Reviving our Infrastructure to Save Lives

Alec Arshavsky

1. Introduction

Though much data exists on vehicle accidents, particularly the more severe ones, there is often difficulty in using the data in ways that effectively prevent further accidents. Even states with the most initiative towards the use of data only analyze intersections that are prone to crashes. Intersections, however, only account for 40% of accidents in the United States, and only 20% of fatal collisions [6,8].

Even there, however, selecting intersections of choice relies mainly on accident frequency or even simply concerned citizens [4]. The process for ranking intersections will then take into account the circumstances of the intersection that causes a trend towards accidents and evaluate cost [5]. The process of identifying elements of the infrastructure in need of improvement relies mainly on the brute numbers and lengthy human analysis.

Therefore, this paper seeks to use the data on roads and traffic accidents to find roads that need to be improved or repaired in order to minimize accidents. This expands the scope from intersections alone and helps automate candidate selection for improving roadways. The dataset will be composed of road-related data and crash-related data, with manually selected improvement candidates serving as targets.

This work will be done in a manner accessible to state governments with more limited scopes of data than the most advanced, partially through testing with Google Maps APIs [7] and partially by limiting the most inaccessible data-types to those relating to traffic accidents, which through the virtue of police reports must be documented fairly well.

2. Related Work

There is much work in the field of analyzing traffic accidents, but most of it focuses on conditions leading to accidents and road features that contribute [3]. The specific work on selecting infrastructure improvement candidates includes mainly state department of transportation documentation and process guidelines as well as a limited collection of software such as Intersection Magic [2]. This software, however is limited in analytical ability and mostly creates frequency mappings and death-rate calculations.

3. Dataset and Features

A series of four datasets was obtained from the Iowa Department of Transportation, constituting a road network layout, average daily traffic, crash data from the past ten years, and intersection improvement candidates (top 1000). The features ultimately extracted from these datasets mapped road information (number of lanes, width of dividing median, road length, speed limit, and traffic volume) as well as crash information (fatalities, damage, severity, influence of alcohol, and number of crashes) to each road. The crash information has been structured into the total number of fatalities and the average damage, severity, and alcohol influence. The intersection improvement candidates

were also mapped to each road as targets in three distinct manners. The first target metric is a binary indicator corresponding to whether the road contains any of the intersection improvement candidates. The next target metric is the minimum ranking value of any intersection found on the road, so that a road that has the #28 and #341 ranked intersections is marked 28. The last metric is the number of intersection improvement candidates the road contains.

There were 176,612 roadways as well as 986,753 crashes in the dataset. There were 1000 intersection improvement candidates. The data was split into 80% training data and 20% testing data.

The intersection improvement candidates creates two complexities for the data. Firstly, it applies specifically to intersections, while the goal of this paper is to generalize to roads. Because of this, the three target metrics are helpful but imperfect. Secondly, the intersections are selected through a manual process, which combines statistical data on crash frequency with manual subjective observations, such as from people who frequently use a particular route [4]. Once many municipal and county governments have compiled lists of intersections they would like funding for to improve, the state department of transportation evaluates if there is a trend in crash circumstances and what repairs may cost [5]. Because the ranking process itself is a manual process, subjective factors may affect this process.

In order to obtain the ultimate sets of features and targets, the raw data was first run through a series of preprocessing steps in order to map the data to specific roads in the network. The datasets did not have a standardized location format, and it was necessary to synthesize Longitude/Latitude, Street/County/City names, and road geometry formats into a coherent and unified matrix. Unfortunately, the format of the traffic volume dataset was such that there were duplicate identifiers, and this created noise in the dataset. Formatting the data into a cohesive combined dataset while minimizing the noise as well as experimenting with feature selection was one of the most challenging and substantial aspects of this work. The roads that did not have any associated accidents were then cut in order to reduce the weight of roads that did not need improvement.

Because the problem at hand is not limited to states like Iowa where detailed data is available, a Google Maps API [7] was tested on location data. This API accurately maps coordinates to name-based locations, and the only direct limitation is that only one of the streets at an intersection is returned. This limitation may be fixed through minor tweaking. Though inaccessible for this project due to daily query limits, this API could be instrumental to replicating similar feature-sets in other states.

4. Models and Techniques

The first model used was the binomial target vector, indicating the presence of an intersection improvement candidate on a given road. This was chosen because it is a robust and efficient model for large datasets. It does not assume a linear relationship between features and targets and makes no assumptions on feature distribution. The feature model was trained using logistic regression, through stochastic gradient ascent, with the following update rule:

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)}$$

This model did not prove to be optimal for the dataset, however.

The next model used was support vector machines, using the same binomial target. SVMs are also good at modeling large datasets and are able to handle noisy datasets, which is important due to the potential error in duplicate road identifiers as well as the imperfection of the target metrics. Unfortunately the SVM model also proved to be less optimal with the settings and several kernels that were tried - radial basis function and polynomial.

The supervised learning algorithm that was found to be most promising was decision trees. The random forest model was settled on because it is able to handle non-linear features, is resilient towards noisy datasets, and can handle multi-class classification as well as regression [9]. It is more robust due to the use of multiple decision trees as compared to other forms of decision tree models. Due to preliminary success of this model, several methods of modeling were used with random forests.

The first technique pursued was a regression based on the intersection rank target metric. A regression-based Random forest was applied, and then a thresholding-based post-processing was applied. The thresholding method is a process developed in this work to convert regression data into binary data indicating whether a road should be considered a candidate for improvement. First, the output data is normalized within a range that excludes outliers on the low end (where the lowest numbers indicate the highest need for improvement), setting them to the minimum value once normalized. A threshold is then found to force values on either side to 0 and 1 depending on the rigor desired from the program. The results and discussion of program rigor are explained in the Results section.

