**Driving Towards a Safer Future: Predicting Accident Severity and Delays Caused**

Mikhail Hershfeld- Spring 2020

**Introduction:**

Every year car accidents cause thousands of deaths, millions of injuries, and billions of dollars of expense in this country and around the world. America’s roadways are a huge source of human and financial loss, one that we have become unfortunately accustomed to and regard as an externality of our lifestyles and our current preferred transportation methods. The fact is, car accidents cause far too much damage to ignore, and efforts should and are being made to engineer our cars and roads to be safer. For my part, I worked with an extensive collection of accident data gathered from navigation software APIs to be able to predict accident severity measured by delays caused based on available information, with an emphasis on being able to predict a minority of serious accidents.

The primary goal of my analysis was constructing a model to predict these accidents, but the large volume of data presented other interesting possibilities for analysis. First, I wanted to make the predictions, based on geospatial, time, weather and other data. The data I am using rates severity on a 1-4 scale, determined by how much of a delay the accident causes (Moosavi, 2019). It is apparent shortly after an accident takes place how detrimental it is based on the ensuing backup, but it is not often apparent right away. I am seeking to be able to predict, based on only information that would be known from the very first accident report, how severe an anticipated delay is so that an accurate estimation can be immediately reported to drivers on the road. In a hypothetical use case, this model could be possibly embedded within GPS software and utilized to warn motorists, which is another reason I chose to use data that is generated from these same sources.

In addition to this, I decided to explore in greater depth any geographic anomalies found. If certain states or regions have higher prevalence of detrimental accidents it could be a result of poor infrastructure, unmanageably high traffic volume, or a number of other possibilities that would be interesting to explore. Based on whatever is found, there could be many opportunities to smoothen out the roadways of the country.

For my methodology, I first conducted a thorough exploratory data analysis after the data was cleaned and ready for use. There is a great deal of highly informative visualization potential given the geographic nature of this data, which led into further research. The nature of the predictive problem was challenging in several ways, which I will discuss. The data is highly imbalanced with multiple classes, and my goal is to correctly identify minority class instances without sacrificing predictive accuracy for other classes, using only a limited subset of features.

**Previous Approaches:**

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 possibility that is now the goal that every player in the mobility business is striving for. Reducing the number of detrimental, delay-causing collisions is an important component of this, and the topic of the relevant literature to review. Both Jianfeng, Hongyu, Jian, Liu & Haizhu and Garrido, Bastos, Almeida, & Elvas 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 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 a path I pursued given that the features I used are very similar to theirs. Also, Jianfeng, et al’s methodology of precise feature selection was 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 (Wang, Li, 2019).

So while existing research centers mainly around predicting severity in terms of vehicle damage or mortality, the data and methodology is similar enough that I can make several takeaways from the approaches. Identifying population levels around accidents is important, and the success of tree-based modeling is encouraging.

**Exploratory Data Analysis and Proposed Solution:**

Accident severity was found to be highly bimodal and imbalanced, with the huge majority of observations classified as either a 2 or 3 out of 4. As mentioned earlier, this complicates modeling significantly. 99.6% are accidents were classified as either a 2 or a 3. Given that 1 and 4 combined for only .4% of all accidents, this begs the question of how exactly severity is defined given that so few fit the criteria for 1 or 4. After further exploration into the source of the data, the publisher acknowledged that there is no exact known criteria for severity. This is a metric that Mapquest and Bing, the sources of the data, do not make public (Moosavi, 2019). The data was specifically chosen from these sources with a possible use case in mind, and I was determined to see what could be done with it. With this known, 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 remove level 1 accidents in my final model and keep level 4’s as they were and move forward with this as a multi-classification problem. Level 1s are the smallest minority class and are of little importance to accurately classify for my goals, so I made the decision to exclude them. Level 4s are closer in similarity to Level 3s, and if my model were to end up mainly misclassifying 3s as 4s and vice versa, this would not be functionally much different than if it were a binary classification problem, and 3s and 4s were lumped into a single category.

With this established, it was shown from previous research that the roadway type appeared to have the strongest influence on severity, so creative feature engineering to extract more location information is a necessity. 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 the roadway type, whether it was a multi-vehicle crash, whether a road/ramp was immediately blocked, whether it occurred in a construction zone, and several more. After the initial feature engineering to increase predictive power, I set to answer my secondary objective of finding geographic anomalies.

West Virginia and Arkansas were both found to have statistically significantly higher percentages of severe accidents, with a satisfactorily large sample size. 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 (Arkansas Road Inventory, 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 (Arkansas Road Inventory, 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 (Arkansas Democrat-Gazette, 2019). 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.

With exploratory data analysis completed, I decided upon a multiclassification random forest model to begin testing. This stems from their versatility, my familiarity with them, and their reported success in previous research.

