DS504 Fall 2020  
Final Project Proposal

Jeffrey Baglioni

# Overview – Driver safety

## Project Background and Description

Approximately 40000 Americans die annually in car accidents and it is one of the leading causes of death in the country[[1]](#footnote-1). The purpose of the project is to provide a model that can help identify the combination of factors most likely to result in a extremely sever or fatal car wreck, and help save lives.

To accomplish this, machine learning is used to analyze car accident data, compare the most influential factors and provide a prediction score. Features include a selected list of pre and post accident variables that are most likely to affect the outcome of an accident. They include environmental factors, vehicle characteristics, as well as attributes of the people involved in the crash. For example, environmental factors may include the weather, time of year, location of the crash, road and lighting conditions. Characteristics about the vehicle involved are things such as vehicle type, make/model, major component failures, type of damage etc. Finally, characteristics of the people involved include their age/experience, sex, whether they were distracted, impaired, or speeding, amongst other factors.

Although somewhat macabre, the predictions may help to save lives. For example, car manufacturers may use the outcomes to help focus on improving automobile safety features by knowing that certain types of vehicles incur fatal accidents when damaged in the same places, or from the same types of equipment failures. First responders can use this analysis to be better prepared for response calls, and public safety campaigns can target the types of behaviors or combinations of behaviors likely to cause an accident, and specific to those drivers likely to be involved.

More directly, driving navigation apps like Waiz could incorporate this type of algorithm to warn drivers in real time when certain conditions are aligning that present an increased and immediate safety risk. For example, a driver in a certain make of car, driving along a certain type of road in certain weather may be warned to slow down or change course because of elevated threat risk. By providing a little bit of personal information, such as their age or how they are feeling, the accuracy could be improved even more. Or better still, it could be combined with phone or car sensors that could help identify driver state or behavior without even requiring input. For example, if the car sensed that repeated un-signaled lane changes were occurring, or the car was moving in erratic directions, it could indicate driver impairment, prompting the app to make the driver aware of a potentially serious conditions.

As autonomous vehicle technology improves, even more proactive measures could be imposed to prevent a likely accident form occurring. For example, if the car were to detect a combination of factors deemed too dangerous, it could be programmed to automatically notify police, limit its speed maximum speed, or take other evasive/defensive measures like put on its hazard lights. As the prediction score increased, even more substantive actions could be taken, for instance the car could take control away from the driver, pull itself over to the side of the road and/or turn itself off. Many newer automobiles already have rudimentary versions of these types of actions, as many cars will automatically break if an object is detected in front of them. This project aims to provide a more sophisticated, data-driven version of this.

## Project Scope

The scope will be to find, collect, prepare, develop and build a model that can effectively identify car accident risk based on the input of key environmental, vehicle and person factors. The model will be developed using the best, freely available tools and technologies available to gain its insights previously not recognized. While a production-ready application will not be developed, a basic interface for testing and demonstrating model output will be provided along with a final presentation.

## Specific Exclusions from the Scope

The project will be limited to development of the model only. Development of a webapp or other software application will not be conducted as part of this work.

*Also, a note on the prediction:* I realize there is a slight logical issue with what we are posing in this project: in order to predict a severe accident, there first must be an accident to begin with. I am taking it for granted that an accident has occurred, and the algorithm will not try to predict that. What I want to explore in this project is:

Is there a difference in the factors contributing to a very severe or fatal accident versus those that are not?

Can these differences be modeled?

If so, can they be articulated and/or recognized independently and used to improve driving safety?

## High-Level Requirements

The algorithm is expected to meet the following requirements:

* Utilize publicly available accident data provided by the US government. The data contains well-documented, standard fields, is updated regularly, and has several years of historical records to train the model on.
* Utilize open-source technologies and tools.
* Provide demonstrable accuracy scores
* Provide results which are repeatable and auditable

## Benefits and applications

The benefit to this algorithm is that it might save lives. Driving is one of the most dangerous activities people participate in on a daily basis, killing approximately 40000 people per year[[2]](#footnote-2). An effective model that identifies the factors causing severe car accidents can help people live longer, healthier lives. A model that could do this well can have many applications.

