Car Crash Prevention
Hotspot Identification and Prevention

Kevin Cha [1] **Jonathan Lam** [1] **Emily Zhou** [1]

Abstract

Emily Zhou

Our research question for this project is how to predict and prevent mobile vehicle crashes in Virginia, especially with crashes being a leading cause of death in the United States. By identifying hotspots for car crashes and the common causes for car crashes, we can work with state officials to better identify how to promote safer driving and reducing fatalities from car crashes.

1. Data

To answer our research question, we have used the VDOT's datatset on crashes within Virginia. https://dashboard.virginiadot.org/pages/safety/crashes.aspx

1.1. Dataset

Jonathan Lam

This dataset consists of motor vehicle crash records in the state of Virginia since 2016. Each record corresponds to a different crash and provides detailed information on roadway conditions, driver factors, environmental conditions, location details, and affected individuals, along with the outcome of the incident. The raw dataset contains 69 variables and over 1 million observations. This dataset was chosen because it provided a comprehensive and reliable overview of all different competing factors involved in a car crash. As such, we can better analyze how they influence the frequency and severity of collisions to build models that can assist government officials with public safety and maintenance decisions.

1.2. Key Variables

Jonathan Lam

In order to build an effective model, we need to identify key variables to serve as features. Because we wish to discern the patterns behind car crashes, variables describing the severity and outcome of the accident, vehicles involved, environmental and weather conditions, road conditions, crash dynamics, traffic control, driver conditions, special populations, and whether it was work zone-related are all included. Variable groups are listed instead of each individual variable due to the vast number of variables. Although driver conditions may not classify as an external contributing factor to a collision, context on whether a given driver was under the influence, distracted, drowsy, unbuckled, or speeding provides a more holistic understanding behind an accident's severity. Remaining variables containing metadata or identifiers are dropped in the data cleaning pipeline.

These variables groups are key because they represent the broad domain of contributing factors in a collision. Variables pertaining to the crash severity and outcome serve as a baseline for predicting crashes, while variables about the vehicles involved (trucks, motorcycles, etc.) may also influence a crash's severity. Environmental and weather conditions have a direct effect on driving performance, and, consequently, the likelihood of an accident; heavy rain or snow limit visibility and influence reaction time. Road conditions play an equally impactful role, where potholes or icy roads could result in serious injuries. Crash dynamics and traffic control provide additional knowledge on an accident's severity. As aforementioned, driver conditions directly relate to incident severity, and special populations and work zone accidents highlight special at-risk groups and conditions.

1.3. Data Wrangling Challenges

Emily Zhou

The first challenge to handle with this dataset was the sheer size of the data from VDOT, as it had data from 2016 and had around 1 million entries. We wanted relevant data, considering our research question of what can be done to prevent crashes in Virginia. Take for example the case of roadway defects such as potholes, since they could have been fixed or new ones have emerged over time and use of the roads. The next step for data cleaning was to determine

[*] Equal contribution [1] Department of Computer Science, University of Virginia, Charlottesville, VA, USA. Correspondence to: Kevin Cha <hpb2gv@virginia.edu>, Jonathan Lam <rgm6yp@virginia.edu>, Emily Zhou <csz6wd@virginia.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

which variables were still useful to our model, since while there were key variables, there were other variables such as OBJECTID which would not support our model so we dropped those columns completely.

The next big steps for cleaning the data were handling both categorical and numerical data, where we had to decide how to handle null values and also turn our categorical data into meaningful numerical data for our model to train on. For null values, it was determined that there were a few variables that only had a few null values so those rows were just dropped but for the following variables: Max Speed Diff, Functional Class, and Facility Type, had more than a few thousand null values and required further analysis. Max Speed Diff was only showing data where the difference between the speed of the car at the incident compared to the speed limit where the crash happened, so it was fair to fill out those null values with 0's. Then for Functional Class, it was a variable that determined where the crash occurred and since 10% of the values were NaN's, imputing them as ”Unknown” sufficed. With the last variable with a significant number of null values being Facility Type, it also explained the format of the road where the crash occurred, so the remaining null values were imputed as ”Unknown” as well.

With our categorical data, we had some issues discerning what the variables mean since there was no specific data dictionary for this dataset but after some thorough search the most confusing variables, K_people, A_people, B_people, C_people were determined to be on a scale referred to as KABCO. This scale indicated the damage to human lives, where K meant fatality, A meant serious injury, B meant minor injury, C meant possible injury, and O meant property damage only. The best way to handle this in a meaningful manner was to assign weight mappings to the variables, since K does have a higher severity due to lives being lost.