EXAMINING PREDICTORS OF TRAFFIC ACCIDENTS WITH RANDOM FOREST CLASSIFIERS

EMMALINE MCKINNON, FLETCHER MURRAY, ASHLEY SPENCER

Abstract. Traffic accidents are common in the United States and analyzing patterns to identify high-risk factors can help cities implement preventative measures. We utilize random forest classifiers and gradient-boosted trees to determine the most important features in road infrastructures and weather conditions. Analysis of our models shows that weather is the strongest predictor of crash severity, but infrastructure also matters, with features like roundabouts linked to less severe crashes

1. Research question and overview of the data

Traffic accidents are among the leading causes of fatalities in the United States, with approximately 39,345 deaths in 2024 [Nat25]. To investigate the contributing factors of these accidents, we analyze the US-Accidents dataset originally introduced by Moosavi et al. [MSPR19] and later referenced in accident risk prediction research [MSP+19]. This dataset covers 49 states from February 2016 to March 2023 and is a conglomeration of records from the US and state departments of transportation, law enforcement, and traffic cameras/sensors. While extensive, it doesn't capture every accident, providing an incomplete view of incidents. Regardless, we set out to answer the following questions:

- Which weather conditions are the strongest predictors of traffic accidents in the US?
- Which infrastructure features are the most strongly associated with traffic incidents?

Previous work has utilized deep neural networks to identify high-risk areas for targeted prevention efforts [MSP+19]. To increase interpretability and facilitate understanding, we take a simplified approach. We utilize random forest and XG-Boost classifiers to identify the main predictors of accident severity.

2. Data Cleaning / Feature Engineering

In our analysis, we reduced the original 46 features to 24 relevant to weather and infrastructure as seen in Figure 2 and Tables 1 and 2

Date: 4 October 2024.

Wind Direction, a categorical feature, was one-hot encoded. Weather Condition is a qualitative description of the weather. Due to the variety of conditions, we engineered a score for these descriptions, giving +1 for keywords like 'rain' or 'thunder'. To account for few outliers with extreme values, scores were capped at 2. Infrastructure features were boolean indicators of accident-related elements. The target labels were the dataset's accident severity score, an ordinal value from 1–4.

Missing values were substituted as 0s on the logic that it was better to interpret a crash with NaN rain or NaN Traffic Signal as having 0 precipitation and no traffic signal.

The original dataset contained over 7 million reported incidents across the US. To facilitate analysis, we trimmed the data by taking a stratified sample across states of 1 million entries. 2023 Census records were utilized from [U.S24] to ensure geographic representation. For numeric values, outliers greater than 1.5 times the IQR of the feature were removed.

3. Data Visualization and Basic Analysis

We begin our analysis with a simple examination of the frequency of severity scores. Figure 1 shows the majority of our data contains crashes of severity two.

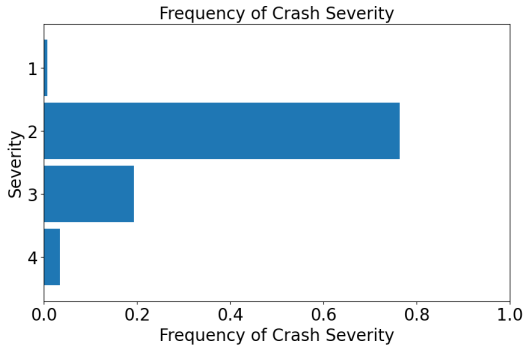


Figure 1. Distribution of crash severity

Weather Condition	Severity 1	Severity 2	Severity 3	Severity 4
0	0.008	0.763	0.195	0.034
1	0.007	0.725	0.235	0.033
2	0.007	0.696	0.266	0.030

Table 1. Proportion of Weather Conditions by Severity

Next, we examine the conditional distributions of weather condition with severity. As shown in Table 1, we can see that the proportion of weather conditions remains constant across severity scores. This mirrors the overall severity distribution, indicating our engineered weather score does not effectively predict crash severity.

Now we turn our attention to numerical weather data for insights. Figure 2 shows box plots of all relevant data. ANOVA tests grouped by severity was run for each feature all resulting in p values near 0. Thus the slight differences in our distributions are statistically significant, but there is little discernible pattern.

