# Review of Fatal Traffic Accidents in DC

**YouJia Chen and Concillia Hleziphi Mpofu**

**School of Public Affairs**

American University

# Table of Contents

# Introduction

## Factors at Play in Fatal Traffic Accidents

In 2017, accidents were noted as the third of DC's leading causes of death. The accident death rate in DC is 61%, which is higher than the U.S. rate of 49.4% (Centers for Disease Control and Prevention ). The CDC further notes that in the year 2020, unintentional injury contributed to more years of potential life lost than any other cause of death in 2020. Years of potential life, described by the CDC as a commonly used measure of premature death, are calculated by multiplying the age-specific and sex-specific alcohol-attributable deaths by the corresponding reduction in years of life potentially remaining for decedents relative to average life expectancies, CDC, (2020). Unintentional injuries include poisoning, motor vehicle traffic, drowning, suffocation, and falls. Motor Vehicle Traffic was the second leading cause of these unintentional injuries.

This paper used data related to traffic accidents and street infrastructures from OpenDC. We conducted Decision Trees, Logistic Regression Analysis, and General Linear Regression as we sought to analyze the factors involved when fatal crashes occur around DC.

## Research Question

What are the factors involved in the occurrences of fatal traffic accidents in DC?

## Data

The data was collected from DC's official data source OpenData. We primarily used Car Crashes and Car Crashes Details from 1975 to the present. After an extensive data exploratory analysis, we focused on the 2000 - 2021 period because it has the most car accident records.

| Data Name | Attributes | Records |
|---|---|---|
| Car Crashes | 58 | 273498 |
| Car Crashes Details | 15 | 721264 |
| Street Lights | 53 | 71,721 |

We filtered out unknown information and NA in each column for better prediction. We went further to combine street light data with car accident data by wards to understand if some infrastructure factors have any kind of influence on car accidents. Due to the challenge of combining car accident data with street light information, we only utilized the variable of the total number of lights in each ward.

| PERSONTYPE | Numbers of Accidents |
|---|---|
| Bicyclist | 4085 |
| ElectricalCar | 28 |
| Passenger | 45676 |
| Pedestrian | 6039 |
| Streetcar | 3 |

### Potential Ethical Issues of the Data

There are several potential ethical issues that arise from using the OpenData DC. These can include invasion of bias, discrimination, and bias in data. The data contains some level of personal information, such as the vehicle license plate numbers, addresses, and contact information of individuals involved in the accidents. If this information is not adequately

protected, it could be accessed and used by unauthorized individuals or organizations. Hence professionals need to be extra cautious and protect the data.

Another ethical issue is the potential for discrimination or bias in the use of the data. For example, if the data is used to identify areas with higher rates of car accidents, this could result in discrimination against individuals or communities living in those areas. Additionally, the use of the data may raise concerns about informed consent. The individuals involved in the car accidents may not have been informed that their information would be used for research purposes and may not have given their consent for their data to be collected and shared.

Overall, it is important to consider and address these ethical issues, we are fully aware we need to use these data in a responsible and respectful manner as well as not perpetuate biases and stereotypes.

## Methods

The goal of this research is to analyze the population and factors that are mostly at risk in car crashes with several classification methods.

### Initial Approach: Logistic Regression

**Initial logistic regression predictors were selected manually and selected by the stepwise function.** The model planned to use Fatal Accident as the response variable (y) and 31 dummy predictor variables (x) that were created by Time Period, Month, Speeding Involved, Age, Ward, and Person Type. During the interpretation process, it is obvious that our data does not fit our model well. It is possible that (1) the data set has multicollinearity issues, (2) there is an extremely unequal distribution of data points, and (3) the wrong approach. It seems hard to deal with multicollinearity and unequal distributions with our data set; therefore, we used the

decision tree method for our prediction model. However, because of high multicollinearity and skewed distribution caused by rare events happening, we noticed that the output from this model was invalid.

