BIOS 735 Project Proposal

Group 3 - CALCCY
Marissa Ashner, Yen Chang, Marco Chen, Weifang Liu, Yi Tang Chen, Joyce Yan
3/25/2020

Dataset: US Car Accidents

- **References for Dataset**

https://www.kaggle.com/sobhanmoosavi/us-accidents
Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.
Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

- **Description of Data**

“This is a countrywide car accident dataset, which covers 49 states of the United States. The accident data are collected from February 2016 to December 2019, using several data providers, including two APIs that provide streaming traffic incident data. Currently, there are about 3.0 million accident records in this dataset.” For the purpose of this project, we will focus on car accidents in North Carolina. There are 49 columns in the original dataset and ~ 140,000 accidents recorded in the North Carolina specific dataset.

- **Outcomes of interest:**

Severity of car accidents (1-4, increasing impact on traffic) in North Carolina

- **Short list of variables that could help classification**:

Weather condition, visibility, time of the accident, traffic, location/zip code, landmarks/road features (stop signs, roundabouts,...etc.)

- **What question(s) do we want to answer?**

1. What variables have the most influence on the severity of car accidents in North Carolina?
2. Can we build a model to predict the severity of car accidents in North Carolina?

Model Building

- **How do we plan on answering this question? What model(s) can we use?**

Module 2: Fit a proportional odds model and use BFGS to optimize and find the parameter estimates and associated standard errors, after picking a subset of the attributes we want to consider. Based on the parameter estimates and SEs, decide what variables to keep in order to pick the most parsimonious model.

Module 3: Alternatively, use machine learning classification methods (i.e. Random Forest) and apply them to our ordinal response variable to select important attributes and build a classification model to predict car accident severity.