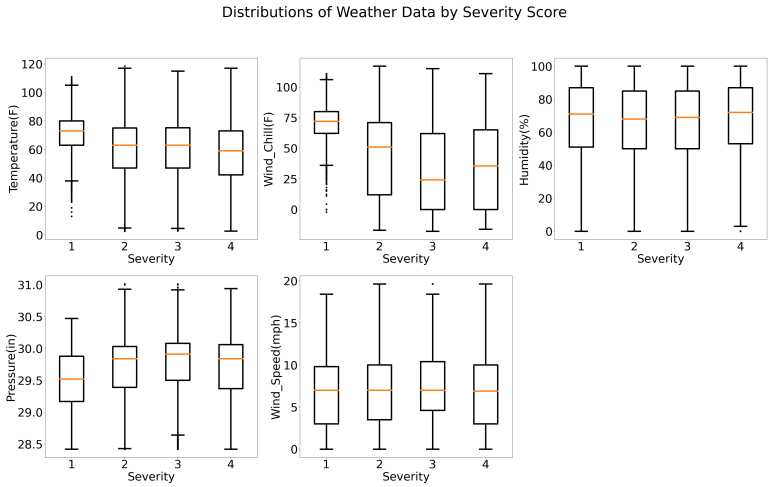


FIGURE 2. Box plots of numerical weather data broken down by severity score

Finally, we look at the frequencies of different infrastructures per severity score as seen in Table 2. Notably, this indicates roundabouts seem to be less related to severe crashes, but more fender benders.

Infrastructure	Severity 1	Severity 2	Severity 3	Severity 4
Amenity	0.0145	0.9132	0.0503	0.0219
Bump	0.0091	0.8973	0.0906	0.0030
Crossing	0.0182	0.8949	0.0691	0.0178
Give_Way	0.0119	0.8185	0.1386	0.0311
Junction	0.0039	0.7015	0.2492	0.0454
No_Exit	0.0163	0.8445	0.1153	0.0239
Railway	0.0176	0.8141	0.1355	0.0328
Roundabout	0.0000	0.9556	0.0444	0.0000
Station	0.0152	0.9010	0.0659	0.0178
Stop	0.0131	0.9096	0.0449	0.0323
Traffic_Calming	0.0068	0.8311	0.1385	0.0236
Traffic_Signal	0.0188	0.8711	0.0900	0.0200

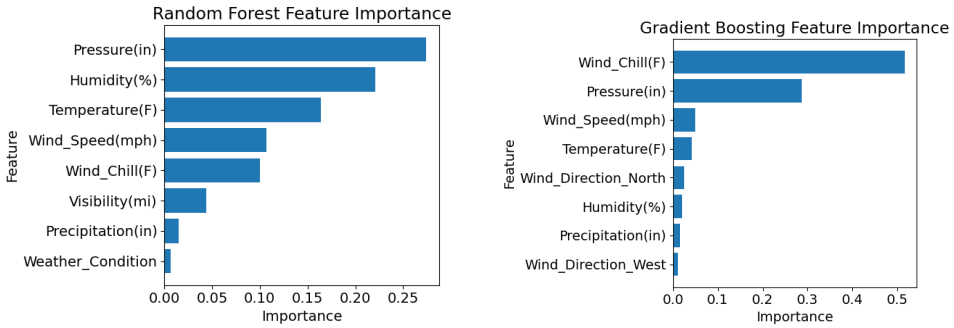
TABLE 2. Proportion of infrastructure features by crash severity

4. Learning Algorithms and In-depth Analysis

To further investigate the effects of weather and infrastructure, we utilize simple random forests to predict feature importance.

4.1. **Weather Analysis.** The random forest trained had an out-of-bag score of 0.7433. This means that the model correctly predicts the severity of the incident 74.33% of the time on data it hasn't seen during the training of the individual trees. We see in Figure 3a that the weather conditions with the highest importance are pressure, humidity, and temperature. This doesn't mean that these factors will cause the worst crashes, simply that these factors are the most predictive of a collision of any severity.

