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

The National Highway Traffic Safety Administration (NHTSA) found that 18.9 percent of fatal crashes involved rollover events in 2014[[1]](#footnote-1). 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 within 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>.

This project will use FARS between 2014 and 2018. FARS 2014-2018 will need to be 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

**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 (<https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812444>)

2. The Effect of ESC on Passenger Vehicle Rollover Fatality Trends (<https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812031>)

3. Characteristics of Fatal Rollover Crashes (<https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/809438>)

**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 |
| 3/26 - 4/3 | Data cleaning & data manipulation |
| 4/4 – 4/10 | Data visualization & variable selection |
| 4/11 – 4/17 | Statistical test for the variable significance |
| 4/18 – 4/24 | Model building |
| 4/25 – 5/1 | Prepare presentation |

1. 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. [↑](#footnote-ref-1)