**Model Implementation:**

Prior to removing level 1 accidents, I left them in and conducted a baseline model test to see what I had to work from. The first iteration of testing was a default random forest model with untuned hyperparameters attempting to predict all 4 classes. Results were accuracies of over 92% for levels 2 and 3 accidents, but 0% for 1 and 4 accidents. Even with oversampled minority instances during training, the model was completely ignoring minority instances. Simply fine tuning this approach would be insufficient to achieve an acceptable level of accuracy, so another approach was necessary. I decided upon synthetically generating more minority instances rather than oversampling them, to improved results. The SMOTE (synthetic minority oversampling technique) algorithm as I implemented it finds the k-nearest neighbors of each minority, picks one of these points, and generates a new data point somewhere in the Euclidian space between the two points. This led to 53% accuracy classifying level 4 accidents, a great improvement from 0%, though at the expense of an 18% accuracy drop in level 3 classification, and level 1 still being ignored. At this point, I decided to drop level 1’s due to their lack of importance.

While results were significantly improved, there was still much work to be done. More features were necessary to differentiate level 4 accidents, so I conducted some text analysis of the descriptions of level 3 and 4 accidents to see if any more information could be extracted. Several words were found to be disproportionally more or less present in level 4 descriptions, which were engineered into new features. “Near”, “ramp”, “slow” and “center” appeared significantly less in level 4 descriptions, while “lanes”, “closed”, “emergency” and “detour” among others appeared significantly more often. Ultimately, I chose to include features indicative of an earlier accident, whether or not the roads had been reopened, whether there was an immediate detour, emergency vehicles immediately on the scene, whether the accident was near another event, and a flag for weekdays/weekend. Also, I added a measure for population density in the area of the accident. This was not done programmatically, rather I put the data into a GIS software, added in population density based on the county the accident was located in, and then exported the data back out. Several extremely spare features were also removed.

With a greater array of features now included in the data, it needed to be determined conclusively which ones were relevant. A chi-squared test was done to determine the statistical significance of categorical features per their distributions across severity levels, and all were shown to be significant. With all prudent preprocessing and feature engineering steps completed, a good model could be now fleshed out.

Along with the challenges of data imbalance, the sheer size of the dataset was a challenge in it of itself. For my modeling I was working on a sample of approximately 500,000 data points, and even this reduced amount required some extra steps to be taken for code to run relatively quickly. Sk-learn performed extremely slowly on the dataset of that size, so I turned to H2O, an ML library with built in parallel processing capabilities, which sped up modelling dramatically, especially useful for grid searches.

The best random forest model produced mixed yet decent results that held up on testing data. As before, the model is very robust at classifying level 2 accidents, but worse for level 3 and 4. Mean per-class-accuracy hovers around 77%, and overall is around 86%. A strong ROC\_AUC score was buoyed by the robustness of the model predicting level 2, and level 3 to an extent. On the flip side, a weak F1 score was caused by the lack of precision in predicting level 4 accidents. A one vs. rest support vector machine approach was tried due to penalized SVM’s strength at detecting minority classes, but the results were not good.

Ultimately, there appears to be a ceiling on how effective a given model can be with this specific data and task, which is something I suspected coming in. Overrepresenting severe accidents increases the amount of these that are correctly classified, but with a very low precision, and at the cost of accuracy for level 3 predictions. However, the results are still reasonably good. Predicting such a small minority class is reminiscent of modeling fraudulent financial transactions, a process in which it can be prudent to trade off precision for recall. Traffic delays are slightly less important to alert to, and this once again raises the question as to the distribution of accident severities. In context, a model similar to mine would function as a “warning system”, indicating a probable moderate delay with a chance of being severe whenever a level 4 accident is predicted. However, the main sticking points with this type of work going forward is with the data and features. Many of the features were very sparse, and did not provide a very large amount of gini when used for splitting, as can be seen from the low optimal maximum depths and the overall model performance. Whether or not the accident occurred on an interstate highway was by far the most important feature, speaking again to the importance of location and road information. Vehicle and specific road data were difficult to find and could have made an impact on the analysis if it could have been. Ultimately, accurately predicting delays from immediately known variables is a very accomplishable task with the toolbox and modeling I have established. However, several important questions about the somewhat black box of data provided by the GPS services must first be clarified.

References

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

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

I-30 work to close lanes for two weeks. (2019, September 7). *Arkansas Democrat-Gazette.* Retrieved from <https://www.arkansasonline.com/news/2019/sep/07/i-30-work-to-close-lanes-for-two-weeks-/?news-arkansas>

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

Moosavi, S. (2020, January 17). US Accidents (3.0 million records). Retrieved from https://www.kaggle.com/sobhanmoosavi/us-accidents

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