Murtaza Jawid

Kevin Norris

Yota Sugai

Nathan Young

21 April 2024

STAT 627

Statistical Machine Learning

**STAT 627 Final Project: Chicago Collision Chronicles – A Multifaceted Statistical Analysis**

**Introduction**

In the city of Chicago, where the urban rhythm never sleeps, vehicular accidents occur with far too much frequency. Each collision tells a story, of not only lives changed and families separated, but also of the many factors that intertwine in a split second to affect the outcome. Our project aims to understand the small details of these crashes, seeking patterns amongst the chaos and insights discovered within the data.

We plan to conduct our analysis through two distinct avenues: regression and classification. With regression, we want to uncover the underlying relationships between various factors and the financial toll of accidents. We hope to decipher the monetary aftermath of a crash, drawing connections between factors such as posted speed limits, weather conditions, lighting changes, and crash types. The information we discover can empower policymakers and safety advocates with actionable insights to reduce risks and allocate resources effectively.

At the same time, the path of classification invites us to distinguish between binary outcomes and multiple class differences. Using the tools available with machine learning, we want to predict the fatal aftermath of a crash, discerning nuances in weather conditions, road defects and traffic control devices. These findings could potentially inform emergency response strategies and preemptive measures, saving the lives of those who regularly engage in urban transit. These extend beyond just drivers, but pedestrians, cyclists, and residents of the community at large.

Our planned approach to analyzing this dataset is multifaceted, involving the framing of the problem(s), the data at our disposal, the methods we wield, and the outcomes we envision. Through the proper setup, we encourage clarity instead of complexity when sifting through the nuances of our inquiries. With methods ranging from regression algorithms to classification frameworks, we traverse the statistical territory in front of us. At the end of the day, we hope to contribute to a safer, more informed urban transit landscape through the work contained herein regarding car crash analysis.

**The Data Source and Cleaning the Data**

The primary data source for our analysis is the Crash Data from the City of Chicago, sourced directly from the electronic crash reporting system (E-Crash) maintained by the Chicago Police Department (CPD). As recorded by CDP, this dataset encompasses information pertaining to every traffic crash occurring on city streets within Chicago’s jurisdiction. The dataset withholds any personally identifiable information, maintaining privacy protocols. The dataset spans from 2017 onwards for citywide data, with some police districts having data available from 2015. Approximately half of the crash reports are self-reported by the involved driver or drivers at the police district, with the remainder are recorded be responding police officers at the crash scene. When analyzing the dataset, it’s crucial to acknowledge potential discrepancies in recorded parameters such as street conditions, weather conditions and posted speed limits, as these are usually based on the reporting officer’s assessment at the time of the incident.

Understanding the limitations of the dataset is paramount. Not all traffic crashes within Chicago’s city limits may be captured in this dataset, as crashes on interstate highways, freeway ramps, and local roads along the city boundary, where CPD is not the responding agency, are excluded. Additionally, while Illinois statue dictates certain criteria for reportable crashes, CPD records every reported traffic crash event, regardless of statutory thresholds. Consequently, the dataset may encompass a broader scope of incidents compared to formal crash datasets released by the Illinois Department of Transportation.

**Stakeholders and Ethical Implications**

The analysis of car crash data in Chicago holds implications for a diverse array of stakeholders, each with their own interests and concerns. These include many different groups and individuals, starting with city officials and policymakers. These city officials, including transportation departments and policymakers, are the primary stakeholders in this analysis. They have a vested interest in recognizing the patterns and determinants of car crashes to inform policy decision related to road safety and traffic management. Law enforcement agencies including the CPD also serve as stakeholders, as they are directly involved in the collection and retention of crash data. Relying on accurate and comprehensive data to allocate resources effectively, they focus their efforts on prioritizing enforcement and enhancing public safety.

Third, emergency response and healthcare teams concern themselves with the consequences of car crashes, specifically in terms of injuries and fatalities. Access to accurate and effective crash data can assist with emergency preparedness and trauma care planning. Fourth, the transportation and automotive industry as a whole serve as stakeholders, as they may utilize crash data analysis to affect product design, conduct risk assessments, and improve safety innovations. Finally, community advocate groups represent those directly involved in and affected by car crashes, including pedestrians, cyclists, and residents of various neighborhoods. They may push for safer streets, improved infrastructure, and more vigilant police presence.

Assessing the ethical considerations of the data directly leads to responsible data science practice and implementation. First, one cannot stress enough the importance of safeguarding the privacy of individuals involved in car crashes. This includes anonymizing personally identifiable information and keeping track of data protection regulations to prevent unauthorized access or misuse of sensitive data. Second, handling bias in data collection and interpretation goes a long way to ensure fairness and equity in the results. This requires addressing likely biases occurring in sample selection, data collection methods, and decision-making processes. Third, maintaining transparency in data collection methods, sources and analysis techniques fosters trust and accountability among stakeholders. By holding the data scientists accountable with regards to assumptions and uncertainties in the results, informed decision-making and proper collaboration can occur between all parties. Finally, upholding the ethical considerations, including social responsibilities and the adherence to legal frameworks, directs the proper use of car crash data. By considering the ethical implications outlined above, scientists can navigate the data science concerns and provide analysis beneficial to stakeholders while promoting the safety of drivers, pedestrians, cyclists, and the community at large.