RISK AND SEVERITY OF U.S. CAR ACCIDENTS: A DATA ANALYSIS

DANIEL PERKINS, DRAKE BROWN, TREVOR GARRITY

Abstract. Every year, tens of thousands of people across America die in automobile accidents. Our analysis aims to provide actionable insights for policymakers and the transportation industry to mitigate accident severity and improve road safety. We base our results on Kaggle's US Accidents dataset, which contains information about the conditions and severity of 7.7 million accidents across America. We use Principal Component Analysis and Spearman's rank correlation to highlight key determinants such as road length affected, wind chill, and the presence of traffic signals. We analyze how accident trends evolved during the global pandemic of 2020. Furthermore, we employ a variety of predictive models, including Multinomial Regression, Random Forest, and K-Nearest Neighbors to reveal key insights from the data.

1. Research question and overview of the data

What conditions exacerbate the risk of accidents? Which factors lead to more severe outcomes? How did accident trends evolve during the global pandemic of 2020? Addressing these questions may provide valuable insights for the transportation industry in developing strategies to enhance automobile safety.

A priori, it can be assumed that hazardous road conditions influence the severity of an accident. For example it stands to reason that hazardous wrecks are more likely to occur if it is dark outside or if there is rain or ice that impairs a driver's ability. However, reasoning alone cannot provide us with enough information to determine which factors are the most significant.

Many papers have been devoted to better understand what causes automobile accidents. For example, [ALHZS24] find that various factors in an accident are correlated, most notably blood alcohol concentration, time of night, and lack of signage. [Fel76] finds that 40-50% of accidents result due to driver decision or recognition errors. [CHK$^+$11] find that driver error is by far the most common cause of automobile accidents among teens (95%). Lastly, [KNRL95] found that alcohol consumption, seatbelt use, and driver error were factors that increased severity of an accident. Our work hopes to complement [KNRL95] by exploring how attributes of the environment contribute directly to the severity of an accident.

Date: 4 October 2024.

For our analysis, we leverage Kaggle's comprehensive US Accidents dataset ([Moo23]), which details about 7.7 million car accidents across the 48 contiguous U.S. states (and D.C.) from February 2016 to March 2023. This dataset offers a severity ranking for accidents on a scale of 1 to 4, providing valuable insights into various types of incidents. It also captures extensive contextual information, including the location, date, time of occurrence, weather conditions (e.g., temperature, humidity, visibility, wind speed), and road conditions at the time of the accident. With 7.7 million entries, the dataset's significant size ensures a robust representation of traffic accident patterns across the United States.

2. Data Cleaning / Feature Engineering

To begin our analysis, we addressed several issues within the dataset, starting with the uneven distribution of severity ratings. Over 79.67% of the accidents were classified as level 2, while only 0.87% were classified as level 1, creating a significant imbalance that could skew classifier performance. To reduce inaccurate predictions for less common severity levels, we employed a down-sampling strategy, randomly removing many of the entries from the most frequent severity levels. This adjustment reduced the dataset to 516,006 training entries and 129,005 test entries, resulting in a more balanced distribution of labels and a significant reduction of the temporal complexity of our algorithms.

Furthermore, many features in the dataset did not contribute significantly to the analysis. We remove the following columns:

- **Irrelevant Features**: Accident ID, airport code, source of information, wind direction
- **Redundant Features**: The street, zip code, city, county, state, time zone, and country are already determined by starting longitude and latitude. The end longitude and latitude are largely determined by starting longitude and latitude and by distance. The weather time stamp is largely determined by the start time. Wind chill is determined by temperature and wind speed. Civil, nautical, and astronomical twilight do not add much more information than the sunrise/sunset feature.
- **Nominal Features**: Accident description is completely nominal and would require a language model to extract any meaningful information.
- **Disproportionate Features**: Any boolean column (even after one-hot encoding) in which less than 10,000 instances of the data were True was removed. For example, a very small percentage of accidents indicated a "roundabout" or "bump" on the road, and many different weather conditions occurred less than 10 times.