**Decision Tree**

***Decision helps us to solve the issues of multidimensional data.*** The model required a match of the sample size for the noninjured person (FATAL = 0, n = 393) with people with fatal injury (FATAL = 1), which initially had 471471 observations for better prediction. The decision tree model was trained on data with four features: PERSON TYPE, SPEEDING_INVOLVED, WARD, and MONTH. The goal of the model was to predict whether a person was injured in an accident (labeled as "Y" or "N").

**Linear Regression**

***Linear Regression on the Numbers of Cases and Street Lights in each Ward in D.C.*** Since the numbers are numeric, linear regression is the most suitable model for this analysis. From the output of logistic regression, we know Night is the time period most affected time period; we compared the street lights with the cases that only happened in the *Evening* and at *Night*.

## Results

### Decision Tress:

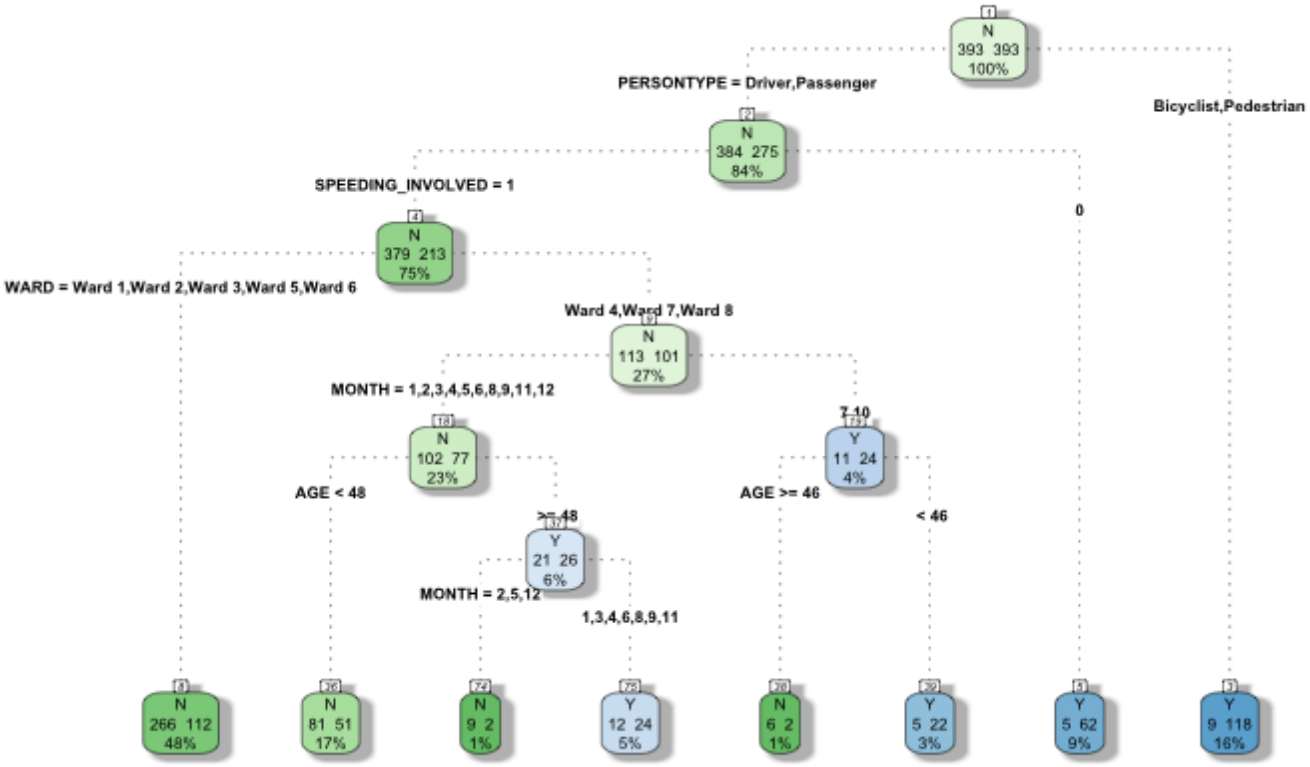



Image [Decision Tree Result]

