**Abstract**

Road crashes each year result in thousands of lives lost and injured victims, and billions of dollars in property damage in the USA. Accurate data are required to support the development, implementation, and assessment of highway safety programs aimed at reducing this toll. The analysis of highway-crash data has long been used as a basis for determining vehicle and highway designs. It also has great influences on directing and implementing a wide variety of regulatory policies aimed at improving safety. With the improvement in statistical methodologies, researchers are able to extract more information from crash databases to guide a wide array of safety design and policy improvements. In this project, Crash Report Sampling System (CRSS) of 2018 data were used which contains all types of police-reported crashes, ranging from property-damage-only crashes to those that result in fatalities. ACCIDENT Data File and PERSON Data File in CRSS was mainly analyzed to find the correlation between injury information and several representative variables like age, sex, alcohol involvement, manner of collision, ejection status, day of the week, light condition and so on. To start with, basis statistic information of crashes was obtained by ACCIDENT Data File. Then we built models to predict the maximum injury severity in an accident on the basis of ACCIDENT Data File, which can be a criterion of how dangerous an accident is. Finally, models predicting whether a person will be injured during the accident are created based on the information of PERSON Data File.

**Introduction and Background**

More than 1.2 million people died every year in highway-related crashes and almost 50 million people are injured all over the world. It is reported that the highway-related crash is the 5th leading causes of the death worldwide. Apart from that, highway-related crashes result in many billions of dollars in property damage. To provide vital information on motor vehicle traffic crashes, the National Highway Traffic Safety Administration (NHTSA) annually publishes nationally representative estimates of police-reported motor vehicle traffic crashes and their characteristics. From 1988 to 2015 NHTSA created national estimates using data from the National Automotive Sampling System General Estimates System (NASS GES), which sampled police crash reports from police jurisdictions across the United States. In 2016 NHTSA replaced NASS GES with the Crash Report Sampling System (CRSS), which 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. The CRSS obtains its data from a nationally representative probability sample selected from the more than six million police-reported crashes which occur annually. Although various sources suggest that there are many more crashes that are not reported to the police, the majority of these unreported crashes involve only minor property damage and no significant personal injury. By restricting attention to police-reported crashes, the CRSS concentrates on those crashes of greatest concern to the highway safety community and the general public. The datasets we are going to use is from CRSS 2018.

Several popular injury severity model structures have been studied in the extant literature. Sequential binary logit models1, ordered-response probit models2 and multinomial or nested logit models3 are typical examples. The sequential binary logit and ordered-response probit models represent the ordinality in the discrete categories of the injury severity. Sequential binary logit models assume that the factors determining the level of the severity change according the level of the severity itself, while ordered-response probit models assume that the same factors are correlated with all levels of injury severity. In multinomial and nested logit models, ordinality is not theoretically implemented, thus information relating to ordering of severities is not inherently captured in those structures. Compared with the previous two models, multinomial and nested logit models are structurally flexible in the sense that independent variables are not forced to be the same across all severities.

In this project, crash statistic information was analyzed, such as which region has the highest number of crashes, which hour of the day has the highest number of injuries and so on. Then we built models to predict the maximum injury severity in an accident and whether a person will be injured during the accident by logistic regression, neural network and random forest classifier method.

**Predictive Problem and Approach**

The maximum injury severity is of great importance to transportation analysis but difficult to predict. We select some representative variables which have strong correlation with injury severity, and then logistic regression, neural network and random forest classifier are applied to generate models to predict injury severity and whether the person involved in accidents is injured. To improve the precision of our models, pipeline is used to determine the best parameters for the random forest classifier method and ROC curve to evaluate our model.

The most difficult part is that maximum injury severity has five categories, including no apparent injury, possible injury, suspected minor injury, suspected serious injury and fatal. It means that models predicting multiclass variables should be built which is different from the binary variables models we learned in class.

**Methodology**

**Data Collection**

In CRSS 2018, totally 48443 representative crashes are selected from over six million police-reported crashes in 2018, involving 86105 vehicles and 120230 people. The ACCIDENT data file contains crash data and PERSON data file includes motorist and non-motorist data. The following two table show the representative variables considered in the maximum injury severity models from ACCIDENT file and if the person involved in accidents is injured or not from PERSON file respectively.

Table 1 representative variables of ACCIDENT file

|  |  |  |
| --- | --- | --- |
| 1. Maximum severity injury | 6. Alcohol involvement | 11. Number of Persons in Motor Vehicles In-Transport |
| 1. Hour | 7. Manner of collision | 12. Number of Persons Not in Motor Vehicles |
| 1. Day of the week | 8. Location of crash (whether in junctions or interchange areas) | 13. Number of vehicles in the crash |
| 1. Light condition | 9. Location of crashes (type of junctions or interchange areas) |  |
| 1. Weather | 10. First harmful event |

Table 2 representative variables of PERSON file

|  |  |  |
| --- | --- | --- |
| 1. Maximum severity injury | 6. Alcohol involvement | 11. Sex |
| 1. Hour | 7. Manner of collision | 12. Age |
| 1. Day of the week | 8. Location of crash (whether in junctions or interchange areas) | 13. Ejection status |
| 1. Light condition | 9. Location of crashes (type of junctions or interchange areas) | 14. Seating position |
| 1. Weather | 10. First harmful event |  |

Weight should be involved when analyzing frequency to ensure unbiased and robust estimate. In the pre-processing part, we transform the hour of the day into two categories, where daytime is from 6am to 8 pm and nighttime is from 8pm to 6am, and day of the week is also divided into the weekday and weekend. The rest of categorical variables are all transformed into dummy binary variables. For the first harmful event related to the maximum injury severity, there are four primary categories and many secondary categories, so we choose fifteen categorical first harmful events whose frequencies are larger than 0.007, including motor vehicle in-transport, parked motor vehicle, live animal and so on. There are totally 88 variables

**Data Analysis Theory**

Logistic regression is a classification algorithm which assigns observations to a discrete set of variables. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to more discrete variables.

Random forest classifier is a meta estimator that fits a number of decision tree classifiers on a variety of subsets of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The size of subsets is always the same as the original input sample size.

~~~~~~~~continued~~~~~~~~

**Reference**

1. Saccomanno, F. F., Nassar, S. A., & Shortreed, J. H. (1996). Reliability of statistical road accident injury severity models. *Transportation Research Record*, *1542*(1), 14-23.
2. O'donnell, C. J., & Connor, D. H. (1996). Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. *Accident Analysis & Prevention*, *28*(6), 739-753.
3. Shankar, V., & Mannering, F. (1996). An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. *Journal of safety research*, *27*(3), 183-194.