Running Head

Anthony Escalona
Seidenberg School of CSIS
New York, Ny 10038
Email: ae50483p@pace.edu

***Abstract*—Roadway traffic safety is a significant concern for transportation governing agencies as well as ordinary citizens. To provide advice for safe driving, careful analysis of road traffic information is important to identify variables closely related to fatal accidents. In this paper, I apply statistical analysis and data mining algorithms on the NYC Open Data portal dataset as an attempt to address these problems. The relationship between fatal rate and other attributes, including collision manner, weather, surface condition, light condition, and noise complaints, were investigated.**

I. Introduction

It is estimated that around 4,000 New Yorkers are seriously injured in New York and more than 250 people are killed in traffic accidents every year. The automobile is the leading cause of injury-related death for children under 14 years of age and the second leading cause for seniors. On average, every two hours, vehicles severely injure or kill a New Yorker. The cost of these deaths and injuries impacts the city's social and economic growth greatly. New York City should no longer consider traffic crashes as mere "accidents," but as preventable incidents that can be addressed systematically. No degree of fatality is unavoidable or appropriate on the streets of the city. New York City's Vision Zero Action Plan[1] is the foundation to reduce traffic deaths and injuries. City of New York will use every available tool to enhance the safety of our streets. With this action plan, it is making a bold new commitment to improving street safety in every neighborhood and district – with increased enforcement of dangerous moving violations such as speed and failure to yield to pedestrians, new street designs and configurations to improve safety, widespread public access and communications, and a comprehensive legislative agenda to increase penalties.

Data mining is a major step in knowledge discovery. It is the process of extraction of non-trivial, valid and potentially useful information from huge databases. Some of the important data mining techniques are classification, association rule mining, segmentation, and clustering.

Predicting where and when road incidents will occur is complicated. It is possible to analyze traffic injury statistics and identify a correlation between variables based on historical traffic event data. On the other hand, visualization of data from traffic accidents provides detailed insights into how it changes over time. This paper focuses on practical issues related to the project to prevent road accidents. Analysis and visualization of data help observe the occurrence of traffic accidents and take appropriate action to enhance safety.

The paper is as follows. Section II addresses motivation. Section III addresses related work. Section IV presents data analysis and evaluation lessons learned. Section V concludes and outlines future work.

II. Motivation

NYC's Vision Zero Action Plan was launched in 2014, detailing 63 different programs that are implemented by the Mayor's Office and several City Agencies to minimize mortality and serious injury. The Vision Zero Task Force has since launched 143 new initiatives for a total of 206 initiatives (40 new initiatives were implemented in 2015, 22 were added in 2016, 26 in 2017, 27 in 2018 and 28 in 2019). City Agencies separate these programs and continue to make progress on their following measures [2]:

1) New York City Department of Transportation (DOT):
    a) Launch an integrated speed reducer installation program
    b) Install speed cameras at additional school zone locations
    c) Expand and enhance People Priority Streets to improve pedestrian safety and access
2) New York Police Department (NYPD):
    a) Expand outreach and enforcement regarding the safe execution of left and right hand turns by all motorists
    b) Expand NYPD's clear bus routes enforcement action plan
    c) Increase safety within the trade waste and private carting industry through outreach and enforcement
3) New York City Taxi & Limousine Commission (TLC):
    a) Ensure TLC-licensed vehicles with outstanding part recalls are fixed in a timely manner
    b) Reduce use of left turns
    c) Introduce predictive analytics relating to driving behaviors and crashes through CRASHStat and the Fleet Office of Real Time Tracking (FORT)

More vigorous enforcement of dangerous driving behavior by the NYPD and the TLC may help to reduce traffic fatalities and serious injuries. In addition to greater enforcement, efforts are being made to upgrade equipment and technology for speed detection, increase the number of staff on the highway unit, and expand the breadth of information and data captured to preserve crash details better. Finally, the purpose of the data analysis is to derive useful data from the data source and to further use it for visualization by using statistical models.