The output shows the structure of the decision tree, with each node representing a data split based on a certain decision rule. For example, the root node (node 1) indicates that the first data split was based on the PERSON TYPE feature, with the data being split into two groups: one for drivers and passengers and one for bicyclists and pedestrians. It also shows the number of samples (n), the loss or error rate (the percentage of samples that were misclassified), and the predicted class (yval) for each node. For example, at node 2, there are 381 samples, and the loss is 15.8%, with the predicted class being "N".

In addition, the output shows the predicted probabilities for each class at each node. For example, at node 2, the predicted probabilities are 0.585 for "N" and 0.415 for "Y". This indicates that the model is relatively confident in its prediction of "N" at this node.

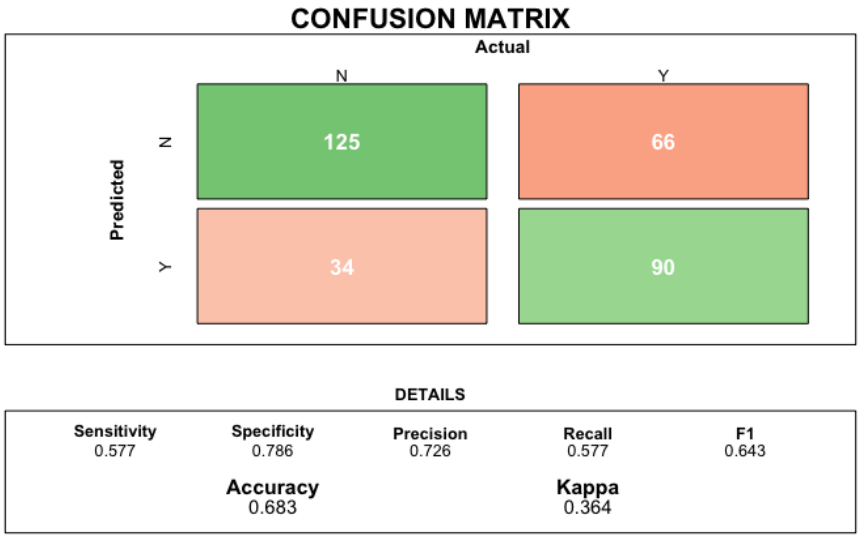



Image [Confustion Matrix and Statistical Summary]

The confusion matrix shows that the model made 125 true negatives, 66 false positives, 34 false negatives, and 90 true positives. This means that the model correctly predicted the "N" class for 125 samples and correctly predicted the "Y" class for 90 samples. However, it also incorrectly predicted the "N" class for 34 samples and incorrectly predicted the "Y" class for 66 samples.

The statistics summary shows that the overall accuracy of the model is 0.683, which means that the model correctly predicted the class for 68.3% of the samples. It also shows that the sensitivity (the proportion of true positives that were correctly predicted) is 0.577, and the specificity (the proportion of true negatives that were correctly predicted) is 0.786.

The full performance of the model is reported in the summary statistics, which shows the confusion matrix and statistics summary of a binary classification model, with the numbers in the matrix indicating the number of correct and incorrect predictions and the statistics providing various metrics for evaluating the performance of the model.

**Linear Regression:**

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 62749.839832 | 48655.652095 | 1.2896722 | 0.2446423 |
| s.lights | 0.886603 | 5.293268 | 0.1674963 | 0.8724824 |

$$\widehat{num.\ case} = 62749.84 + 0.887 \times Num.\ Streelights$$

Our initial expectation was that the more street lights, the less likely to have car accidents in each ward. The intercept coefficient is 62747, which means that the expected value of the number of fatal accidents is 62747 when the number of street lights is zero. The estimated coefficient for the number of street lights is 0.88, this means the expected number of fatal accidents is expected to increase by 0.88 for each unit increase in the number of street lights. This suggests that increasing the number of street lights could potentially reduce the number of fatal accidents. Nonetheless, there is only an 87% probability that the observed value of the predictor's coefficient could have occurred by chance if the true value was actually zero. This indicates there is no strong statistical evidence (p.value > 0.1) to conclude that number of street lights has to do with car accident cases from 5 PM to 6 AM (Evening and Night).