We also ran gradient boosting on our data. The boosted tree had an accuracy score of 76.28%(slightly better than the random forest); however, the runtime of boosting was much longer than that of the random forest. The strength of our dataset lies in the sheer magnitude of records, thus for computational feasibility we run random forests with bagging for the remainder of this project. As seen in Figure 3b the gradient boosted tree showed wind chill, pressure, and wind speed as the most predictive weather conditions.



(A) Random Forest Feature Importance (B) Boosting Feature Importance

FIGURE 3. Comparison of Feature Importance Methods

4.2. **Infrastructure Analysis.** We trained a random forest classification model to look at the severity of car crashes against the features of the road. Each decision tree in the forest is trained on different subsets of data and features, which will help prevent overfitting. The random forest model had an out-of-bag score of 76.31%. This means that the model correctly predicts the severity of the incident 76.31% of the time on data it hasn't seen during the training of the individual trees.

The infrastructure with the highest feature importance is traffic signals, crosswalks, and stop signs. This raises the question of whether traffic signals are the most predictive or if they are just more common. To investigate this, we implement permutation importance. This method evaluates each feature's contribution to the model's performance by randomly shuffling the values of the feature in the dataset and measuring the resulting drop in accuracy. Figure 4 shows the road features ranked by permutation importance.

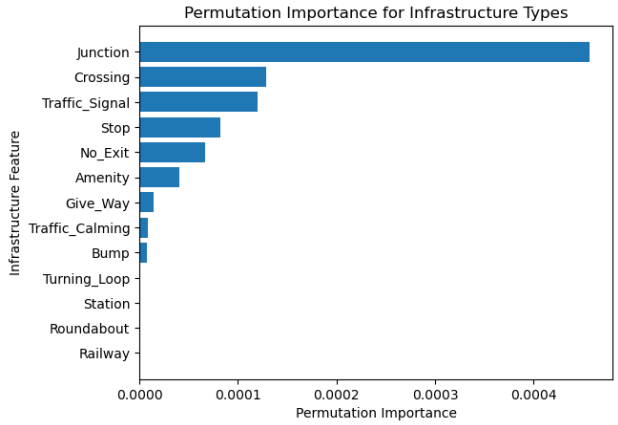


FIGURE 4. Feature Importance for Infrastructure

Overall, we see that junctions (intersections) have the highest importance with respect to car crash severity. Cross walks and traffic signals are ranked as the following two most important features.

5. Ethical Implications and Conclusions

When examining car crash severity, the junctions, traffic signals, and cross walks emerged as the most predictive features. Roundabouts, traffic calming signs, and yield signs (give way) were significantly less associated with crashes. To enhance the safety of cities, we can incorporate more roundabouts in place of traffic signals. For weather, wind chill, pressure, and wind speed were the strongest predictors in the gradient boosting tree. While the weather is outside our control, when hazardous weather conditions are expected, cities can issue alerts to help drivers stay aware and cautious.

While this model can identify the most important predictors of a car crash, we must consider the ethical implications and limitations of our model. Our analysis is sorely simplified by categorizing all crashes in severity 1-4. Additionally, this dataset included only information from the United States, meaning our findings cannot be reliably generalized to cities in other countries. Differences in driving culture and road construction practices introduce inherent biases that limit broader applicability. While models such as ours can influence infrastructure and safety decisions, they should not be used as the sole consideration for policy decisions. It is essential that decisions impacting individuals and communities remain firmly rooted in responsible human judgment.

Our models examine predictors of any accident, even if it is simply a scratch. However, to prevent fatalities, we hope to focus our analysis on more severe car accidents. In the future, we will examine the most predictive factors of crashes of severity three and four.

References

[MSP+19] Sobhan Moosavi, 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.

[MSPR19] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. A countrywide traffic accident dataset. *arXiv preprint arXiv:1906.05409*, 2019.

[Nat25] National Highway Traffic Safety Administration. Nhtsa estimates 39,345 traffic fatalities in 2024. `https://www.nhtsa.gov/press-releases/nhtsa-estimates-39345-traffic-fatalities-2024`, April 2025. Accessed: 2025-11-24.

[U.S24] U.S. Census Bureau. State Population Totals and Components of Change: 2020–2024. `https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html`, 2024. Accessed: 2025-11-15.