* Automobile insurance providers may use it to calculate more accurate and equitable premium rates that reward drivers who are truly less dangerous.
* First responders may use predication models to better prepare for calls, calculate staffing needs, or train paramedics.
* Public safety campaigns can identify a combination of risky or preventable behaviors, and these can be communicated to the public for greater awareness and to promote behavior changes. This has worked in the past, for example seatbelt use was not as prevalent until the 1980s until ad campaigns focused attention on the benefits (remember “You could learn a lot from a dummy”?)[[3]](#footnote-3).
* Car manufacturers may use the outcomes to help focus on improving automobile safety features based on failures that have occurred, or specific
* Driving/navigation applications could incorporate the algorithm to warn drivers in real time when certain conditions are aligning that present an increased and immediate safety risk.
* Sensors that measure driver or car behavior (such as high speed, erratic movements) could be used to provide warnings to the driver, or passengers.
* As self-driving vehicles become more prevalent, conditions posing risk could automatically detected and remediated. For example, if the car detects a drowsy driver, it may automatically be enabled to pull itself over, or take over control of the vehicle.

## Data

Accident data published by the United States National Highway Traffic Safety Administration (NHTSA) will be used. The NHTSA publishes several datasets related to traffic safety, going back to the 1970’s. The dataset that will be used for this project is the “Crash Report Sampling System”. Per the NHTSA website, this dataset is “a sample of police-reported crashes involving all types of motor vehicles, pedestrians, and cyclists, ranging from property-damage-only crashes to those that result in fatalities. CRSS is used to estimate the overall crash picture, identify highway safety problem areas, measure trends, drive consumer information initiatives, and form the basis for cost and benefit analyses of highway safety initiatives and regulations”[[4]](#footnote-4).

The dataset is published in its entirety online and made available via FTP download[[5]](#footnote-5). In its current iteration, the CRSS dataset has been in place since 2016, and up to 2018. Prior to that it was part of a different dataset (the General Estimates System (GES)). In the interest of time and wish to avoid widely divergent data formatting, only the three years of the CRSS dataset will be used for this project (2016-2018). Future iterations may use the older datasets, depending on the availability of all features)[[6]](#footnote-6).

Per the NHTSA website, CRSS obtains its data from a nationally representative probability sample selected from the estimated 5 to 6 million police-reported crashes that occur annually. These crashes include those that result in a fatality or injury and those involving property damage. These crash reports are randomly chosen from 60 selected areas across the United States that reflect the geography, population, miles driven, and crashes in the United States.

In the three years that will be used as part of this project, there are approximately 150,000 accidents sampled, that include over 120 data elements are coded into 22 separate csv formatted files per year. The data contains accidents that are both fatal and non-fatal, and actually scales each accident on a severity rating form 1-4, with 4 being fatal. It is this column that will be leveraged to train and test and model for classifying accident severity.

## Tools and Pre-work

Standard Python libraries will be used to import, clean and analyze the data, and build the model. A Jupyter Notebook will be the primary application used for coding and testing. Libraries for cleaning and analyzing the data are well known, and include numpy, pandas, and scikit learn. Several machine learning algorithms will be tested as part of the project. The use of ensemble methods is anticipated. All pre-work, data cleaning and other transformation steps will be documented in the final Jupyter notebook, available for review upon request.

The use of Spark was investigated, but not anticipated for use at this time due to its current lack of support for ensemble methods. Should data processing be overly slow or cumbersome on a standalone machine, services offered from providers such as AWS or Google will be leveraged.

## Deliverables

* A pickled model that this portable and applied to future applications.
* A presentation that outlines the data analysis, model function and outcome.

## High-Level Timeline/Schedule

The project timelines will be governed by overall class due dates. The major milestones and intermediate goals are listed below.

* Final Project Proposal due by October 1
  + High level project goals expected to be in place
* Final Project Status Report #1 due by October 8
  + By this time all data should be clean and imported in Jupyter Notebook. Beginning exploration of machine learning models
  + Any issues and errors should be identified
* Final Project Status Report #2 due by October 22
  + By this time several models should be tested
  + Identified issues and errors should be nearly remediated
* Final Project Deliverable (Strategy Consulting Presentation Deck) due by October 29
  + Complete solution developed, tested and error free
  + Presentation materials with outlining actions and final results

1. <https://www.iii.org/fact-statistic/facts-statistics-mortality-risk> [↑](#footnote-ref-1)
2. <https://www.iii.org/fact-statistic/facts-statistics-mortality-risk> [↑](#footnote-ref-2)
3. <https://aef.com/classroom-resources/social-responsibility/ad-council-campaigns-made-difference/seat-belt-education/> [↑](#footnote-ref-3)
4. <https://www.nhtsa.gov/crash-data-systems/crash-report-sampling-system> [↑](#footnote-ref-4)
5. <https://www.nhtsa.gov/content/nhtsa-ftp> [↑](#footnote-ref-5)
6. The datasets have changed over time, and things have more around and been added. If time permits, including older data will done if all needed features are present and mappable. [↑](#footnote-ref-6)