In addition, precipitation and wind speed were missing when there was no precipitation or no wind. We therefore filled in their missing values with 0.

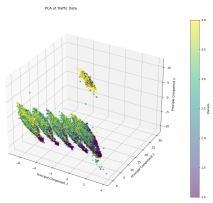

FIGURE 1. The PCA was run on a subset of the data for better visualization, but on the whole data it looks similar. Color indicates the severity of the accident with yellow being most severe and blue being the least.

We dropped any remaining instances with missing values (about 200,000 of the 7.7 million). We also applied one-hot encoding to categorical features, such as weather conditions and sunrise/sunrise. Finally, to ensure equal weighting across columns, we shifted and scaled each non-boolean feature to have mean 0 and standard deviation 1.

It is important to note that these design choices are not the only methods of feature engineering in our dataset. Although significantly reducing data size may have some unintended consequences, the benefits of removing extraneous columns, one-hot encoding categorical variables, and normalizing numerical data were substantial. These steps provided a clearer framework for interpreting our data with numerical classifiers.

3. Data Visualization and Basic Analysis

We performed Principal Component Analysis (PCA) on the scaled data (see Figure 1), reducing it first to two components and then to three. From these plots, we observe that the second principal component divides the data based on the availability of sunset/sunrise information, a factor that appears correlated with accident severity. In the x-z plane, representing the first and third principal components, the data form stratified disc-shaped structures, which seem to correspond to varying levels of accident severity. We believe it is likely these disc structures also correspond to accidents that occur near each other in location.

Notice that the number of severity 1 car crashes had a spike near the time that Covid-19 began spreading throughout the United States. We believe that this is likely due to a combination of over-reporting during Covid-19 as well as quarantine encouraging shorter low risk drives (e.g. to the supermarket and back).

4. Learning Algorithms and In-depth Analysis

4.1. **Feature Correlation.** To determine which characteristics play the greatest role in determining the severity level of car accidents, we calculated

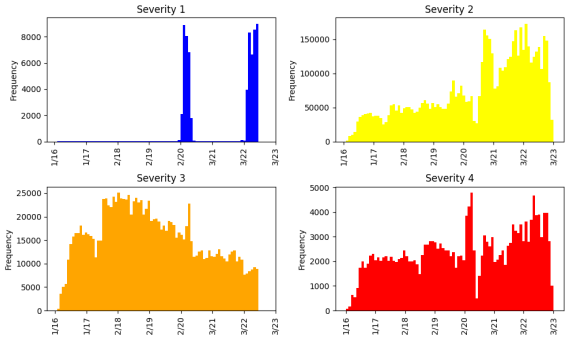

FIGURE 2. Car crashes per severity level over time.

the Spearman rank correlation using the formula:

$$r_s = \frac{\sum(R(X_i) - \overline{R(X)})(R(Y_i) - \overline{R(Y)})}{\sqrt{\sum(R(X_i) - \overline{R(X)})^2 \sum(R(Y_i) - \overline{R(Y)})^2}}$$

where X_i is the value of the severity of the accident of the ith data point and R is a measure of the rank ([Spe04]). This measurement for correlation is a modified version of Pearson's correlation coefficient. It is better suited for ordinal data, making it a good choice for our analysis.

Some of the features with the highest magnitude correlations were the length of the road affected by the accident (0.4000), the start time (-0.2254), the elapsed time of the accident (0.2182), and the presence of a traffic signal (-0.1961). Each of these factors had a moderate impact on the severity of the accident. The features with the smallest impact, or the lowest correlation, were the presence of certain weather conditions such as cloudy and windy.

