DATS 6103 Final Project Proposal - Group 8

Jichong Wu
Ethan Litman
Jia-Ern Pai

What problem did you select and why did you select it?

On a global scale, approximately 1.35 million people die annually as a result of motor vehicle collisions[1]. Road traffic injuries are the leading cause of death among people ages 5-29[1]. In the United States there were more than 33,244 fatal motor vehicle collisions[2]. Anyone who has been involved in a motor vehicle collision understands that collisions seemingly occur at random, without warning and take an emotional, physical and financial toll on the parties involved.

The National Highway Traffic Safety Administration (NHTSA) found that 18.9 percent of fatal crashes involved rollover events in 2014[3]. Rollover events happened in various types of crashes, such as single-vehicle crashes and multi-vehicle crashes. This project will only focus on the rollover events in fatal single-vehicle crashes.

The purpose of this project is to study how to reduce the likelihood of rollover events in fatal single-vehicle crashes.

What database/dataset will you use? Does it need to be cleaned?

This project will use the Fatality Analysis Reporting System (FARS) provided by NHTSA.

FARS is a census of fatal traffic crashes that includes the 50 States, the District of Columbia and Puerto Rico since 1975. The crashes in FARS must result in the death of at least one person within 30 days of the crash. NHTSA has a cooperative agreement with an agency in each State government to provide information in a standard format on fatal crashes occurring in the State. The data observations in FARS came from police crash reports in the States, death certificates, State coroners and medical examiners, State driver and the vehicle registration records and emergency medical service records. NHTSA's FARS datasets can be downloaded at https://www.nhtsa.gov/content/nhtsa-ftp/251.

FARS database includes the following datasets each year:

[1] World Health Organization (WHO). Global Status Report on Road Safety 2018. December 2018. [cited 2020 October 28]. Available from URL: https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/external icon
[2] Federal Highway Administration. 2020. Highway statistics, 2019. Washington, DC: US Department of Transportation
[3] Traffic Safety Facts 2014: A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System. (Report No. DOT HS 812 261). Washington, DC: National Highway Traffic Safety Administration.

- Accident dataset: The variables in the accident dataset describe the circumstances of a fatal crash. Variables, such as the weather condition, roadway condition and local speed limit are belonged to the accident dataset.

- Person dataset: The variables in the person dataset describe the characteristics of occupants that experienced a fatal crash. Variables, such as the number of occupants in a crash event, driver's gender and driver's age are belonged to the person dataset.

- Vehicle dataset: The variables in the vehicle dataset describe the characteristics of vehicles in a fatal crash. Variables, such as the model year and first impact direction are belonged to the vehicle dataset.

Taking FARS year 2018 dataset as an example, it includes the following sub-datasets:

- "accident" has 33,919 observations and 91 features.
- "Vehicle" has 522,86 observations and 197 features.
- "Person" has 84,344 observations and 118 features.

This project will use FARS between 2014 and 2018. FARS 2014-2018 will need to be sliced, merged and cleaned, since missing values exist in FARS.

What data mining algorithm will you use? Will it be a standard form, or will you have to customize it?

This project will use the standard K-nearest-neighbor model and logistic regression model, since this is a supervised study, the variables are categorical.

What packages will you use to implement the network? Why?

1. Pandas and NumPy: Data manipulation and data cleaning

2. SciPy and Statsmodels: Statistical hypothesis test and statistical model building

3. Matplotlib, Seaborn and Plotly: Visualize the analysis results

4. PyQt: GUI

What reference materials will you use to obtain sufficient background on applying the chosen network to the specific problem that you selected?

1. Trends and RolloverReduction Effectiveness of Static Stability Factor in Passenger Vehicles[4]

[4] Jia-Ern Pai. DOT HS 812 444, August 2017. Washington, DC: National Highway Traffic Safety Administration. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812444

2. The Effect of ESC on Passenger Vehicle Rollover Fatality Trends[5]

3. Characteristics of Fatal Rollover Crashes[6]

How will you judge the performance of your results? What metrics will you use?

This project will use the value of ROC curve to judge the performance of K-nearest-neighbor model and logistic regression model.

Provide a rough schedule for completing the project.

Week	Work
3/20 - 4/3	Proposal – problem specification and understanding
4/4 – 4/10	Data cleaning & data mining
4/11 – 4/17	Data visualization & variable selection Statistical test for the variable significance GUI
4/18 – 4/24	Model building GUI
4/25 – 5/1	Presentation slides Presentation recording Presentation practice

[5] Bob Sivinski. DOT HS 812 031, June 2014. Washington, DC: National Highway Traffic Safety Administration. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812031

[6] William Deutermann. DOT HS 809 438, April 2002. Washington, DC: National Highway Traffic Safety Administration.
https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/809438