**Driving Towards a Safer Future- Part 3: Model Construction**

Mikhail Hershfeld

Since the EDA/ Data exploration stages of the project, quite a lot has been added. All major trends have been identified in the data, and major steps have been made towards the main goal of predictive analysis. However, there were some interesting finds from data exploration that also merited further analysis. My intent was to further explore geographic anomalies if any were found, and West Virginia and Arkansas were found to have a highly disproportionate rate of very severe accidents. While data was available on traffic volumes in these states, it was very difficult to find any hard data specifically related to the roadways within these states that were prone to these accidents. Given this, any conclusions are difficult to firmly assert given the lack of quantitative evidence available. Strangely as well was West Virginia, which contained none of the possible explanatory factors Arkansas had, and therefore its severe accident rate is still unexplained.

With what information was available, I dug into Arkansas. Almost all of the level 4 accidents were found to have occurred on the two major interstate highways that run through the state, I-40 and I-30. These highways have a very high volume of traffic, with I-40 receiving a daily average of 3,698,290 cars, and I-30 receiving 3,859,140 cars daily on average (2017). Note that these figures are from 2017, the most recent year data was available. Following these are I-49 with 3,011,310 daily cars and Route 71 with 2,602,570 cars. I-40 and I-30 are also among the busiest freight routes in the U.S., with rough estimates of at least 50,000 trucks on these routes per day (2017). However, traffic volume alone is not always correlated with more severe accidents, as many states and localities with notoriously high volumes of traffic did not register abnormally high amounts of severe accidents.

After this is where I ran into a lack of quantitative information and had to turn to qualitative sources. Numerous news reports were easy to find regarding road widening projects for different stretches of I-40 and I-30, with a general opinion that these efforts were much-needed and likely overdue. Also, the presence of large amounts of national forest in the state appeared to be a contributor towards more narrow highways and high traffic volumes. While it is impossible to make any firm conclusions given the lack of quantitative backing, it can be hypothesized that the amount of severe accidents in AR can be attributed to a combination of high traffic and freight volumes with roadways poorly equipped to handle these.

After exploration of these outliers, modeling could begin to predict severity. First, features with very little explanatory power were removed. Numerous features were then added based on information available from the accident descriptions that would be appropriate to use in a scenario when an accident is immediately reported. Such features included whether it was a multi-vehicle crash, whether a road/ramp was immediately blocked, whether it occurred in a construction zone, and several more. While more useful features were added, the data itself presented a challenge in model construction. The distribution by severity is extremely imbalanced, with 99.6% of accidents being classified as either a 2 or a 3. 1 and 4 combined for only .4% of all accidents, which begs the question of how exactly severity is defined given that so few fit the criteria for 1 or 4. In fact, there is no exact criteria for severity, just a general statement that it represents delay caused by the accident. With these factors, I considered lumping in 1s with 2s and 4s with 3s and turning this into a binary classification problem given the extreme imbalance and lack of clarity on the differentiation between classes. Ultimately, I decided to go forward with this as a multi-classification problem.

Initially, I ran an untuned random forest model with roughly balanced classes via resampling minority instances. Results were accuracies of over 92% for levels 2 and 3 accidents, but 0% for 1 and 4 accidents. The model was completely ignoring minority instances. At a crossroads here, fine-tuning my model would not fundamentally fix this issue. With many possible options, I chose to try synthetically generating minority instances rather than resampling them. This made a large difference in my model, as training on this data improved accuracy of level 4 accidents to 53.3% rather than 0%. This came at the cost of increased variance, as level 3 accuracy dropped to 73% from 92%, and level 1 accidents remained ignored. While this method showed promise, more work was necessary. Level 4 accidents need to be differentiated more from level 3 via their attributes, and I performed text analysis of the descriptions to this end. The results found were interesting and will be highly useful, as there are numerous words that appear at far different frequencies in level 4 descriptions vs level 3, and these can be engineered into features. In the final phase, my intent is to finalize a robust dataset with very precisely chosen features, and fine tune the synthetic data generation and random forest algorithms.

Data Source

Arkansas Road Inventory (2017). *Arkansas GIS Office.* Retrieved from <https://gis.arkansas.gov/product/arkansas-road-inventory/>