III. Literature Review

A number of studies have been conducted to determine the factors leading to serious road accidents and to reduce the number and severity of injuries by removing or regulating these factors. As the traffic accident is large and heterogeneous, most scientists adopted data mining methods to carry out their studies.

Chong, Abraham, and Paprzycki [3] have applied artificial neural networks and decision trees to a specific data collection from the National Automotive Sampling Program and General Estimates Systems including traffic incident information from 1995 to 2000. The collection of data was limited only to head-on collisions. The findings revealed that neural networks were outperformed by the decision tree approach. The findings found that seat belt use, highway lighting condition, and driver alcohol use were the most important variables in fatal injuries.

Authors in [4] have applied their work using multivariate logistic regression to determine the independent contribution of driver, crash, and vehicle characteristics to driver' fatality. The result showed that increased use of seatbelts, reduced vehicle speed, and reduced number of and severity of drivers' side-impact might prevent deaths.

The purpose of paper [5] was to employ logistic regression models to develop crash-related injury prediction models. They analyzed traffic crash data in Kentucky during 2001 using logistic regressions. They concluded that the occupant's risk factors for the high level of injury severity were age, gender, and non-use of restraints. In [6], the authors used the same model to quantify the association of driver's age with traffic injury severity. Wisconsin crash data from 2000 to 2004 was used to study 602,694 drivers of a car or truck who were involved in a motor vehicle crash. It was discovered that the oldest drivers, especially those older than 85 drivers 85 years and older, had the highest risks for serious injury or fatality.

IV. Data Analysis and Evaluation

A. Background

The following information should provide enough background for the reader unfamiliar with Logistic Regression to understand the points made in this paper. Readers already familiar with Logistic Regression can skip to the next section. It is a statistical method used to evaluate a dataset in which one or more independent variables calculate an outcome. The result is measured by a dichotomous variable in which there are only two possible outcomes. It uses a formula, like linear regression, as the representation. However, the outcome variable is categorical like two-valued outcomes like true/false, pass/fail, or yes/no. To predict value y, input values x are combined linearly using weights or coefficient values as beta. A key difference from linear regression is that the computation output value is a binary number (0 or 1) instead of a numerical value. Below is an example logistic regression equation:

$$y = e\hat{(}b0 + b1*x) / (1 + e\hat{(}b0 + b1*x))$$

Where y is the predicted output, b0 is the bias or intercept term, and b1 is the coefficient for the single input value x. Each column in the input data has an associated b coefficient, a constant real value, that must be learned from data.

V. Conclusion and Future Work

The conclusion goes here.

Acknowledgment

The authors would like to thank...

References

[1] "Nyc vision zero," 2019, last access November 2019. [Online]. Available: https://www1.nyc.gov/site/visionzero/index.page

[2] "Vision zero scorecard and initiatives," 2019, last access November 2019. [Online]. Available: https://www1.nyc.gov/site/visionzero/initiatives/initiatives.page

[3] M. M. Chong, A. Abraham, and M. Paprzycki, "Traffic accident analysis using decision trees and neural networks," 2004.

[4] M. Bedard, G. H. Guyatt, M. J. Stones, and J. P. Hirdes, "The independent contribution of driver, crash, and vehicle characteristics to driver fatalities," *Accident Analysis and Prevention*, vol. 34, no. 6, p. 717, 2002.

[5] M. Singleton, H. Qin, and J. Luan, "Factors associated with higher levels of injury severity in occupants of motor vehicles that were severely damaged in traffic crashes in kentucky, 2000-2001," *Traffic Injury Prevention*, vol. 5, no. 2, pp. 144–150, 2004.

[6] R. B. Hanrahan, P. M. Layde, S. Zhu, C. E. Guse, and S. W. Hargarten, "The association of driver age with traffic injury severity in wisconsin." *Traffic Injury Prevention*, vol. 10, no. 4, pp. 361 – 367, 2009.