**Predicting Accident Severity: Literature Review and Exploratory Data Analysis**

With advances in technology that have led to modern assisted and autonomous driving technologies, there has been a great deal of research interest in the reduction of traffic accidents. The “Race to Zero”, as in zero accidents, zero emissions, and zero congestion was a previously unthinkable goal that is now the benchmark that every player in the mobility business is striving for. Reducing the number of severe collisions is a huge component of this, and that is the topic of the relevant literature that I have reviewed. Both Jianfeng, et al and Garrido, et al carried out studies attempting to predict accident severity based on a conglomeration of vehicle, driver, environment, and accident information. My goal fundamentally differs from theirs in that I am seeking to predict accident severity based solely on information available immediately upon accident reporting in concordance with the use case of a GPS program being able to immediately notify motorists. Also, severity in my data is defined as the delay an accident causes, rather than vehicle damage or injury sustained. Nevertheless, previous research proved very insightful towards my goals.

Jianfeng, et al and Garrido, et al both wrote very favorably about the effectiveness of ordinal logistic regression and tree models on this data, which is the path I will pursue given that the features I plan to use are very similar to theirs. Also, Jianfeng, et al’s methodology of factor analysis was very insightful, and it was reflected in a positive increase in model performance. In addition, even though my research does not focus on this, given the predicted emergence of autonomous cars, it was prudent to find what research existed on AV collisions. While data here is extremely limited, there was a study performed in 2019 on 114 AV accident records from California that can serve as a starting point for research into the topic. In general, it found that the majority of severe accidents occurred during autonomous driving mode and were the fault of the AV system, while nearly every minor accident was because of human error.

In my exploratory data analysis, I found accident severity to be highly bimodal, with most observations classified as either a 2 or 3 out of 4. The roadway type appears to have the strongest influence on this, and I may have to do more creative feature engineering to extract more location information given that Garrido, et al found more severe accidents occurred in rural areas. 2 interesting outliers were also found, West Virginia and Arkansas both had statistically significantly higher percentages of severe accidents, with a satisfactorily large sample size.

References

* Jianfeng, X., Hongyu, G., Jian, T., Liu, L., & Haizhu, L. (2019). “A classification and recognition model for the severity of road traffic accidents”. *Advances in Mechanical Engineering*, 11(5), 1687814019851893
* Garrido, R., Bastos, A., Almeida, A. D., & Elvas, J. P. (2014). “Prediction of Road Accident Severity Using the Ordered Probit Model”. *Transportation Research Procedia*, 3, 214–223. doi: 10.1016/j.trpro.2014.10.107
* Wang, S., & Li, Z. (2019). “Exploring the mechanism of crashes with automated vehicles using statistical modeling approaches”. *Plos One*, 14(3). doi: 10.1371/journal.pone.0214550