It is important to note that this is only an estimated relationship based on the data used in the regression analysis. Other factors may also be involved in the relationship between the

number of street lights and the number of fatal accidents. For instance, the factors could be the fact that these places are bigger in terms of area with more roads and a high volume of traffic.

## Discussion

During our analysis, we noticed that there's an imbalanced/ skewed distribution. This can also be known as a rare event prediction. By definition, rare events refer to events that occur much less frequently than commonly occurring events (Maalouf and Trafalis, 2011). There is a risk that the fatal crashes may have been treated as noise by the algorithm. Standard classifiers such as logistic regression, Support Vector Machine (SVM), and decision trees are suitable for balanced training sets. When facing imbalanced scenarios, these models often provide suboptimal classification results, i.e., good coverage of the majority examples, whereas the minority examples are distorted (López et al., 2013). This is because there are more nonfatal cases than fatal, which is the reality of data and fatal cases are a rare event in the crashes dataset.

To guide our recommendations for the factor that are significant as causes of fatal accidents, we chose the decision trees method. In this method, we note that due to a skewed distribution in our data, there is an overwhelming classification of NOs from the decision tree for most response variables. From the decision tree, we note that pedestrians are more likely to die in a fatal crash than any other group of road users. This model has more detail that is useful in drafting and recommending policies and bills relating to making DC roads safe for its residents and visitors. From the model, we note that passengers and pedestrians are at a higher risk of dying in the event of a motor vehicle accident, 16%, while drivers have a 9% risk of dying in a motor vehicle crash. The model showed us that people of ages 48 and younger have a risk of 3% dying in a car crash more than any other age group. From the model, we note that the wards

and months do not have much effect in causing fatal crushes. It is important for the district to ensure safety for other road users, such as cyclists and pedestrians.

## Conclusion

In conclusion, we can note that there is a need for the DC government to ramp up road safety education, especially for pedestrians and cyclists, as they are at higher risk of dying in the event of a motor vehicle car crash than any other person type involved in car crashes. Even from the statistics, we noted that passengers are the highest casualties in motor vehicle accidents compared to drivers and other road users. However, there was not enough statistical evidence that other factors, such as age, ward, speeding, month, mar score, etc., had an influence on fatal car crashes in DC. The DC government has already passed a Safer Streets Amendment Act, which is a great stride towards preserving potential life lost in motor vehicle accidents.

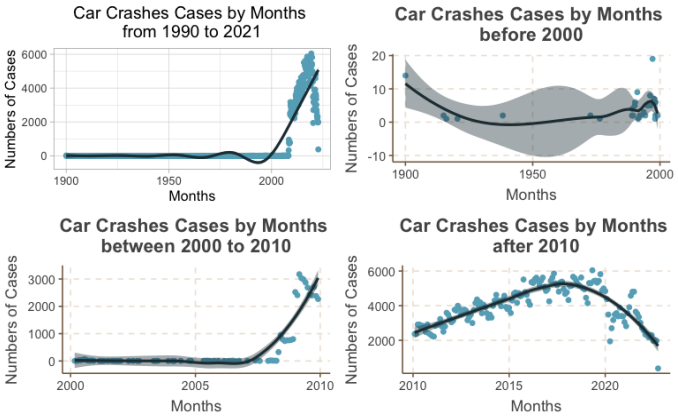
Other models that we ran for this analysis, such as linear regression and logistic regression, were statistically insignificant as such, their algorithms could not be used to predict fatal crashes.

