Project Proposal

On a global scale, approximately 1.35 million people die annually as a result of motor vehicle collisions. Road traffic injuries are the leading cause of death among people ages 5-29. In the United States there were more than 33,244 fatal motor vehicle collisions. Anyone who has been involved in a motor vehicle collision understands that collisions seemingly occur at random, without warning, and are often physically, emotionally, and financially scarring.

While there are many well known risk factors that increase the chance of an individual being involved in a fatal collision, such as lack of restraint use and substance use, there are currently no well defined predictive algorithms for high-risk scenarios that are likely to result in fatal collision. For example, using vehicle type, roadway condition, weather, and drive demographic information, an algorithm could be used to provide the driver with a real-time calculation of their risk of being involved in a fatal collision.

We plan to utilize the National Highway Traffic Safety Administration (NHTSA) publicly available dataset on fatal crashes in the United States. Annual data organized by state is publically available. There are over 50 variables that are contained in the dataset and we will create a customized dataset of variables of interest.

We will likely utilize an expectation-maximization data mining algorithm to predict the probability of a fatal collision given weather conditions, road condition, time of day, type of vehicle, and demographic information of driver.

We plan to randomly select 20% of our data to test the performance of our algorithm. We propose that a model capable of correctly predicting fatal collision in 70% of cases is clinically relevant.

We plan to spend 1-2 weeks cleaning the data and collecting the dataset and 2-3 weeks implementing the data mining algorithm and testing the dataset.

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

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

##################

• What problem did you select and why did you select it?
• What database/dataset will you use? Does it need to be cleaned?
• What data mining algorithm will you use? Will it be a standard form, or will you have to customize it?
• What packages will you use to implement the network? Why?
• What reference materials will you use to obtain sufficient background on applying the chosen network to the specific problem that you selected?
• How will you judge the performance of your results? What metrics will you use?
• Provide a rough schedule for completing the project.