The second regression-based technique used with the target metric that corresponded to the quantity of intersections requiring improvement on a given road. The thresholding method was applied here as well, with less outliers in the normalization process.

The final decision tree model used was a six class multi-classification system where roads where split into categories based on the prominence of improvement candidates and their rank. Five of the classes split roads into severity of improvement need in a discretized gradient, while the last class indicated no improvement necessary.

5. Results and Discussion

Before presenting the results, it is important to understand what they mean. In this work, the range of scrutiny is being expanded from intersections in need of improvement to entire roads. Therefore, while it is necessary for the program to have a baseline accuracy, the overall accuracy is not the most important metric. What is important to look at are the false positives and false negatives. Most importantly, a low number of false negatives indicates that there is indeed a good accuracy in the classification as well as further potential candidates for road repair. This is good because it can help spot improvements that need to happen that were not previously detected. A high number of false positives indicates a lesser degree of precision in the classification process, and while potentially these would be valid candidates, it was beyond the scope of this work to perform such extensive manual analysis, and results showing extremely high false positives were considered to be less favorable. False negatives are the second factor to consider in the results, which tended towards a high number of false negatives. False negatives indicate that the program is misclassifying many

roads as not in need of improvement as compared to their associated targets. Too many false negatives may signify that the model is missing some of the roads that need improvement, but this number must be compared with the false positives.

The logistic regression model had a high accuracy, but ultimately a high number of false positives (Table 1). Because manual analysis on that quantity of false positives was not reasonable within time constraints, this method was not deemed optimal.

Accuracy	92.47%
False Positives	68.09%
False Negatives	91.77%

Table 1. Results for Logistic Regression. High accuracy but high false positives.

The SVM model was ultimately the worst performing of the tested methods (Table 2).

Accuracy	86.46%
False Positives	89.92%
False Negatives	87.79%

Table 2. Results for the best of the SVM models. Low accuracy in all areas.

The random forest models yielded the best results, but the thresholds could be adjusted to include more candidates for improvement or less. Throughout adjustment, the accuracy was relatively stable, but the false positives and negatives correlated inversely. When a more rigorous threshold was applied, yielding a lower overall number of estimated road repair candidates, there was a high number of false negatives, but a low number of false positives (example can be seen in Table 3, Figure 4). On the other hand, relaxing the threshold greatly lowered false negatives but also increased false positives (Table 5).

Accuracy	93.54%
False Positives	17.20%
False Negatives	89.44%

Table 3. Shows results of a multiclass random forest model, where the threshold has been adjusted to

		Estimated Repair Priority:					
		No repair	Low	Moderate	Medium	High	Highest
Actual	No Repair	9844	1	1	1	5	8
	Low	147	2	0	0	2	3
	Moderate	119	1	11	0	0	1
	Medium	151	0	0	4	0	1
	High	118	0	0	0	21	1
	Highest	117	0	1	1	2	23

Table 4. Confusion matrix for the multi-class random forest results in Table 3. The rigorous threshold has signified a select few roads as in need of improvement. Further manual analysis of these false positive results yielded promising results.

Further manual analysis of the false positive results in Table 4 showed promising results as roads that may be in need of improvement, though further analysis of cost and other factors should be applied by someone who understands the current system of improvement candidate selection.

Accuracy	89.48%
False Positives	67.36%
False Negatives	50.34%

Table 5. These results are for a repair rank regression based model where the threshold is significantly loosened. The number of false negatives is significantly decreased, but there are many false positives, which would need to be further manually analyzed.

Accuracy	93.57%
False Positives	38.12%
False Negatives	82.85%

Table 6. An example of repair count regression, based on the number of intersection candidates on a given road. The threshold is between those of Tables 3 & 5.

These results as well as the manual analysis performed show much promise, but ultimately extensive further manual analysis is necessary in order to characterize the false positives as needing improvement or not. This same analysis would determine the validity of the target metrics used in these models. Another possible course of action is to confirm the results using unsupervised learning, which would be unhindered by the possible flaws of the targets. Ultimately, once the results are manually analyzed, this work will greatly help the automation of improving roadways to decrease accidents, potentially increasing the amount of lives saved with a given budget. Knowing which roads cause accidents can help implement many solutions, including ones that don’t cost money such as encouraging alternate routes.

References

[1] "Explore Open Data." *Home | Iowa Department of Transportation - Open Data Portal*. Iowa DOT, n.d. Web.

[2] "Pd' Programming's Intersection Magic." *Pd' Programming's Intersection Magic*. N.p., n.d. Web.

[3] Chong, Miao, Ajith Abraham, and Marcin Paprzycki. "Traffic Accident Analysis Using Machine Learning Paradigms." *Informatica* 29th ser. 89.98 (2005): n. pag. Web.

[4] "Improvement Process." *Improvement Process - Unsignalized Intersection Improvement Guide*. N.p., n.d. Web.

[5] "Crash Analysis." *CTRE*. Iowa State, n.d. Web.

[6] "Crash Factors in Intersection-Related Crashes: An On-Scene Perspective." *Nhtsa.dot.gov*. NHTSA, n.d. Web.

[7] "Getting Started | Google Maps Geocoding API | Google Developers." *Google Developers*. Google, n.d. Web.

[8] "Analysis of Fatal Motor Vehicle Traffic Crashes and Fatalities at Intersections, 1997 to 2004." *Nhtsa.dot.gov*. NHTSA, n.d. Web.

[9] Breiman, Leo, and Adele Cutler. "Random Forests Leo Breiman and Adele Cutler." *Random Forests - Classification Description*. UC Berkeley, n.d. Web.