROC AUC Score: 0.63

Confusion Matrix:

precision recall f1-score support

0 0.89 0.98 0.94 135913

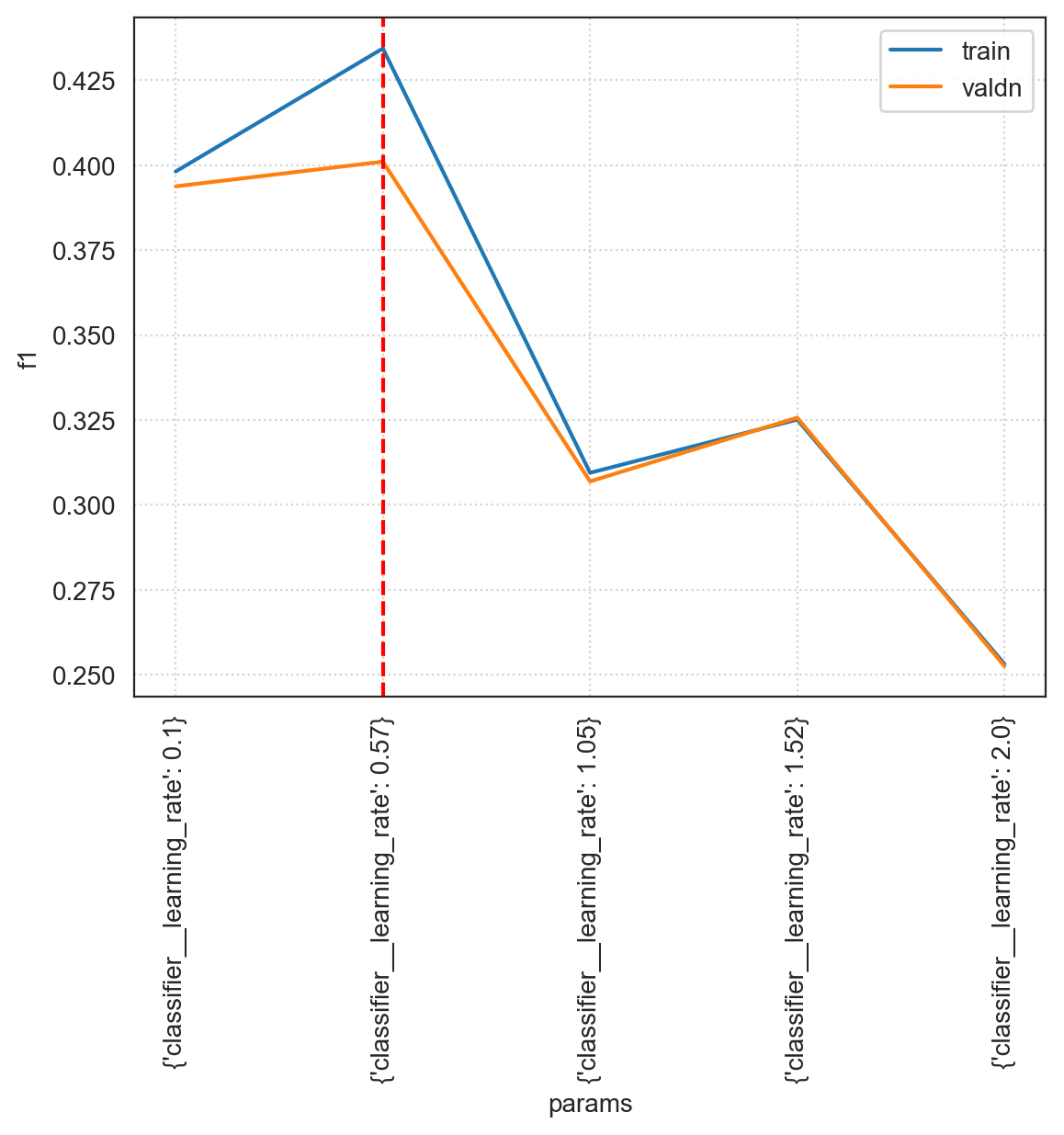
1 0.72 0.28 0.40 21773

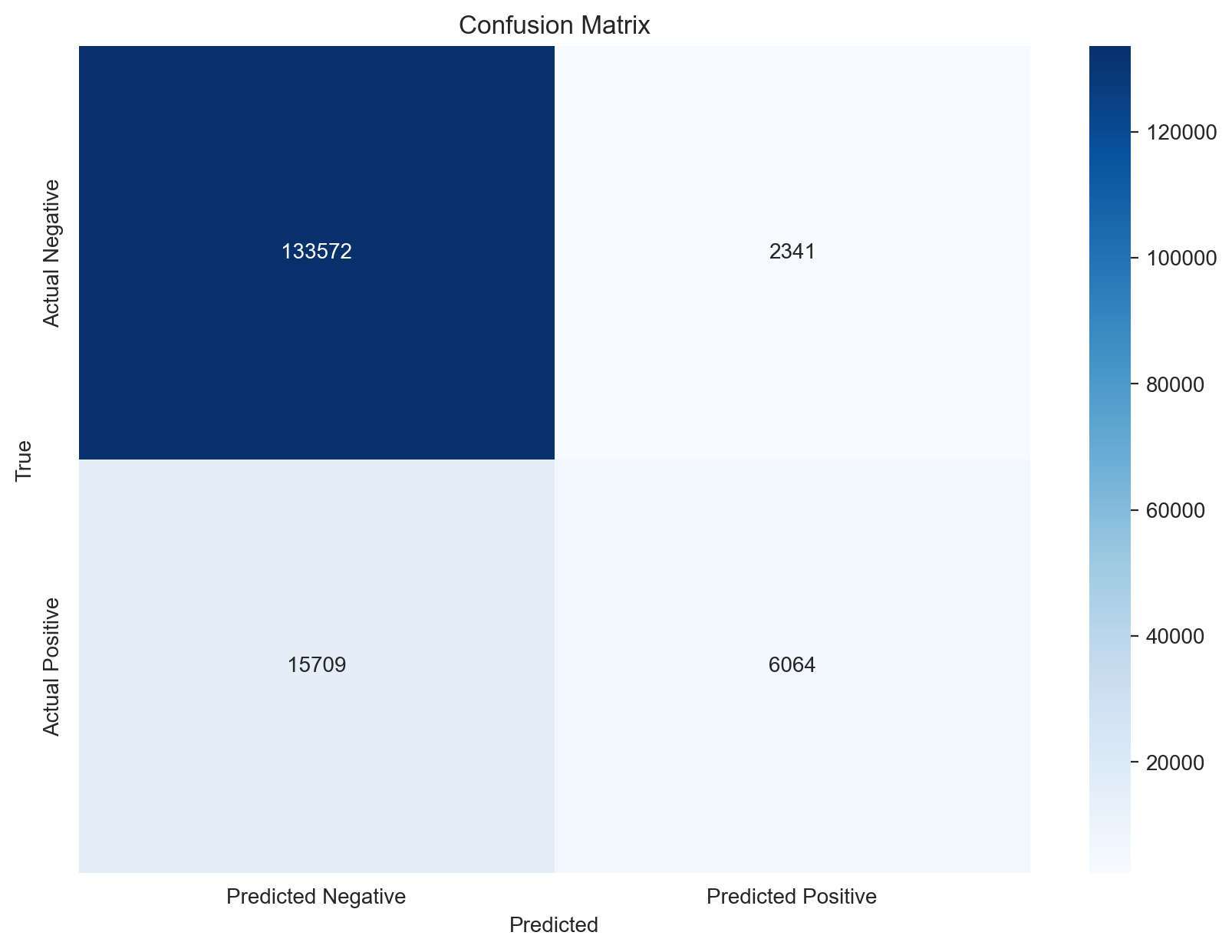
accuracy 0.89 157686

macro avg 0.81 0.63 0.67 157686

weighted avg 0.87 0.89 0.86 157686

{'classifier\_\_learning\_rate': 0.57}





**REFERENCES**

[1]Jkh-Code. (n.d.). JKH-code/predicting-traffic-accident-injuries: This project will predict the number of injuries per traffic accident in the City of Chicago. GitHub. https://github.com/jkh-code/predicting-traffic-accident-injuries

[2] Predicting injuries for Chicago traffic crashes. Julia Silge. (2021, January 4). https://juliasilge.com/blog/chicago-traffic-model/