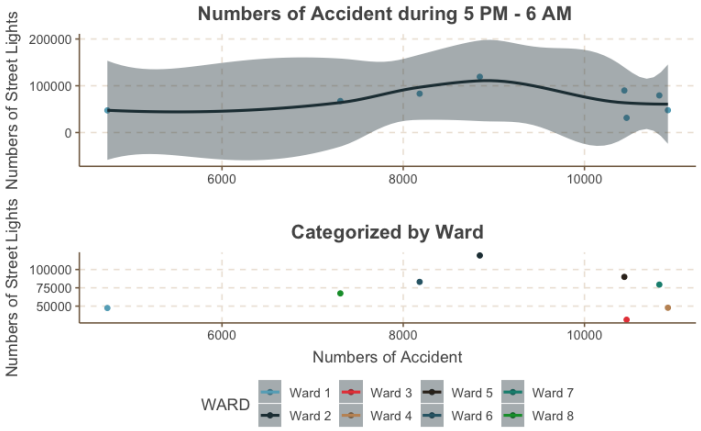
## Recommendations for future study

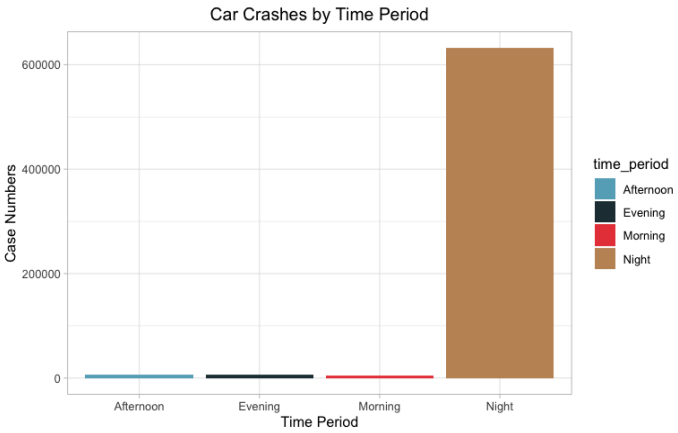
Due to time constraints, we had to remove some high dimensional variables such as type of vehicles and state. However, going forward, we would like to run the models with that data included, especially for motor vehicles, as they have a high chance of contributing to fatalities as some vehicles may have factory faults that can go unnoticed till such data is analyzed. The vehicle can take for a review. Secondly, our initial area of interest was analyzing the ability of traffic lights to reduce traffic fatalities, but we could not join the map of the datasets. Hence we would like to explore the different joins to explore our research area.