4.2. **Learning Algorithms.** To predict the severity levels of car accidents, we evaluated a variety of machine learning models. The models selected for this analysis included multinomial regression, support vector machine (SVM), random forest, XGBoost, and K-nearest neighbors (KNN). First, we fine tuned model hyperparameters by randomly drawing hyperparameters from a sensible distribution (such as uniform, loguniform, or exponential) and then comparing the accuracy of the various choices of hyperparameters using 5-fold cross validation. After finding the corresponding hyperparameters, each model was trained and tested to assess its ability to accurately predict the severity of the accident (compared to random chance, which is 25% accurate). The following table summarizes the results.

XGBoost outperformed all other models by a significant margin. Random forest classifiers performed similarly. It therefore appears that the XGBoost regularization terms improve both speed and performance (compared to random forests). These models also found the most important features to be the distance of road affected by the accident, the precense of a traffic signal, and the starting time of the accident.

Model	Train Accuracy	Test Accuracy	Train time (s)
Multinomial Regression	53.0%	53.3%	178
Support Vector Machine	52.6%	52.8%	48
Random Forest	77.3%	76.8%	698
XGBoost	88.2%	81.7%	95
K-Nearest Neighbors	51.8%	51.7%	0.006

TABLE 1. Learning Model Results

It is interesting to note that multinomial logistic regression, support vector machines, and k-nearest neighbors all attained nearly identical accuracies. This seems to imply that most classifiers are capable of extracting some basic information from the dataset, but only random forests and XGBoost were able to extract further information. It is also interesting that k-nearest neighbors trained so quickly, but it should be denoted that it took several minutes for it to make predictions on the full training and test data (offsetting any saved training time).

5. Ethical Implications and Conclusions

Although our analysis offers valuable information that could help government agencies improve automobile safety, it is essential to carefully consider the ethical implications associated with the dataset and our approach. There are a number of potential issues: privacy of data, potential imbalanced distribution of resources, bias toward highly populated areas, harm to local economies, and negative feedback loops. For brevity, we discuss two, but invite the reader to consider these situations thoughtfully.

It should be noted that areas with larger populations are more likely to report accidents simply because of the greater presence of people and infrastructure. Thus, areas with fewer reports but similar risks may not receive necessary infrastructure improvements. This inherent bias risks reinforcing assumptions that urban areas alone require resources for safety measures, potentially neglecting rural or less populated areas.

Prioritizing funding based solely on accident reports might prompt increased police presence and reporting in those areas. This in turn could amplify perceived accident rates, creating a self-sustaining cycle of investment and reporting that neglects areas with genuine but underreported needs. For this reason, we urge caution in applying our analysis to policy-making. High reporting rates in a specific location do not necessarily indicate a proportionally high occurrence of accidents in that area.

To mitigate these potential misinterpretations, we emphasize that our study explores correlations rather than causation between factors. Although we hope our findings are able to lower accident severity, our findings should be interpreted with care. Any actions taken should be the result of thoughtful planning and forsight.

References

[ALHZS24] Mohammad Reza Abbaszadeh Lima, Md Mahmud Hossain, Huaguo Zhou, and Yukun Song. Data mining approach to explore the contributing factors to fatal wrong-way crashes by local and non-local drivers. *Future Transportation*, 4(3):985–999, 2024.

[CHK+11] Allison E Curry, Jessica Hafetz, Michael J Kallan, Flaura K Winston, and Dennis R Durbin. Prevalence of teen driver errors leading to serious motor vehicle crashes. *Accident Analysis & Prevention*, 43(4):1285–1290, 2011.

[Fel76] James C Fell. A motor vehicle accident causal system: the human element. *Human Factors*, 18(1):85–94, 1976.

[KNRL95] Karl Kim, Lawrence Nitz, James Richardson, and Lei Li. Personal and behavioral predictors of automobile crash and injury severity. *Accident Analysis & Prevention*, 27(4):469–481, 1995.

[Moo23] Sobhan Moosavi. Us accidents (2016 - 2023), 2023.

[Spe04] Charles Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101, 1904.