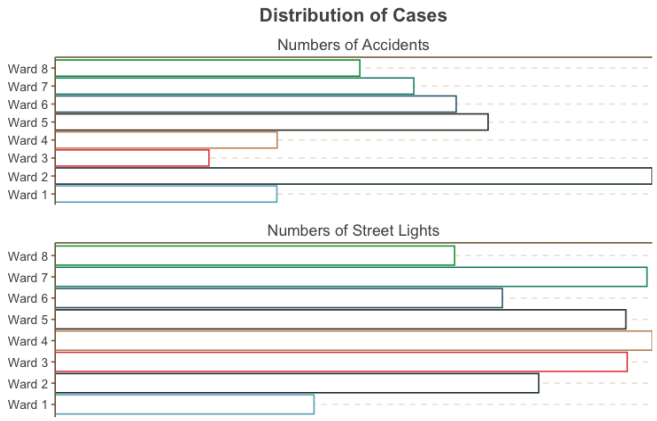
# Index

## Data Exploratory Analysis

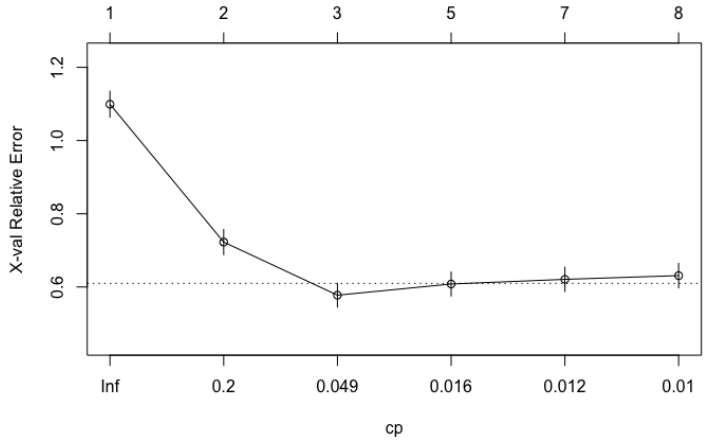


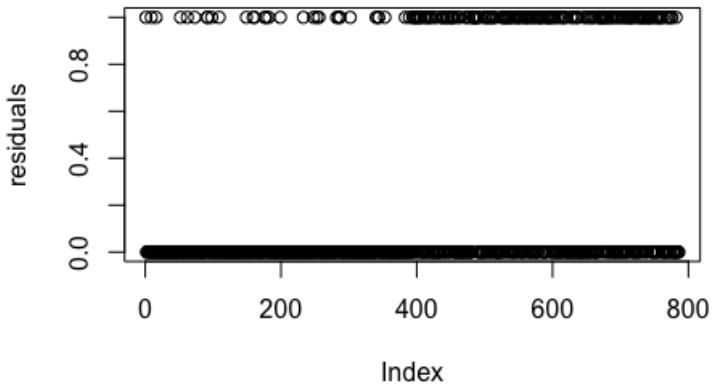




## Decision Tree Diagnosis



**References**

Barba, Carolina Tripp, Miguel Angel Mateos, Pablo Reganas Soto, Ahmad Mohamad Mezher, and Mónica Aguilar Igartua. 2012. "Smart City for VANETs Using Warning Messages, Traffic Statistics and Intelligent Traffic Lights." In 2012 IEEE Intelligent Vehicles Symposium, 902–7. IEEE.

Centers for Disease Control and Prevention. (2018, April 13). *Stats of the District of Columbia*. Centers for Disease Control and Prevention. Retrieved December 10, 2022, from https://www.cdc.gov/nchs/pressroom/states/dc/dc.htm

Centers for Disease Control and Prevention. (2018, April 13). *Stats of the District of Columbia*. Centers for Disease Control and Prevention. Retrieved December 10, 2022, from https://www.cdc.gov/injury/wisqars/LeadingCauses.html

Hayakawa, Hiroshi, Paul S. Fischbeck, and Baruch Fischhoff. 2000. "Traffic Accident Statistics and Risk Perceptions in Japan and the United States." Accident Analysis & Prevention 32 (6): 827–35.

Keep, Matthew, and Tom Rutherford. 2013. "Reported Road Accident Statistics." Commons Library Standard Note, SN/SG/2198.

Kingham, Simon, Clive E Sabel, and Phil Bartie. 2011. "The Impact of the 'School Run'on Road Traffic Accidents: A Spatio-Temporal Analysis." Journal of Transport Geography 19 (4): 705–11.

Marzoug, R, N Lakouari, O Oubram, H Ez-Zahraouy, A Khallouk, M Limón-Mendoza, and JG Vera-Dimas. 2018. "Impact of Traffic Lights on Car Accidents at Intersections." International Journal of Modern Physics C 29 (12): 1850121.

Ma, Y. and He, H. eds., 2013. Imbalanced learning: foundations, algorithms, and applications

Renouf, MA. 1991. “A Car Accident Injury Database: Overview and Analyses of Entrapment and Ejection.”

S. Maldonado, J. López, Imbalanced data classification using second-order cone programming support vector machines

Pattern Recognition, 47 (5) (2014), pp. 2070-2079

V. López, A. Fernández, S. García, V. Palade, F. Herrera

An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics

Information Sciences, 250 (2013), pp. 113-141